Axiomatizing $\mathcal{EL}^\bot$-Expressible Terminological Knowledge from Erroneous Data

Daniel Borchmann*
TU Dresden, Germany
daniel.borchmann@mailbox.tu-dresden.de

ABSTRACT

In a recent approach, Baader and Distel proposed an algorithm to axiomatize all terminological knowledge that is valid in a given data set and is expressible in the description logic $\mathcal{EL}^\bot$. This approach is based on the mathematical theory of *formal concept analysis*. However, this algorithm requires the initial data set to be *free of errors*, an assumption that normally cannot be made for real-world data. In this work, we propose a first extension of the work of Baader and Distel to handle errors in the data set. The approach we present here is based on the notion of *confidence*, as it has been developed and used in the area of data mining.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Representation Languages*; I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge Acquisition, Concept Learning*

1. INTRODUCTION

Every logic-based expert system needs a certain amount of knowledge about the domain it is supposed to act in. This knowledge must be represented in a way suitable for the expert system to take actions and to make decisions. One of the most popular ways to present such knowledge is to use *description logic knowledge bases* (also called *ontologies*). Therein, description logics make use of *concept descriptions* to express two kinds of knowledge: *assertional knowledge* stating that certain *individuals* satisfy a particular concept description, and *terminological knowledge* stating certain relationships between concept descriptions. For example, we could state that the individual Tom is a cat, and that every cat is a mammal that hunts a mouse. This could be written in the description logic $\mathcal{EL}^\bot$ as

$$\mathsf{Cat}(\mathsf{Tom}),$$
$$\mathsf{Cat} \sqsubseteq \mathsf{Mammal} \sqcap \exists \mathsf{hunts}.\mathsf{Mouse}.$$

*Supported by DFG Graduiertenkolleg 1763 (QuantLA)

K-CAP '13, June 23-26, 2013, Banff, Canada.

These are two forms of *axioms*, where the former is called an *assertional axiom* and the latter is called a *$\mathcal{EL}^\bot$-general concept inclusion (GCI)* and is particular example of a *terminological axiom*. An ontology is then just a collection of assertional axioms (collected in the $ABox$ of the ontology) and terminological axioms (collected in the $TBox$).

One of the biggest obstacles in using ontologies lies in the difficulty of their construction. One of the most popular example of an ontology that is used in real-world applications is the Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT) [17], a medical ontology that is used to guarantee a common language for medical treatment, among others. The development of SNOMED-CT took more than forty years and the ontology itself now contains more than 311,000 concept descriptions. Clearly, developing and maintaining SNOMED-CT is an expensive and time-consuming task, which is mostly conducted by human experts.

On the other hand, ontologies are only a different way to represent knowledge which is already available, mostly in a form which is not suitable for computers to be processed easily, e. g. as textual publications. It would be advantageous to provide methods to automatically extract ontologies from such data. Of course, this would be a challenging task and the resulting ontologies may not be as good as one would like them to be. However, starting from these ontologies, one could use other methods to refine these ontologies, as for example *axiom pinpointing* [7], to find reasons for errors in the ontology, or *completion methods* [6] to ensure that all relevant facts are present in this ontology.

A recent approach in the direction of learning ontologies from data is the work of Baader and Distel [12, 5, 4]. This work concentrates on extracting compact representations (*bases*) of all valid $\mathcal{EL}^\bot$-GCIs of a finite *interpretation*. Description logic interpretations are used to define the semantics of description logics, and are essentially vertex- and edge-labeled graphs. As such, they can also be seen as a different flavor of *linked data* [8]. If the interpretation contains instances of a particular domain of interest for which we would like to have an ontology, then GCIs valid in this interpretation can be seen as terminological knowledge which should be present in our ontology. Thus, we can view the work of Baader and Distel as a way to extract terminological knowledge from linked data, and thus as a way to (partially) construct ontologies form linked data.

A disadvantage of the approach of Baader and Distel is that it requires the initial data to be *perfect*, i. e. free of errors, simply due to the fact that only valid GCIs are considered.

In other words, errors in the data may lead to certain GCIs not being found, and a human expert would have to add them manually.

A natural attempt to circumvent the restriction of perfect data is the following: instead of considering only valid GCIs of an interpretation, one could also consider GCIs which are *almost* valid, i. e. valid except for a small set of *counterexamples*. Of course, this may lead to certain GCIs being found which are not correct, but whose set of counterexample is just too small. However, one could argue that it is much more difficult for a human expert to come up with missing GCIs than identifying GCIs which are not correct.

To formalize the idea of GCIs being "almost true," one can use the notion of *confidence* from association rule mining [1]. The goal of this paper is to generalize results of Baader and Distel to GCIs with *high confidence*, i. e. to find a base of all GCIs which have high confidence in a given interpretation. With respect to practical applications, this would mean that we can extract from a certain amount of linked data terminological knowledge while ignoring rarely occurring errors. The resulting set of GCIs can then be used for further refinement to obtain suitable ontologies.

This paper is structured as follows. We shall first introduce the necessary definitions from the fields of description logics and formal concept analysis. Due to technical reasons, we shall also introduce the description logic $\mathcal{EL}^{\perp}_{\text{gfp}}$, an extension of $\mathcal{EL}^{\perp}$ by cyclic concept descriptions. We shall also review the main results of [12] that are relevant for our considerations. Thereafter, we present the notion of confident GCIs and confident bases in Section 5.1 and illustrate the use of these notions by means of an example. After this, we describe a way to obtain confident $\mathcal{EL}^{\perp}_{\text{gfp}}$-bases in Section 5.2, and how to transform them into $\mathcal{EL}^{\perp}$-bases in Section 5.3. We finish this paper with an outlook on further research.

The results of this paper are based on two technical reports [10, 11].

2. RELATED WORK

As already noted, our work is largely based on the work of Baader and Distel. However, it also bears some similarities to *Statistical Schema Induction*, as presented in [19]. Statistical Schema Induction is a method constructing terminological axioms expressible in $\mathcal{EL}^{+}$ from sets of RDF triples. For this, *data tables* are constructed from the RDF triples whose attributes are certain concept descriptions. From these data tables terminological axioms are extracted by means of association rule mining, and these rules are then used to yield terminological axioms.

While this approach is quite similar to ours, it has some fundamental differences. The most notable one is that in Statistical Schema Induction only *certain* axioms are extracted (namely those constructible from attributes in the data tables), whereas in our approach a compact representation of *all* terminological axioms with high confidence and expressible in $\mathcal{EL}^{\perp}$ is found. Thus in contrast to Statistical Schema Induction, our approach enjoys a certain form of *completeness*.

As our approach makes use of formal concept analysis in conjunction with data mining methods it is also worth to note previous works in this direction. Most notably are Stumme et. al. [18], which use formal concept analysis to obtain compact representations of association rules by means of Luxenburger's base [15], and Zaki et. al. [20], which use formal concept analysis to provide a theoretical foundation of the theory of association rules.

3. PRELIMINARIES

In the following subsection we shall introduce the necessary definitions from the fields of description logics and formal concept analysis, as they are needed for our further considerations.

3.1 The Description Logics $\mathcal{EL}^{\perp}$ and $\mathcal{EL}^{\perp}_{\text{gfp}}$

In this section, we shall introduce the description logic $\mathcal{EL}^{\perp}$, and to a certain extent also the description logic $\mathcal{EL}^{\perp}_{\text{gfp}}$. However, we shall leave out certain details due to space restrictions.

Let N_C and N_R be two disjoint sets. We think of N_C as the set of *concept names* and of N_R as the set of *role names*. An $\mathcal{EL}$-concept description C is then formed according to the rule

$$C ::= A \mid \top \mid C \sqcap C \mid \exists r.C$$

where $A \in N_C$ and $r \in N_R$. For example, if we choose $N_C = \{\, \mathsf{Cat}, \mathsf{Mouse} \,\}$ and $N_R = \{\, \mathsf{hunts} \,\}$, then some valid $\mathcal{EL}$-concept descriptions are

$$\mathsf{Cat},\ \mathsf{Cat} \sqcap \mathsf{Mouse},\ \mathsf{Cat} \sqcap \exists \mathsf{hunts}.\mathsf{Mouse}.$$

An $\mathcal{EL}^{\perp}$-concept description is either $\perp$ or an $\mathcal{EL}$-concept description.

The semantics of $\mathcal{EL}^{\perp}$-concept descriptions are defined through the notion of *interpretations*. An interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ consists of a set $\Delta^{\mathcal{I}}$ of *elements* and an *interpretation function* $\cdot^{\mathcal{I}}$ which maps concept names $A \in N_C$ to subsets $A^{\mathcal{I}}$ of $\Delta^{\mathcal{I}}$ and role name $r \in N_R$ to subsets $r^{\mathcal{I}}$ of $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. Equivalently, interpretations can be thought of as vertex- and edge-labeled graphs. Such a graph consists of the vertex set $\Delta^{\mathcal{I}}$ and the labels of a vertex $x \in \Delta^{\mathcal{I}}$ are all $A \in N_C$ such that $x \in A^{\mathcal{I}}$. An edge (x, y) between two elements $x, y \in \Delta^{\mathcal{I}}$ exists if and only if there exists an $r \in N_R$ such that $(x, y) \in r^{\mathcal{I}}$. The edge (x, y) is then labeled with all $r \in N_R$ such that $(x, y) \in r^{\mathcal{I}}$.

The interpretation function $\cdot^{\mathcal{I}}$ can naturally be extended to all $\mathcal{EL}^{\perp}$-concept descriptions C, D in the following way:

$$\begin{aligned} \perp^{\mathcal{I}} &= \varnothing \\ \top^{\mathcal{I}} &= \Delta^{\mathcal{I}} \\ (C \sqcap D)^{\mathcal{I}} &= C^{\mathcal{I}} \cap D^{\mathcal{I}} \\ (\exists r.C)^{\mathcal{I}} &= \{\, d \in \Delta^{\mathcal{I}} \mid \exists e \in \Delta^{\mathcal{I}} \colon (d, e) \in r^{\mathcal{I}} \wedge e \in C^{\mathcal{I}} \,\}, \end{aligned}$$

where $r \in N_R$. We shall call the set $C^{\mathcal{I}}$ the *extension* of C in $\mathcal{I}$. For each $x \in \Delta^{\mathcal{I}}$ we shall say that x *satisfies* C if and only if $x \in C^{\mathcal{I}}$.

As we shall see in Section 4, the description logic $\mathcal{EL}^{\perp}$ is not sufficient for our needs, as it in general does not allow us to represent *model-based most-specific concept description* (which will be introduced then). We therefore also want to introduce the description logic $\mathcal{EL}^{\perp}_{\text{gfp}}$, an extension of $\mathcal{EL}^{\perp}$ which uses greatest fixpoint semantics.

The main distinction of $\mathcal{EL}^{\perp}_{\text{gfp}}$ over $\mathcal{EL}^{\perp}$ is that it allows for cyclic concept descriptions. More formally, let N_D be a set disjoint to both N_C and N_R. We call this set the set of *defined concept names*. A *concept definition* is then an expression of the form $A \equiv C$, where $A \in N_D$ and C is

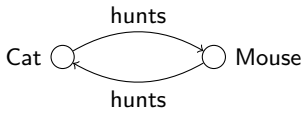

Figure 1: Graph of an $\mathcal{EL}^{\bot}_{\mathbf{gfp}}$-concept description

an $\mathcal{EL}^{\bot}$-concept description which can use in the place of concept names from N_C also defined concept names from N_D. Let $\mathcal{T}$ be a finite set of concept definitions. Then the set $N_D(\mathcal{T})$ is defined as the set of all defined concept names from N_D that appear in some concept definition in $\mathcal{T}$. $\mathcal{T}$ is called a *cyclic TBox* if and only if each element in $N_D(\mathcal{T})$ appears exactly once on the left-hand side of a concept definition from $\mathcal{T}$.

An $\mathcal{EL}_{\text{gfp}}$-concept description is a pair $(A, \mathcal{T})$, where $\mathcal{T}$ is a cyclic TBox and $A \in N_D(\mathcal{T})$. An $\mathcal{EL}^{\bot}_{\text{gfp}}$-concept description is either of the form $\bot$ or is an $\mathcal{EL}_{\text{gfp}}$-concept description. As an example of an $\mathcal{EL}^{\bot}_{\text{gfp}}$-concept description we can consider the concept description C, where

$$C = (A, \{\, A \equiv \mathsf{Cat} \sqcap \exists \mathsf{hunts}.B \tag{1}$$
$$B \equiv \mathsf{Mouse} \sqcap \exists \mathsf{hunts}.A \,\}). \tag{2}$$

Intuitively, C represents a cat A, which hunts a mouse B which again hunts A, a common situation in the old cartoon series "Tom and Jerry."

To define the semantics for $\mathcal{EL}^{\bot}_{\text{gfp}}$ it is necessary to resolve the cyclic dependencies within $\mathcal{EL}^{\bot}_{\text{gfp}}$-concept descriptions. This is done using *greatest fixpoint semantics* [16, 3]. Since the exact definition of the semantics of $\mathcal{EL}^{\bot}_{\text{gfp}}$ is not necessary for our further considerations, we leave out the details and refer the reader to the corresponding publications.

Instead, we want to describe the technique of *unravelling*, which shall allow us to transform $\mathcal{EL}^{\bot}_{\text{gfp}}$-concept descriptions into $\mathcal{EL}^{\bot}$-concept descriptions in a suitable way. Let $C = (A_C, \mathcal{T}_C)$ be an $\mathcal{EL}^{\bot}_{\text{gfp}}$-concept description. We assume that C is *normalized*, i. e. every concept definition in $\mathcal{T}_C$ is of the form

$$B \equiv P_1 \sqcap \ldots \sqcap P_n \sqcap \exists r_1.Q_1 \sqcap \ldots \exists r_m.Q_m,$$

where $P_1, \ldots, P_n$ are concept names from N_C, $r_1, \ldots, r_m \in N_R$ and $Q_1, \ldots, Q_m$ are defined concept names from N_D. Normalizing C can be achieved in polynomial time [2]. We then can associate with C a labeled graph with vertices $N_D(\mathcal{T})$. Roughly speaking, the vertex B from above is labeled with $P_1, \ldots, P_n$ and has edges to $Q_1, \ldots, Q_m$, each labeled with $r_1, \ldots, r_m$, respectively.

If now $k \in \mathbb{N}$ is given, we can unravel this graph up to depth k, keeping all labels. It is not hard to see that for this graph we can associate an $\mathcal{EL}^{\bot}$-concept description C_k that is represented by the graph. For example, we can consider the concept description C from (1), which is already normalized. Its graph is shown in Figure 1. Then

$$C_2 = \mathsf{Cat} \sqcap \exists \mathsf{hunts}.(\mathsf{Mouse} \sqcap \exists \mathsf{hunts}.\mathsf{Cat}).$$

For convenience, we define $C_0 = \top$.

We are going to use unravelling to transform $\mathcal{EL}^{\bot}_{\text{gfp}}$-bases into $\mathcal{EL}^{\bot}$-bases. For this, the following property of unravelling will be helpful.

3.1 Lemma (Lemma 5.19 of [12]) *Let C and D be $\mathcal{EL}^{\bot}_{gfp}$-concept descriptions. Then for each $k \in \mathbb{N}$ it is true that*

i. $(\exists r.C)_k \equiv \exists r.C_{k-1}$ *for* $k > 0$ *and* $r \in N_R$*;*

ii. $(C \sqcap D)_k \equiv C_k \sqcap D_k$.

It can be shown that $C \sqsubseteq C_k$ is true for each $k \in \mathbb{N}$. Even more, if $k_1 < k_2$, then $C_{k_2} \sqsubseteq C_{k_1}$. However, as the interpretation $\mathcal{I}$ is finite, the sequence

$$(C_0)^{\mathcal{I}} \supseteq (C_1)^{\mathcal{I}} \supseteq (C_2)^{\mathcal{I}} \supseteq \ldots$$

must eventually stabilize. As the following lemma shows, this can be achieved uniformly for all $\mathcal{EL}^{\bot}_{\text{gfp}}$-concept descriptions.

3.2 Lemma (Lemma 5.5 of [12]) *Let $\mathcal{I}$ be a finite interpretation. Then there exists a number $d \in \mathbb{N}$ such that $(C_d)^{\mathcal{I}} = C^{\mathcal{I}}$ is true for all $\mathcal{EL}^{\bot}_{gfp}$-concept descriptions C.*

3.2 General Concept Inclusions

As we have already mentioned, description logics can be used to formalize knowledge in the form of *description logic knowledge bases*. In particular, these knowledge bases contain *terminological knowledge*, whose automatic extraction from data is the main focus of this work.

Terminological knowledge is represented in the form of *general concept inclusions* (GCIs). These are expressions of the form $C \sqsubseteq D$, where C and D are concept descriptions. We speak of $\mathcal{EL}^{\bot}$-GCIs if both C and D are $\mathcal{EL}^{\bot}$-concept descriptions, and likewise for other description logics. An example of an $\mathcal{EL}^{\bot}$-GCI would be $\mathsf{Cat} \sqcap \mathsf{Mouse} \sqsubseteq \bot$, intuitively stating that nothing can be at the same time a cat and a mouse.

The semantics of GCIs is again defined via interpretations. We say that an interpretation $\mathcal{I}$ is a *model* of a GCI $C \sqsubseteq D$ if and only if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$. If $\mathcal{I}$ is a model of $C \sqsubseteq D$, then we shall also say that $C \sqsubseteq D$ is *valid* in $\mathcal{I}$. Intuitively, $C \sqsubseteq D$ states that every element of $\mathcal{I}$ that satisfies C also satisfies D. If $C \sqsubseteq D$ is valid in every possible interpretation we shall say that C is *subsumed* by D. This fact is commonly also denoted by $C \sqsubseteq D$ (as a statement, not an expression).

If $\mathcal{L}$ is a set of GCIs and $C \sqsubseteq D$ is another GCI, then we say that $\mathcal{L}$ *entails* $C \sqsubseteq D$ and write $\mathcal{L} \models (C \sqsubseteq D)$, if and only if for every interpretation $\mathcal{J}$ which is a model of all GCIs in $\mathcal{L}$, the interpretation $\mathcal{J}$ is also a model of $C \sqsubseteq D$.

Since we are going to extract terminological knowledge from interpretations, we can ask for the set of all $\mathcal{EL}^{\bot}_{\text{gfp}}$-GCIs for which $\mathcal{I}$ is a model. We shall denote this set $\text{Th}(\mathcal{I})$ and call it the *theory* of $\mathcal{I}$. A *base* of $\mathcal{I}$ is a set $\mathcal{B} \subseteq \text{Th}(\mathcal{I})$ such that every GCI from $\text{Th}(\mathcal{I})$ is already entailed by $\mathcal{B}$.

3.3 Formal Concept Analysis

In this section we briefly introduce some basic notions from formal concept analysis [14] that are necessary for our considerations.

The fundamental notion of formal concept analysis is the one of a *formal context*. A formal context is a triple $\mathbb{K} = (G, M, I)$ where G and M are sets and $I \subseteq G \times M$. Intuitively, we think of the set G as the set of *objects*, the set M as the set of *attributes*, and of the set I as an *incidence relation* between objects and attributes. If $g \in G$ and $m \in M$, we say that g *has* the attribute m if and only if $(g, m) \in I$.

For a set of objects $B \subseteq G$, we can ask for the set B' of common attributes of all objects in B, i. e.

$$B' = \{\, m \in M \mid \forall g \in B\colon (g, m) \in I \,\}.$$

The set B' is called the *derivation* of B in $\mathbb{K}$. Dually, we define for $A \subseteq M$ the set A' of all objects satisfying all attributes in A.

In the formal context $\mathbb{K}$ we can ask the question whether an object g that has all attributes from a set A_1 always also has all attributes from a set A_2, i. e. whether it is true that $g \in A_1'$ implies $g \in A_2'$. We can formalize this question as follows: we call the pair (A_1, A_2) with $A_1, A_2 \subseteq M$ an *implication*, usually written as $A_1 \to A_2$. We shall say that the implication $A_1 \to A_2$ is *valid* in $\mathbb{K}$ if and only if $A_1' \subseteq A_2'$. The set of all implications $A_1 \to A_2$ with $A_1, A_2 \subseteq M$ is denoted by $\mathrm{Imp}(M)$, and the set of all valid implications of $\mathbb{K}$ is called its *theory* and is denoted by $\mathrm{Th}(\mathbb{K})$.

A set $\mathcal{L} \subseteq \mathrm{Imp}(M)$ of implications *entails* an implication $A_1 \to A_2$ if and only if for all contexts in which $\mathcal{L}$ is true, the implication $A_1 \to A_2$ is true as well. A set $\mathcal{L} \subseteq \mathrm{Th}(\mathbb{K})$ is called a *base* of $\mathbb{K}$ if and only if all valid implications of $\mathbb{K}$ are entailed by $\mathcal{L}$.

It is obvious that we can extend each set $\mathcal{K} \subseteq \mathrm{Th}(\mathbb{K})$ to a base of $\mathbb{K}$. We call $\mathcal{L} \subseteq \mathrm{Th}(\mathbb{K})$ a *base with background knowledge* $\mathcal{K}$ if and only if $\mathcal{L} \cup \mathcal{K}$ is a base of $\mathbb{K}$. If $\mathcal{K} = \varnothing$, then bases with background knowledge $\mathcal{K}$ are just bases of $\mathbb{K}$.

A particularly interesting base is the so called *canonical base* $\mathrm{Can}(\mathbb{K}, \mathcal{K})$ of $\mathbb{K}$, for some given background knowledge $\mathcal{K}$. Making the definition of this base understandable is hardly possible in the given amount of space, and we refer the reader to [14] for further details. However, we still note that it is well known that $\mathrm{Can}(\mathbb{K}, \mathcal{K})$ is a base of smallest cardinality with background knowledge $\mathcal{K}$, i. e. every set of implications with less elements than $\mathrm{Can}(\mathbb{K}, \mathcal{K})$ cannot be a base of $\mathbb{K}$ with background knowledge $\mathcal{K}$.

4. AXIOMATIZING VALID $\mathcal{EL}^\bot$-GCIS OF FINITE INTERPRETATIONS

In this section we briefly introduce the main results from [12] for axiomatizing all valid $\mathcal{EL}^\bot$-GCIs of a finite interpretation. Essentially, these results are given in Theorem 4.3 for finding an optimal $\mathcal{EL}^\bot_{\mathrm{gfp}}$-base, and in Theorem 4.4, which ensures that we can effectively turn such $\mathcal{EL}^\bot_{\mathrm{gfp}}$-bases into equivalent $\mathcal{EL}^\bot$-bases.

In the following, if not stated otherwise, we shall denote with $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ a finite interpretation. The terminological knowledge we are now interested in is just the set of all valid $\mathcal{EL}^\bot$-GCIs of $\mathcal{I}$. However, it is quite easy to see that the number of valid $\mathcal{EL}^\bot$-GCIs of $\mathcal{I}$ is infinite in general, for if $C \sqsubseteq D$ is such a GCI, then $\exists r.C \sqsubseteq \exists r.D$ for $r \in N_R$ is a valid $\mathcal{EL}^\bot$-GCI as well. Therefore, we cannot simply use the set $\mathrm{Th}(\mathcal{I})$ as a TBox for an ontology. Instead, we try to find a *finite* base $\mathcal{B}$ of $\mathrm{Th}(\mathcal{I})$. Such a base would contain the same information as $\mathrm{Th}(\mathcal{I})$, and since it is finite it could be used as a TBox for an ontology. One of the main results from [12] is to prove that such bases always exists, and also to give an effective method to compute them. These results have been achieved using formal concept analysis.

The central notion that has been introduced in [12] for bringing together the description logic $\mathcal{EL}^\bot$ and formal concept analysis is the one of *model-based most-specific concept descriptions*. Roughly, for a set $X \subseteq \Delta^\mathcal{I}$ we are looking for a concept description that *describes* the individuals in X in the best way possible. More formally, we call an $\mathcal{EL}^\bot$-concept description C a *model-based most-specific concept description* for X (in $\mathcal{EL}^\bot$) if and only if

- $X \subseteq C^\mathcal{I}$ and
- for all $\mathcal{EL}^\bot$-concept descriptions D such that $X \subseteq D^\mathcal{I}$, it is true that $C \sqsubseteq D$.

It is clear that, if a model-based most-specific concept description for X exists, it is unique up to equivalence. In this case, we shall denote it with $X^\mathcal{I}$, to stress similarities to the formal concept analysis derivation operators. If C is a concept description, we shall write $C^{\mathcal{I}\mathcal{I}}$ instead of $(C^\mathcal{I})^\mathcal{I}$.

However, it may happen that model-based most-specific concept descriptions in $\mathcal{EL}^\bot$ may not exists. To see this, consider the example interpretation

$$N_C = \varnothing \qquad N_R = \{\, r \,\}$$
$$\Delta^\mathcal{I} = \{\, x \,\} \qquad r^\mathcal{I} = \{\, (x, x) \,\}.$$

Then all $\mathcal{EL}^\bot$-concept descriptions $\exists r.\exists r.\ldots\exists r.\top$ have the set $X = \{\, x \,\}$ in their extension, but there does not exist a most specific one. On the other hand, it can be seen quite easily that the $\mathcal{EL}^\bot_{\mathrm{gfp}}$-concept description

$$(A, \{\, A \equiv \exists r.A \,\})$$

is a model-based most-specific concept description of X, if we consider $\mathcal{EL}^\bot_{\mathrm{gfp}}$ instead of $\mathcal{EL}^\bot$ in the above definition. Indeed this not a coincidence, as the following result shows.

4.1 Theorem (Lemma 4.5 of [12]) *Let $\mathcal{I}$ be a finite interpretation and $X \subseteq \Delta^\mathcal{I}$. Then there exists a model-based most-specific concept description of X in $\mathcal{EL}^\bot_{\mathrm{gfp}}$.*

Because of this result we shall implicitly assume from now on that when we are talking about model-based most-specific concept descriptions, that we are actually meaning model-based most-specific concept descriptions in $\mathcal{EL}^\bot_{\mathrm{gfp}}$.

Before we continue, let us note two facts about model-base most-specific concept descriptions. Firstly, if C is an $\mathcal{EL}^\bot_{\mathrm{gfp}}$-concept description, then $C \sqsubseteq C^{\mathcal{I}\mathcal{I}}$ is always a valid GCI of $\mathcal{I}$. Furthermore, $C^{\mathcal{I}\mathcal{I}}$ is subsumed by C, again for each $\mathcal{EL}^\bot_{\mathrm{gfp}}$-concept description C. Establishing these two facts is not difficult, see [12].

Using the notion of model-based most-specific concept descriptions, we can now define a formal context $\mathbb{K}_\mathcal{I}$ which captures all relevant information on the valid $\mathcal{EL}^\bot_{\mathrm{gfp}}$-GCIs of $\mathcal{I}$. For this, we define

$$M_\mathcal{I} := \{\, \bot \,\} \cup N_C \cup \{\, \exists r.X^\mathcal{I} \mid r \in N_R, X \subseteq \Delta^\mathcal{I}, X \neq \varnothing \,\}.$$

The set $M_\mathcal{I}$ has the particular property that all model-based most-specific concept descriptions in $\mathcal{I}$ are *expressible in terms of* $M_\mathcal{I}$. For brevity, let us denote for a set $U \subseteq M_\mathcal{I}$ with $\bigsqcap U$ the concept description that is either $\top$, when U is empty, or $V_1 \sqcap \ldots \sqcap V_n$, when $U = \{\, V_1, \ldots, V_n \,\}$. Then a concept description C is expressible in terms of $M_\mathcal{I}$ if and only if there exists a set $N \subseteq M_\mathcal{I}$ such that $C \equiv \bigsqcap N$. Equivalently, C is expressible in terms of $M_\mathcal{I}$ if and only if

$$C \equiv \bigsqcap \{\, D \in M_\mathcal{I} \mid C \sqsubseteq D \,\}.$$

Having defined the set $M_{\mathcal{I}}$, we are now ready to introduce the notion of the *induced context* of $\mathcal{I}$. This is the formal context $\mathbb{K}_{\mathcal{I}} = (\Delta^{\mathcal{I}}, M_{\mathcal{I}}, \nabla)$, where for all $x \in \Delta^{\mathcal{I}}$ and $C \in M_{\mathcal{I}}$, it is true that $x \nabla C$ if and only if $x \in C^{\mathcal{I}}$.

The derivation operators in $\mathbb{K}_{\mathcal{I}}$, the interpretation function $\cdot^{\mathcal{I}}$ and model-based most-specific concept descriptions are closely related.

4.2 Proposition *Let $A \subseteq \Delta^{\mathcal{I}}, B \subseteq M_I$. Then $A^{\mathcal{I}} \equiv \bigsqcap A'$ and $B' = (\bigsqcap B)^{\mathcal{I}}$, where the derivations are conducted in $\mathbb{K}_{\mathcal{I}}$.*

With some more technical machinery it can even be shown that $(\bigsqcap A)^{\mathcal{I}\mathcal{I}} = \bigsqcap A''$ is true for each $A \subseteq M_{\mathcal{I}}$. With this we can see that the intents of $\mathbb{K}_{\mathcal{I}}$ are in a close correspondence to the model-based most-specific concept descriptions of $\mathcal{I}$. This connection also extends to the canonical base of $\mathbb{K}_{\mathcal{I}}$ and $\mathcal{EL}^{\bot}_{\text{gfp}}$-bases of $\mathcal{I}$.

4.3 Theorem (5.13 and 5.18 of [12]) *Let $\mathcal{I}$ be a finite interpretation and define*

$$S_{\mathcal{I}} = \{\, \{\, C \,\} \to \{\, D \,\} \mid C, D \in M_{\mathcal{I}}, C \sqsubseteq D \,\}.$$

Then the set

$$\mathcal{B}_{\text{Can}} := \{\, \bigsqcap U \sqsubseteq (\bigsqcap U)^{\mathcal{I}\mathcal{I}} \mid (U \to U'') \in \text{Can}(\mathbb{K}_{\mathcal{I}}, S_{\mathcal{I}}) \,\}$$

is a finite $\mathcal{EL}^{\bot}_{\text{gfp}}$-base of $\mathcal{I}$ of minimal cardinality.

Note that the set $S_{\mathcal{I}}$ contains knowledge which is trivially true in every interpretation, but not necessarily in every formal context. More precisely, if $C \sqsubseteq D$, then we do not need to state this GCI explicitly in a base. However, the corresponding implication $\{\, C \,\} \to \{\, D \,\}$ may not necessarily be valid in every formal context, and therefore bases of $\mathbb{K}$ have to contain information to entail such implications. However, we are only interested in bases of $\mathcal{I}$, which is why we add the set $S_{\mathcal{I}}$ as background knowledge.

The drawback of this result is that it only yields an $\mathcal{EL}^{\bot}_{\text{gfp}}$-base, because the usage of $\mathcal{EL}^{\bot}_{\text{gfp}}$-concept description bears two major problems when using them as terminological axioms in knowledge bases: they employ a different TBox semantics than normal knowledge bases do (i. e. greatest fixpoint semantics instead of descriptive semantics), which results in a knowledge bases with two TBox semantics, making reasoning unnecessarily complicated. Additionally, the cyclic nature of $\mathcal{EL}^{\bot}_{\text{gfp}}$-concept descriptions makes it very hard for human expert to understand them, making the maintainability of ontologies an even more complex task

Because of this it is desirable to obtain an $\mathcal{EL}^{\bot}$-base instead. Indeed, doing so is not that difficult. For this we utilize the technique of unravelling as it has been described in Section 3.1.

4.4 Theorem (5.21 of [12]) *Suppose $\mathcal{B}$ is a finite $\mathcal{EL}^{\bot}_{gfp}$-base of $\mathcal{I}$. Let $d \in \mathbb{N}$ as in Lemma 3.2 and define*

$$\begin{aligned}\mathcal{B}_{\text{unravel}} = \{\, C_d \sqsubseteq (C^{\mathcal{I}\mathcal{I}})_d \mid (C \sqsubseteq D) \in \mathcal{B} \,\} \\ \cup \{\, (X^{\mathcal{I}})_d \sqsubseteq (X^{\mathcal{I}})_{d+1} \mid X \subseteq \Delta^{\mathcal{I}}, X \neq \varnothing \,\}.\end{aligned}$$

Then $\mathcal{B}_{\text{unravel}}$ is a finite $\mathcal{EL}^{\bot}$-base of $\mathcal{I}$.

5. AXIOMATIZING $\mathcal{EL}^{\bot}$-GCIS USING CONFIDENCE

In this section we shall generalize the results we have mentioned in the previous section to the setting of *confident GCIs*. For this, we introduce this notion of *confidence* of GCIs in the next section. Thereafter, we present a way to extract $\mathcal{EL}^{\bot}_{\text{gfp}}$-bases of confident GCIs from interpretation. In the final section, we discuss how such bases can be converted into $\mathcal{EL}^{\bot}$-bases.

5.1 Confident GCIs and Bases

The notion of confidence has been used in the area of data mining to declare certain, implication-like rules as interesting if their corresponding confidence is "high," i. e. above a certain, user-defined threshold. We shall use the same idea here and introduce the notion of confidence for GCIs as follows. Let $\mathcal{I}$ be a finite interpretation and let $C \sqsubseteq D$ a ($\mathcal{EL}^{\bot}$ or $\mathcal{EL}^{\bot}_{\text{gfp}}$) GCI. We define the *confidence* of $C \sqsubseteq D$ in $\mathcal{I}$ as

$$\text{conf}_{\mathcal{I}}(C \sqsubseteq D) := \begin{cases} 1 & \text{if } C^{\mathcal{I}} = \varnothing, \\ \frac{|(C \sqcap D)^{\mathcal{I}}|}{|C^{\mathcal{I}}|} & \text{otherwise.} \end{cases}$$

In other words, the confidence of $C \sqsubseteq D$ in $\mathcal{I}$ is the relative frequency that for an individual $x \in C^{\mathcal{I}}$ it is true that $x \in D^{\mathcal{I}}$.

If $c \in [0, 1]$, then we say that $C \sqsubseteq D$ is a *confident GCI* of $\mathcal{I}$ if and only if $\text{conf}_{\mathcal{I}}(C \sqsubseteq D) \geqslant c$. The set of all confident GCIs is denoted by $\text{Th}_c(\mathcal{I})$. A set $\mathcal{B}$ of GCIs is called a *base* of $\text{Th}_c(\mathcal{I})$ if and only if $\mathcal{B}$ entails exactly those GCIs that are entailed by $\text{Th}_c(\mathcal{I})$. We call $\mathcal{B}$ a *confident base* of $\text{Th}_c(\mathcal{I})$ if and only if $\mathcal{B}$ is a base of $\text{Th}_c(\mathcal{I})$ and $\mathcal{B} \subseteq \text{Th}_c(\mathcal{I})$. Note that $\text{Th}_c(\mathcal{I})$ is not necessarily closed under entailment, so these two notions are different in general.

To illustrate the notion of confident GCIs, we shall consider an example using data from the DBpedia data set [9], a popular collection of linked data. More precisely, the interpretation $\mathcal{I}_{\text{DBpedia}}$ we want to consider here arises from the DBpedia data set by considering the child[1] and all the individuals that occur in a child relation with this data set. The interpretation $\mathcal{I}_{\text{DBpedia}}$ then contains 5262 individuals, and the base $\mathcal{B}_{\text{Can}}$ for this interpretation contains 1252 $\mathcal{EL}^{\bot}_{\text{gfp}}$-GCIs.

However, the DBpedia data set contains occasional errors, which are mostly due to the fact that its source, Wikipedia Infoboxes, are quite difficult to parse. Among others, the interpretation $\mathcal{I}_{\text{DBpedia}}$ contains the individuals `Teresa_Carpio`, `Charles_Heung`, `Adam_Cheng` and `Lydia_Shum`, which are not instances of the concept name Person, although they are certainly denoting persons. These erroneous individuals inhibit the GCI

$$\exists \mathsf{child}.\top \sqsubseteq \mathsf{Person}$$

to be valid in $\mathcal{I}_{\text{DBpedia}}$, resulting in some special "circumscriptions" of this GCI to appear in the base $\mathcal{B}_{\text{Can}}$ of $\mathcal{I}_{\text{DBpedia}}$. However, if we consider the confident GCIs of $\mathcal{I}_{\text{DBpedia}}$ for $c = 0.95$, then this GCI is extracted from $\mathcal{I}_{\text{DBpedia}}$. This shows that our approach can yield reasonable results on realistic data sets. See also [11] for more discussions on this.

5.2 A Confident $\mathcal{EL}^{\bot}_{\text{gfp}}$-Base

We now consider the question how to find confident bases of $\text{Th}_c(\mathcal{I})$ effectively, for some chosen $c \in [0, 1]$. It is apparent that valid GCIs of $\mathcal{I}$ are just those with confidence exactly

[1]Actually, we consider `http://dbpedia.org/ontology/child`, but we have shortened the name for readability

1, and that $\mathrm{Th}_1(\mathcal{I}) \subseteq \mathrm{Th}_c(\mathcal{I})$. Thus, an approach to find such a base could be to separately find bases for $\mathrm{Th}_1(\mathcal{I})$ and $\mathrm{Th}_c(\mathcal{I}) \setminus \mathrm{Th}_1(\mathcal{I})$, an idea which goes back to Luxenburger and his work on *partial implications* [15]. One of the main observation from this work, translated to our setting, is the following: if $\mathcal{B}$ is a base of $\mathcal{I}$, and $(C \sqsubseteq D) \in \mathrm{Th}_c(\mathcal{I}) \setminus \mathrm{Th}_1(\mathcal{I})$, then

$$\mathcal{B} \cup \{\, C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}} \,\} \models (C \sqsubseteq D),$$

simply because $\mathcal{B} \models (C \sqsubseteq C^{\mathcal{I}\mathcal{I}})$ and $\varnothing \models (D^{\mathcal{I}\mathcal{I}} \sqsubseteq D)$. Thus, let us define

$$\mathrm{Conf}(\mathcal{I}, c) := \{\, C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}} \mid C, D \ \mathcal{EL}^{\bot}_{\mathrm{gfp}}\text{-concepts}, \\ \mathrm{conf}_{\mathcal{I}}(C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}}) \in [c, 1) \,\},$$

where we only consider GCIs up to equivalence. We can then obtain the following result.

5.1 Theorem *Let $\mathcal{I}$ be a finite interpretation, $c \in [0, 1]$ and $\mathcal{B}$ be a base of $\mathcal{I}$. Then $\mathcal{B} \cup \mathrm{Conf}(\mathcal{I}, c)$ is a finite, confident base of $\mathrm{Th}_c(\mathcal{I})$.*

Proof It is clear that $\mathcal{B} \cup \mathrm{Conf}(\mathcal{I}, c) \subseteq \mathrm{Th}_c(\mathcal{I})$. Furthermore, it has been shown in [12] that there exist only finitely many non-equivalent model-based most-specific concept descriptions of $\mathcal{I}$. It therefore remains to show that every GCI in $\mathrm{Th}_c(\mathcal{I}) \setminus \mathrm{Th}_1(\mathcal{I})$ is entailed by $\mathcal{B} \cup \mathrm{Conf}(\mathcal{I}, c)$.

Let $(C \sqsubseteq D) \in \mathrm{Th}_c(\mathcal{I}) \setminus \mathrm{Th}_1(\mathcal{I})$ be such a GCI. It is true that $\mathcal{B} \models (C \sqsubseteq C^{\mathcal{I}\mathcal{I}})$. Furthermore, an easy calculation shows that $\mathrm{conf}_{\mathcal{I}}(C \sqsubseteq D) = \mathrm{conf}_{\mathcal{I}}(C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}})$, hence $(C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}}) \in \mathrm{Conf}(\mathcal{I}, c)$ up to equivalence. Finally, $\varnothing \models (D \sqsubseteq D^{\mathcal{I}\mathcal{I}})$, and thus

$$\mathcal{B} \cup \mathrm{Conf}(\mathcal{I}, c) \models C \sqsubseteq C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}} \sqsubseteq D. \qquad \square$$

From the proof of the previous theorem we can easily infer that we can weaken the prerequisites in the following way.

5.2 Corollary *Let $\mathcal{C} \subseteq \mathrm{Conf}(\mathcal{I}, c)$ be such that all GCIs in $\mathrm{Conf}(\mathcal{I}, c)$ are already entailed by $\mathcal{C}$, and let $\mathcal{B} \subseteq \mathrm{Th}_1(\mathcal{I})$ be such that $\mathcal{B} \cup \mathcal{C}$ entails every valid $\mathcal{EL}^{\bot}_{\mathrm{gfp}}$-GCI of $\mathcal{I}$. Then $\mathcal{B} \cup \mathcal{C}$ is a finite, confident base of $\mathrm{Th}_c(\mathcal{I})$.*

Using Theorem 5.1, we have obtained an effective method to compute confident bases of $\mathrm{Th}_c(\mathcal{I})$. In fact, the computation of all model-based most-specific concept descriptions, which is needed to compute both $M_{\mathcal{I}}$ and $\mathrm{Conf}(\mathcal{I}, c)$, can be achieved using the NextClosure algorithm [13]. This is mostly due to the fact that the mapping $X \mapsto X^{\mathcal{I}\mathcal{I}}$ for $X \subseteq \Delta^{\mathcal{I}}$ is a closure operator. See also [12] for more details on this.

However, although this approach is effective, it may not be very efficient. In particular, the computation of $\mathrm{Conf}(\mathcal{I}, c)$ requires, if done naively, the computation of the confidence of $C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}}$ in $\mathcal{I}$. The computation of the confidence in $\mathcal{I}$ itself might be costly, if the interpretation $\mathcal{I}$ is too large. Thus, computing the set $\mathrm{Conf}(\mathcal{I}, c)$ may be too expensive, as the interpretation $\mathcal{I}$ would have to be accessed too often.

To handle this situation it may be worthwhile to transform the task of finding a confident base of $\mathrm{Th}_c(\mathcal{I})$ into a task purely formulated in formal concept analysis. The reason for this is that formal concept analysis is closely related to the field of data mining [20], and a translation of our original problem into a problem of formal concept analysis might allow us to use algorithms from data mining for our task. Furthermore, as accessing databases is expensive, those algorithms are designed to access the database as less as possible.

To make use of this tight connection between data mining and formal concept analysis, we are going to show the following claim: given a finite interpretation $\mathcal{I}$ and $c \in [0, 1]$, it suffices to find a *confident base* $\mathrm{Th}_c(\mathbb{K}_{\mathcal{I}})$, i. e. a set of implications with confidence at least c in $\mathbb{K}_{\mathcal{I}}$ such that every implication with confidence at least c in $\mathbb{K}_{\mathcal{I}}$ is already entailed by it. As we shall see, such a base can easily be transformed into a confident base of $\mathrm{Th}_c(\mathcal{I})$. Using this approach, we can utilize data mining algorithms to extract confident bases of $\mathrm{Th}_c(\mathbb{K}_{\mathcal{I}})$ to obtain confident bases of $\mathrm{Th}_c(\mathcal{I})$.

One of the crucial observations for the following considerations involves a connection between entailment of GCIs and entailment of implications. For brevity, let us define for a set M of concept descriptions and $\mathcal{L} \subseteq \mathrm{Imp}(M)$

$$\bigsqcap \mathcal{L} = \{\, \bigsqcap X \sqsubseteq \bigsqcap Y \mid (X \to Y) \in \mathcal{L} \,\}.$$

Then the following statement is true.

5.3 Lemma *Let M be a set of concept descriptions and let $\mathcal{L} \subseteq \mathrm{Imp}(M), (X \to Y) \in \mathrm{Imp}(M)$. Then $\mathcal{L} \models (X \to Y)$ implies $\bigsqcap \mathcal{L} \models (\bigsqcap X \sqsubseteq \bigsqcap Y)$.*

Proof Let $\mathcal{J} = (\Delta^{\mathcal{J}}, \cdot^{\mathcal{J}})$ be an interpretation such that $\mathcal{J} \models \bigsqcap \mathcal{L}$. Let us define a formal context $\mathbb{K}_{\mathcal{J},M} = (\Delta^{\mathcal{J}}, M, \nabla)$ via

$$x \nabla C \iff x \in C^{\mathcal{J}}$$

for all $x \in \Delta^{\mathcal{J}}, C \in M$.

We shall show now that $\mathbb{K}_{\mathcal{J},M} \models \mathcal{L}$. Let $(E \to F) \in \mathcal{L}$. Then $(\bigsqcap E)^{\mathcal{J}} \subseteq (\bigsqcap F)^{\mathcal{J}}$, since $\mathcal{J} \models \bigsqcap \mathcal{L}$. It is not hard to see that $(\bigsqcap E)^{\mathcal{J}} = E'$, where the derivation has been done in $\mathbb{K}_{\mathcal{J},M}$. Therefore, it is true that $E' \subseteq F'$, and thus $\mathbb{K}_{\mathcal{J},M} \models (E \to F)$.

Since $\mathcal{L} \models (X \to Y)$, it is true that $\mathbb{K}_{\mathcal{J},M} \models (X \to Y)$, i. e. $X' \subseteq Y'$. As $(\bigsqcap X)^{\mathcal{J}} = X'$, it is thus true that $(\bigsqcap X)^{\mathcal{J}} \subseteq (\bigsqcap Y)^{\mathcal{J}}$, i. e. $\bigsqcap \mathcal{L} \models (\bigsqcap X \sqsubseteq \bigsqcap Y)$. $\square$

Note that we cannot expect the converse of this lemma to be true in general, as it is illustrated by the following example.

5.4 Example Consider the sets $N_C = \{\, \mathsf{A}, \mathsf{B} \,\}, N_R = \{\, \mathsf{r} \,\}$ and $M = \{\, \mathsf{A}, \mathsf{B}, \exists \mathsf{r}.\mathsf{A}, \exists \mathsf{r}.\mathsf{B} \,\}$. Define $\mathcal{L} = \{\, \{\, \mathsf{A} \,\} \to \{\, \mathsf{B} \,\} \,\}, X = \{\, \exists \mathsf{r}.\mathsf{A} \,\}, Y = \{\, \exists \mathsf{r}.\mathsf{B} \,\}$. Then clearly $\mathcal{L} \not\models (X \to Y)$, but since $\bigsqcap \mathcal{L} \models (\mathsf{A} \sqsubseteq \mathsf{B})$, it is true that $\bigsqcap \mathcal{L} \models (\bigsqcap X \sqsubseteq \bigsqcap Y)$. $\diamondsuit$

We shall now show the main result of this section, which will allow us to obtain confident bases of $\mathrm{Th}_c(\mathcal{I})$ from *confident bases* of $\mathrm{Th}_c(\mathbb{K}_{\mathcal{I}})$. For this let us introduce the notion of confidence for implications, i. e.

$$\mathrm{conf}_{\mathbb{K}}(X \to Y) := \begin{cases} 1 & \text{if } X' = \varnothing \\ \frac{|(X \cup Y)'|}{|X'|} & \text{otherwise} \end{cases}$$

for some formal context $\mathbb{K} = (G, M, I)$ and $(X \to Y) \in \mathrm{Imp}(M)$. If $c \in [0, 1]$, then the set of $\mathrm{Th}_c(\mathbb{K})$ denotes the set of all implications whose confidence in $\mathbb{K}$ is at least c. Furthermore, the notions of *base* and *confident base* are defined in the obvious way.

5.5 Theorem *Let $\mathcal{I}$ be a finite interpretation and let $c \in [0, 1]$. Let $\mathcal{L}$ be a confident base of $\mathrm{Th}_c(\mathbb{K}_{\mathcal{I}})$. Then $\bigsqcap \mathcal{L}$ is a confident base of $\mathrm{Th}_c(\mathcal{I})$.*

Proof (Sketch) It is easy to see that $\bigcap \mathcal{L} \subseteq \mathrm{Th}_c(\mathcal{I})$, and we shall not show this here. Instead, we show that every confident GCI $(C \sqsubseteq D) \in \mathrm{Th}_c(\mathcal{I})$ already follows from $\bigcap \mathcal{L}$. To this end we shall show that

i. $\bigcap \mathcal{L} \models (\bigcap U \sqsubseteq (\bigcap U)^{\mathcal{I}\mathcal{I}})$ for all $U \subseteq M_{\mathcal{I}}, U \neq \varnothing$. This in particular then implies that $\bigcap \mathcal{L} \models \mathcal{B}_{\mathrm{Can}}$.

ii. $\bigcap \mathcal{L} \models \mathrm{Conf}(\mathcal{I}, c)$.

If we establish the validity of these two claims, then $\bigcap \mathcal{L} \models \mathcal{B}_{\mathrm{Can}} \cup \mathrm{Conf}(\mathcal{I}, c)$. By Theorem 5.1, $\mathcal{B}_{\mathrm{Can}} \cup \mathrm{Conf}(\mathcal{I}, c)$ entails all GCIs in $\mathrm{Th}_c(\mathcal{I})$, and thus so does $\bigcap \mathcal{L}$.

For the first case let $U \subseteq M_{\mathcal{I}}$. Since $\mathcal{L}$ entails all implications from $\mathrm{Th}_c(\mathbb{K}_{\mathcal{I}})$, it entails all valid implications, hence

$$\mathcal{L} \models (U \to U'').$$

By Lemma 5.3, we obtain from this $\bigcap \mathcal{L} \models (\bigcap U \sqsubseteq \bigcap U'')$. Since $\bigcap U'' \equiv (\bigcap U)^{\mathcal{I}\mathcal{I}}$, this shows $\bigcap \mathcal{L} \models (\bigcap U \sqsubseteq (\bigcap U)^{\mathcal{I}\mathcal{I}})$ as required.

For the second case, let $(C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}}) \in \mathrm{Conf}(\mathcal{I}, c)$. Define $U := C^{\mathcal{I}}, V := D^{\mathcal{I}}$. It can be shown that $V^{\mathcal{I}} \equiv \bigcap V', U^{\mathcal{I}} = \bigcap U'$, thus

$$\bigcap \mathcal{L} \models (U^{\mathcal{I}} \sqsubseteq V^{\mathcal{I}}) \iff \bigcap \mathcal{L} \models (\bigcap U' \sqsubseteq \bigcap V').$$

It is straight-forward to verify that $\mathrm{conf}_{\mathcal{I}}(\bigcap U' \sqsubseteq \bigcap V') = \mathrm{conf}_{\mathbb{K}_{\mathcal{I}}}(U' \to V')$. Since $\mathcal{L}$ entails all implications from $\mathrm{Th}_c(\mathbb{K}_{\mathcal{I}})$, it is true that $\mathcal{L} \models (U' \to V')$. Lemma 5.3 yields $\bigcap \mathcal{L} \models (\bigcap U' \sqsubseteq \bigcap V')$, hence $\bigcap \mathcal{L} \models (U^{\mathcal{I}} \sqsubseteq V^{\mathcal{I}})$ as required. ▫

Of course, we can also add the background knowledge $S_{\mathcal{I}}$ to our construction, as it has been done in Theorem 4.3.

5.6 Corollary *Let $\mathcal{I}$ be a finite interpretation, $c \in [0, 1]$ and $\mathcal{L} \subseteq \mathrm{Imp}(M_{\mathcal{I}})$ be such that $\mathcal{L} \cup S_{\mathcal{I}}$ is a confident base of $\mathrm{Th}_c(\mathbb{K}_{\mathcal{I}})$. Then $\bigcap \mathcal{L}$ is a confident base of $\mathrm{Th}_c(\mathcal{I})$.*

5.3 Unravelling $\mathcal{EL}^{\perp}_{\mathbf{gfp}}$-Bases into $\mathcal{EL}^{\perp}$-Bases

We have seen in the previous section how we can obtain confident $\mathcal{EL}^{\perp}_{\mathrm{gfp}}$-bases of $\mathrm{Th}_c(\mathcal{I})$, and we now want to address the problem of how to transform this base into an $\mathcal{EL}^{\perp}$-base. To this end, we shall consider in this section the question how we can extend the result of Theorem 4.4 to our general setting of confident GCIs. In other words, we shall see how we can generalize this result to obtain an "unravelling" of confident $\mathcal{EL}^{\perp}_{\mathrm{gfp}}$-bases. The argumentation of this section is a generalization of the corresponding proof from [12].

Let us consider Theorem 4.4 again in more detail. There it is stated that if $\mathcal{B}$ is a finite $\mathcal{EL}^{\perp}_{\mathrm{gfp}}$-base, then the set

$$\begin{aligned}\mathcal{B}_{\mathrm{unravel}} = &\{\, C_d \sqsubseteq (C^{\mathcal{I}\mathcal{I}})_d \mid (C \sqsubseteq D) \in \mathcal{B} \,\} \\ &\cup \{\, (X^{\mathcal{I}})_d \sqsubseteq (X^{\mathcal{I}})_{d+1} \mid X \subseteq \Delta^{\mathcal{I}}, X \neq \varnothing \,\},\end{aligned}$$

where d is chosen as in Lemma 3.2.

Intuitively, we can think of the second set in the definition of $\mathcal{B}_{\mathrm{unravel}}$ as the part which "compensates" for our unravelling of the cyclic concept descriptions. To make this more precise, let us define

$$\mathcal{X}_{\mathcal{I}} = \{\, (X^{\mathcal{I}})_d \sqsubseteq (X^{\mathcal{I}})_{d+1} \mid X \subseteq \Delta^{\mathcal{I}}, X \neq \varnothing \,\}.$$

Then the following statement is true.

5.7 Lemma *For each $Y \subseteq \Delta^{\mathcal{I}}$ and $k \geqslant d$, it is true that*

i. $\mathcal{X}_{\mathcal{I}} \models ((Y^{\mathcal{I}})_k \sqsubseteq (Y^{\mathcal{I}})_{k+1})$,

ii. $\mathcal{X}_{\mathcal{I}} \models ((Y^{\mathcal{I}})_k \sqsubseteq Y^{\mathcal{I}})$.

The set $\mathcal{X}_{\mathcal{I}}$ captures the necessary entailments we need to make our unravelling work. More precisely, let us assume that we have given a base $\mathcal{B} \cup \mathcal{C}$ of $\mathrm{Th}_c(\mathcal{I})$ such that $\mathcal{B}$ contains only valid GCIs and $\mathcal{C} \cap \mathrm{Th}_1(\mathcal{I}) = \varnothing$. We furthermore assume that $\mathcal{B}$ contains only GCIs of the form $E \sqsubseteq E^{\mathcal{I}\mathcal{I}}$. This is not a restriction of the general case, as it can be shown that

$$\{\, E \sqsubseteq E^{\mathcal{I}\mathcal{I}} \,\} \models (E \sqsubseteq F)$$

for every valid GCI $E \sqsubseteq F$ of $\mathcal{I}$. Let $d \in \mathbb{N}$ as in Lemma 3.2. We can then define the unravelling of $\mathcal{B} \cup \mathcal{C}$ as follows:

$$\begin{aligned}\mathcal{B}_0 = &\{\, E_d \sqsubseteq (E^{\mathcal{I}\mathcal{I}})_d \mid (E \sqsubseteq E^{\mathcal{I}\mathcal{I}}) \in \mathcal{B} \,\} \\ &\cup \{\, C_d \sqsubseteq (C^{\mathcal{I}\mathcal{I}})_d \mid (C \sqsubseteq D) \in \mathcal{C} \,\} \\ \mathcal{C}_0 = &\{\, (C^{\mathcal{I}\mathcal{I}})_d \sqsubseteq (D^{\mathcal{I}\mathcal{I}})_d \mid (C \sqsubseteq D) \in \mathcal{C} \,\}.\end{aligned}$$

Note that $\mathcal{B}_0 \cup \mathcal{C}_0$ contains only $\mathcal{EL}^{\perp}$-GCIs. Additionally, $\mathcal{B}_0$ consists of valid GCIs of $\mathcal{I}$ only.

5.8 Theorem *Let $\mathcal{I}$ be a finite interpretation, $c \in [0, 1]$ and let $d \in \mathbb{N}$ be as in Lemma 3.2. Let $\mathcal{B} \cup \mathcal{C}$ be a confident base of $\mathrm{Th}_c(\mathcal{I})$ such that $\mathcal{B} \subseteq \mathrm{Th}_1(\mathcal{I}), C \cap \mathrm{Th}_1(\mathcal{I}) = \varnothing$ and $\mathcal{B}$ only contains GCIs of the form $E \sqsubseteq E^{\mathcal{I}\mathcal{I}}$. Define $\mathcal{B}_0$ and $\mathcal{C}_0$ as before. Then*

i. $\mathcal{C}_0 \subseteq \mathrm{Th}_c(\mathcal{I})$ *and* $\mathcal{B}_0 \cup \mathcal{C}_0 \cup \mathcal{X}_{\mathcal{I}} \models \mathcal{C}$;

ii. $\mathcal{B}_0 \cup \mathcal{C}_0 \cup \mathcal{X}_{\mathcal{I}} \models \mathcal{B}_0$.

In particular, the set $\mathcal{B}_0 \cup \mathcal{C}_0 \cup \mathcal{X}_{\mathcal{I}}$ is a confident $\mathcal{EL}^{\perp}$-base of $\mathrm{Th}_c(\mathcal{I})$.

6. CONCLUSIONS AND FURTHER WORK

Starting from the initial work by Baader and Distel on finding bases of finite interpretation $\mathcal{I}$, we have extended this work in the direction of considering confident GCI instead of valid GCIs only. The idea behind this approach is to circumvent occasional errors in the initial interpretation $\mathcal{I}$, which may inhibit otherwise interesting GCIs to be extracted by the original approach.

The extension we have presented in this paper is twofold. Firstly, we have described two methods how a base of $\mathrm{Th}_c(\mathcal{I})$ can effectively be computed. The crucial ideas in this direction where based on previous work on partial implications in the field of formal concept analysis. The first possibility to obtain such a base was a simple translation of these ideas into our setting of description logics. The second method went further and showed that extracting confident bases of $\mathrm{Th}_c(\mathbb{K}_{\mathcal{I}})$ always yields confident bases of $\mathcal{I}$. This is particularly interesting as it allows algorithms from data mining to assist in constructing the desired base.

However, all these results only yielded $\mathcal{EL}^{\perp}_{\mathrm{gfp}}$-bases of $\mathrm{Th}_c(\mathcal{I})$. We therefore have discussed a way to transform given $\mathcal{EL}^{\perp}_{\mathrm{gfp}}$-bases of $\mathrm{Th}_c(\mathcal{I})$ into $\mathcal{EL}^{\perp}$-bases of $\mathrm{Th}_c(\mathcal{I})$. For this, we have generalized ideas from the original approach to our setting of confident GCIs.

It should be obvious that to make considering confident GCIs a practicably useful approach that further research is necessary. The main reason for this is that considering

confident GCIs is an inherently heuristic approach, and the results always have to be subject to further processing. On the other hand, if we want to deal with errors, which appear mostly randomly, such heuristic methods are the best one can hope for.

A particular problem that arises from confident GCIs is the one of *rare counterexamples*. A classical example for this is the knowledge that all birds fly *except for penguins*. If we would consider a data set from which we want to learn something about birds, and if we consider confident GCIs, then it is highly likely that we learn that all birds fly, just because penguins do not occur often enough. In other words, penguins are treated as errors although they are not.

The fundamental problem one has to solve here is then to tell the difference between errors and rare counterexamples. Of course, such a distinction has to be made outside of the original source, i. e. extra information is needed to decide this. Such external information could be provided by (human) experts, which possess valid (implicit) knowledge about the domain of interest. Following this approach it would be most interesting to generalize the algorithm of *attribute exploration* from formal concept analysis to confident GCIs. This algorithm has been designed to assist experts in making their implicit knowledge explicit. A generalized version of attribute exploration for confident GCIs would then allow an expert to tell errors from valid counterexamples, and could even assist in systematically fixing errors in the original data set.

7. REFERENCES

[1] Agrawal, R., Imielinski, T., and Swami, A. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (May 1993), pp. 207–216.

[2] Baader, F. Least common subsumers, most specific concepts, and role-value-maps in a description logic with existential restrictions and terminological cycles. LTCS-Report LTCS-02-07, Chair for Automata Theory, Institute for Theoretical Computer Science, Dresden University of Technology, Germany, 2002. See `http://lat.inf.tu-dresden.de/research/reports.html`.

[3] Baader, F. Terminological cycles in a description logic with existential restrictions. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence* (August 2003), G. Gottlob and T. Walsh, Eds., Morgan Kaufmann, pp. 325–330.

[4] Baader, F., and Distel, F. A Finite Basis for the Set of $\mathcal{EL}$-Implications Holding in a Finite Model. In *Proceedings of the 6th International Conference on Formal Concept Analysis* (February 2008), R. Medina and S. A. Obiedkov, Eds., vol. 4933 of *Lecture Notes in Computer Science*, Springer, pp. 46–61.

[5] Baader, F., and Distel, F. Exploring Finite Models in the Description Logic $\mathcal{EL}_{\text{gfp}}$. In *Proceedings of the 7th International Conference on Formal Concept Analysis* (May 2009), S. Ferré and S. Rudolph, Eds., vol. 5548 of *Lecture Notes in Computer Science*, Springer, pp. 146–161.

[6] Baader, F., Ganter, B., Sattler, U., and Sertkaya, B. Completing description logic knowledge bases using formal concept analysis. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence* (2007), AAAI Press, pp. 230–235.

[7] Baader, F., and Peñaloza, R. Axiom pinpointing in general tableaux. In *TABLEAUX* (2007), N. Olivetti, Ed., vol. 4548 of *Lecture Notes in Computer Science*, Springer, pp. 11–27.

[8] Bizer, C., Heath, T., and Berners-Lee, T. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS) 5*, 3 (March 2009), 1–22.

[9] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. DBpedia - A Crystallization Point of the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web 7*, 3 (September 2009), 154–165.

[10] Borchmann, D. Axiomatizing Confident $\mathcal{EL}^{\perp}_{\text{gfp}}$-GCIs of Finite Interpretations. Report MATH-AL-08-2012, Chair of Algebraic Structure Theory, Institute of Algebra, Technische Universität Dresden, Dresden, Germany, September 2012.

[11] Borchmann, D. On Confident GCIs of Finite Interpretations. LTCS-Report 12-06, Institute for Theoretical Computer Science, TU Dresden, Dresden, 2012. See `http://lat.inf.tu-dresden.de/research/reports.html`.

[12] Distel, F. *Learning Description Logic Knowledge Bases from Data Using Methods from Formal Concept Analysis*. PhD thesis, TU Dresden, 2011.

[13] Ganter, B. Two basic algorithms in concept analysis. In *Proceedings of the 8th Interational Conference of Formal Concept Analysis* (March 2010), L. Kwuida and B. Sertkaya, Eds., vol. 5986 of *Lecture Notes in Computer Science*, Springer, pp. 312–340.

[14] Ganter, B., and Wille, R. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin-Heidelberg, 1999.

[15] Luxenburger, M. *Implikationen, Abhängigkeiten und Galois-Abbildungen*. PhD thesis, TH Darmstadt, 1993.

[16] Nebel, B. Terminological Cycles: Semantics and Computational Properties. In *Principles of Semantic Networks* (1991), Morgan Kaufmann, pp. 331–362.

[17] Price, C., and Spackman, K. SNOMED Clinical Terms. *British Journal of Healtcare Computing and Information Management 17* (2000), 27–31.

[18] Stumme, G. Efficient data mining based on formal concept analysis. In *Database and Expert Systems Applications* (Heidelberg, 2002), A. Hameurlain, R. Cicchetti, and R. Traunmüller, Eds., vol. 2453 of *LNCS*, Springer, pp. 534–546.

[19] Völker, J., and Niepert, M. Statistical schema induction. In *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, Proceedings, Part I* (May–June 2011), G. Antoniou, M. Grobelnik, E. P. B. Simperl, B. Parsia, D. Plexousakis, P. D. Leenheer, and J. Z. Pan, Eds., vol. 6643 of *Lecture Notes in Computer Science*, Springer, pp. 124–138.

[20] Zaki, M. J., and Ogihara, M. Theoretical foundation of association rules. In *SIGMOD'98 Workshop on Research Issues in Data Mining and Knowledge Discovery (SIGMOD-DMKD'98)* (June 1998), pp. 1–8.