Towards an Error-Tolerant Construction of $\mathcal{EL}^{\bot}$-Ontologies from Data using Formal Concept Analysis

Daniel Borchmann*

TU Dresden

Abstract. In the work of Baader and Distel, a method has been proposed to axiomatize all general concept inclusions (GCIs) expressible in the description logic $\mathcal{EL}^{\bot}$ and valid in a given interpretation $\mathcal{I}$. This provides us with an effective method to learn $\mathcal{EL}^{\bot}$-ontologies from interpretations. In this work, we want to extend this approach in the direction of handling *errors*, which might be present in the data-set. We shall do so by not only considering *valid* GCIs but also those whose *confidence* is above a given threshold c. We shall give the necessary definitions and show some first results on the axiomatization of all GCIs with confidence at least c. Finally, we shall provide some experimental evidence based on real-world data that supports our approach.

Keywords: Formal Concept Analysis, Description Logics, Ontology Learning

1 Introduction

Description logic ontologies provide a practical yet formally well-defined way of representing large amounts of knowledge. They have been applied especially successfully in the area of medical and biological knowledge, one example being SNOMED CT [13], a medical ontology used to standardize medical nomenclature.

A part of description logic ontologies, the so called *TBox*, contains the *terminological knowledge* of the ontology. Terminological knowledge constitutes connections between *concept descriptions* and is represented by *general concept inclusions* (GCIs). For example, we could fix in an ontology the fact that everything that has a child is actually a person. Using the description logic $\mathcal{EL}^{\bot}$, this could be written as

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

Here, $\exists \mathsf{child}.\top$ and Person are examples of concept descriptions, and the $\sqsubseteq$ sign can be read as "implies." General concept inclusions are, on this intuitive level, therefore quite similar to implications.

The construction of TBoxes of ontologies, which are supposed to represent the knowledge of a certain domain of interest, is normally conducted by human experts.

* Supported by DFG Graduiertenkolleg 1763 (QuantLA)

Although this guarantees a high level of quality of the resulting ontology, the process itself is long and expensive. Automating this process would both decrease the time and cost for creating ontologies and would therefore foster the use of formal ontologies in other applications. However, one cannot expect to entirely replace human experts in the process of creating domain-specific ontologies, as these experts are the original source of this knowledge. Hence constructing ontologies completely automatically does not seem reasonable.

A compromise for this would be to devise a *semi-automatic* way of constructing ontologies, for example by *learning* relevant parts of the ontology from a set of *typical examples* of the domain of interest. The resulting ontologies could be used by ontology engineers as a starting point for further development.

This approach has been taken by Baader and Distel [8,2] for constructing $\mathcal{EL}^\bot$-ontologies from finite interpretations. The reason why this approach is restricted to $\mathcal{EL}^\bot$ is manifold. Foremost, this approach exploits a tight connection between the description logic $\mathcal{EL}^\bot$ and *formal concept analysis* [9], and such a connection has not been worked out for other description logics. Moreover, the description logic $\mathcal{EL}^\bot$ can be sufficient for practical applications, as, for example, SNOMED CT is formulated in a variant of $\mathcal{EL}^\bot$. Lastly, $\mathcal{EL}^\bot$ is computationally much less complex than other description logics, say $\mathcal{ALC}$ or even $\mathcal{FL}_0$.

In their approach, Baader and Distel are able to effectively construct a *base* of all valid GCIs of a given *interpretation*, where this interpretation can be understood as the collection of typical examples of our domain of interest. This base therefore constitutes the complete terminological knowledge that is valid in this interpretation. Moreover, these interpretations can be seen as a different way to represent *linked data* [3], the data format used by the semantic web community to store its data. Hence, this approach allows us to construct ontologies from parts of the linked data cloud, providing us with a vast amount of real-world data for experiments and practical applications.

In [7], a sample construction has been conducted on a small part of the DBpedia data set [4], which is part of the linked open data cloud. As it turned out, the approach is effective. However, another result of these experiments was the following observation: in the data set extracted from DBpedia, a small set of errors were present. These errors, although very few, greatly influenced the result of the construction, in the way these errors invalidated certain GCIs, and hence these GCIs were not extracted by the algorithm anymore. Then, instead of these general GCIs, more special GCIs were extracted that "circumvent" these errors by being more specific. This not only lead to more extracted GCIs, but also to GCIs which may be hard to comprehend.

As the original approach by Baader and Distel considers only valid GCIs, even a single error may invalidate an otherwise valid GCI. Since we cannot assume from real-world data that it does not contain any errors, this approach is quite limited for practical applications. Therefore, we want to present in this work a generalization to the approach of Baader and Distel which does not only consider valid GCIs but also those which are "almost valid." The rationale behind this is that these GCIs should be much less sensitive to a small amount of errors than valid GCIs. To decide whether

a GCIs is "almost valid," we shall use its *confidence* in the given interpretation. We then consider the set of all GCIs of a finite interpretation whose confidence is above a certain threshold $c \in [0, 1]$, and try to find a base for them. This base can then be seen as the terminological part of an ontology learned from the data set.

This paper is structured as follows. Firstly, we shall introduce some relevant notions of formal concept analysis and description logics in the following section. In this, we shall also review some of the basic definitions of [8] we are going to need for our discussions. After this, we describe our experiment with the DBpedia data set in more detail and introduce the notion of *confidence* for general concept inclusions. Then we discuss ideas and present first results on how to find bases for the GCIs whose confidence is above a certain threshold. Finally, we shall revisit our experiment with the DBpedia data set and examine in how far the approach of considering confident GCIs was helpful (for this particular experiment).

2 Preliminaries

The purpose of this section is to recall and introduce some of the basic notions needed in this paper. For this, we shall firstly consider relevant parts of formal concept analysis. After this, we introduce the description logic $\mathcal{EL}^\bot$, interpretations and general concept inclusions.

Please note that the sole purpose of this section is to provide these definitions for use in this paper. For thorough treatments of these topics, we refer the reader to [9] for an introduction to formal concept analysis and [1] for an introduction to description logics.

2.1 Formal Concept Analysis

Formal concept analysis studies the relationships between properties of formal contexts and properties of their associated concept lattices. A *formal context* is a triple $\mathbb{K} = (G, M, I)$ of sets such that $I \subseteq G \times M$. The elements $g \in G$ are called *objects*, the elements $m \in M$ are called *attributes* and an object g is said to *have* an attribute m if and only if $(g, m) \in I$. We may also write $g\ I\ m$ instead of $(g, m) \in I$.

For a set $A \subseteq G$ of objects, we can ask for the set of *common attributes* of A, i. e. the set of all attributes in M that all objects in A share. Formally, we denote this set as A' and define it as follows:

$$A' := \{\, m \in M \mid \forall g \in G\colon g\ I\ m \,\}.$$

Likewise, for a set $B \subseteq M$ of attributes, we denote with B' the set of all objects shared by all attributes in B (*common objects* of B), formally

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

We write A'' instead of $(A')'$ and B'' instead of $(B')'$.

An *implication* $X \to Y$ is just a pair (X, Y) such that $X, Y \subseteq M$. The implication $X \to Y$ *holds* in $\mathbb{K}$, written $\mathbb{K} \models (X \to Y)$, if and only if $X' \subseteq Y'$, or

equivalently $Y \subseteq X''$. Therefore, $X \to Y$ holds in $\mathbb{K}$ if and only if whenever an object has all attributes from X, it also has all attributes from Y as well. This also explains the name "implication."

Let $\mathcal{L}$ be a set of implications. A set $A \subseteq M$ is said to be *closed* under $\mathcal{L}$ if and only if for each $(X \to Y) \in \mathcal{L}$ it is true that $X \nsubseteq A$ or $Y \subseteq A$. We denote with $\mathcal{L}(A)$ the $\subseteq$-smallest superset of A that is closed under $\mathcal{L}$. Such a set always exists.

$\mathcal{L}$ is said to be *sound* for $\mathbb{K}$ if and only if each implication in $\mathcal{L}$ holds in $\mathbb{K}$. Furthermore, $\mathcal{L}$ is said to be *complete* for $\mathbb{K}$ if and only if every valid implication of $\mathbb{K}$ already follows from $\mathcal{L}$. Thereby, an implication $X \to Y$ *follows* from $\mathcal{L}$ if and only if for each formal context $\mathbb{L}$ in which all implications $\mathcal{L}$ hold, the implication $X \to Y$ holds as well. This is the case if and only if $Y \subseteq \mathcal{L}(X)$ and we write $\mathcal{L} \models (X \to Y)$ in this case. Finally, $\mathcal{L}$ is called a *base* of $\mathbb{K}$ if and only if $\mathcal{L}$ is sound and complete for $\mathbb{K}$.

2.2 The Description Logic $\mathcal{EL}^\bot$

Description logics are formal languages, whose purpose is to represent knowledge and to provide methods to effectively reason about this knowledge. Thereby, description logics come in different flavors of expressibility and computational complexity. The logic we are mainly interested in in this work is $\mathcal{EL}^\bot$, which we shall introduce now.

Let us fix two countably (finite or infinite) and disjoint sets N_C and N_R, denoting *concept names* and *role names*, respectively. Then, an $\mathcal{EL}$-concept description C is either of the form $C = A$ for $A \in N_C$, $C = \top$, $C = C_1 \sqcap C_2$ or $C = \exists r.C_1$, for C_1, C_2 being $\mathcal{EL}$-concept descriptions and $r \in N_R$. C is an $\mathcal{EL}^\bot$-concept description if $C = \bot$ or C is an $\mathcal{EL}$-concept description.

As an example, suppose that Person, Male are concept names and child is a role name. Then an example for an $\mathcal{EL}^\bot$-concept description is

$$\mathsf{Person} \sqcap \mathsf{Male} \sqcap \exists \mathsf{child}.\top$$

denoting a male person who has children, i. e. a father.

Semantics of $\mathcal{EL}^\bot$-concept descriptions are defined through *interpretations*. An interpretation $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ consists of a set $\Delta^\mathcal{I}$ of *elements* and a mapping $\cdot^\mathcal{I}$ such that $A^\mathcal{I} \subseteq \Delta^\mathcal{I}$ for all $A \in N_C$ and $r^\mathcal{I} \subseteq \Delta^\mathcal{I}\Delta^\mathcal{I}$ for all $r \in N_R$. We can inductively extend this mapping to the set of all $\mathcal{EL}^\bot$-concept description using the following rules, where C_1, C_2 are again $\mathcal{EL}$-concept descriptions and $r \in N_R$:

$$\begin{aligned} \top^\mathcal{I} &= \Delta^\mathcal{I} \\ \bot^\mathcal{I} &= \varnothing \\ (C_1 \sqcap C_2)^\mathcal{I} &= C_1^\mathcal{I} \cap C_2^\mathcal{I} \\ (\exists r.D)^\mathcal{I} &= \{\, x \in \Delta^\mathcal{I} \mid \exists y \in \Delta^\mathcal{I} \colon (x, y) \in r^\mathcal{I} \wedge y \in D^\mathcal{I} \,\} \end{aligned}$$

We say that an element $x \in \Delta^\mathcal{I}$ *satisfies* C if and only if $x \in C^\mathcal{I}$.

Similar to the notion of an implication, we shall define a *general concept inclusion* $C \sqsubseteq D$ to be a pair (C, D) of $\mathcal{EL}^\bot$-concept descriptions. A GCI $C \sqsubseteq D$ *holds* in $\mathcal{I}$ if and only if $C^\mathcal{I} \subseteq D^\mathcal{I}$, i. e. if every element that satisfies C also satisfies D.

Examples of GCIs are

$$\exists \mathsf{child}.\top \sqsubseteq \mathsf{Person}$$
$$\mathsf{Mouse} \sqcap \mathsf{Cat} \sqsubseteq \bot.$$

Intuitively, the first GCI expresses the fact that everything having a child is actually a person. The second GCI states that there are no things which are both a mouse and a cat.

If $\mathcal{B}$ is a set of GCIs, then $\mathcal{B}$ is said to be *sound* for $\mathcal{I}$ if and only if every GCI in $\mathcal{B}$ holds in $\mathcal{I}$. $\mathcal{B}$ is said to be *complete* for $\mathcal{I}$ if every GCI valid in $\mathcal{I}$ is already entailed by $\mathcal{B}$. Thereby, the set $\mathcal{B}$ *entails* a GCI $C \sqsubseteq D$ if and only if for each interpretation $\mathcal{J}$ where all GCIs in $\mathcal{B}$ hold, the GCI $C \sqsubseteq D$ holds as well, or in other words,

$$\forall \mathcal{J} : (\mathcal{J} \models \mathcal{B} \implies \mathcal{J} \models (C \sqsubseteq D)).$$

We write $\mathcal{B} \models (C \sqsubseteq D)$ in this case. As in the case of implications, the set $\mathcal{B}$ is a *base* of $\mathcal{I}$ if and only if $\mathcal{B}$ is sound and complete for $\mathcal{I}$.

In some cases, a GCI $C \sqsubseteq D$ may be valid in all interpretations. In this case, we say that C is *subsumed by* D, or that C is *more specific* than D. In this case, we simply write $C \sqsubseteq D$ (note that there is no risk of confusion, as a GCI $C \sqsubseteq D$ is an expression, while the fact that C is subsumed by D is an statement.) We call two concept descriptions C and D *equivalent* if and only if $C \sqsubseteq D$ and $D \sqsubseteq C$. We shall write $C \equiv D$ in this case.

If $\mathcal{C}$ is another set of GCIs, we say that $\mathcal{C}$ and $\mathcal{B}$ are *equivalent* if and only if every GCI from $\mathcal{B}$ is entailed by $\mathcal{C}$ and vice versa. We say that $\mathcal{C}$ is *complete* for $\mathcal{B}$ if and only if every GCI from $\mathcal{B}$ already follows from $\mathcal{C}$.

Finally, a flavor of $\mathcal{EL}^\bot$ we shall mention here is $\mathcal{EL}^\bot_{\mathrm{gfp}}$, an extension of $\mathcal{EL}^\bot$ using *greatest fixpoint semantics*. Intuitively, $\mathcal{EL}^\bot_{\mathrm{gfp}}$ can be understood as extending $\mathcal{EL}^\bot$ with *cyclic concept descriptions*. Although this description logic is crucial for our technical considerations, it is not necessary to introduce it formally here. We refer interested readers to [12,8,5].

3 Confident General Concept Inclusions

We have now the necessary definition in place to motivate and introduce the notion of *confidence* of general concept inclusions. It shall turn out that this definition is a straight-forward generalization from the definition of confidence for implications. However, before we shall come to this, we want to describe in more detail the experiment with the DBpedia data set mentioned in the introduction. We also want to discuss the similarities between linked data and description logic interpretations.

One of the main parts of the linked open data cloud is the DBpedia data set. This is a collection of RDF Triples extracted from the Infoboxes of Wikipedia articles. Two examples for such triples are[1]

[1] Strictly speaking, we consider *serializations* of RDF Triples here

```
<http://dbpedia.org/resource/Aristotle>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/Philosopher> .

<http://dbpedia.org/resource/Aristotle>
<http://dbpedia.org/ontology/influenced>
<http://dbpedia.org/resource/Western_philosophy> .
```

The first triple states the fact that Aristotle was a philosopher, and the second triple encodes that Aristotle influenced Western Philosophy. Every RDF Triple in the DBpedia data set considered here has either of these forms. Let us call RDF Triples like the first one above *instance triples* (or *typing triples*) and RDF Triples like the second one *role triples*.

We can understand a set R of RDF Triples as a vertex- and edge-labeled graph G. Intuitively, we use role triples as edges of this graph, and instance triples provide the labels of the edges. For example, the two triples mentioned above would yield the following graph:

where node 1 denotes Aristotle and node 2 denotes Western_philosophy.

Let us suppose we have given such a vertex- and edge-labeled graph G (not necessarily, but possibly constructed form a set of RDF Triples). From the graph G, we can easily construct two sets N_C, N_R of concept- and role-names and an interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ that represents this graph. As the set N_C we just collect all labels of vertices in G. Furthermore, as set N_R of role names, we collect all labels of edges in G. For the set $\Delta^{\mathcal{I}}$ of elements of $\mathcal{I}$ we just collect the vertices of G. Finally, we define the interpretation mapping $\cdot^{\mathcal{I}}$ for $A \in N_C$ and $r \in N_R$ as follows:

$$A^{\mathcal{I}} := \{\, x \in \Delta^{\mathcal{I}} \mid x \text{ is labeled with } A \text{ in } G \,\},$$
$$r^{\mathcal{I}} := \{\, (x, y) \in \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \mid \text{an edge between } x \text{ and } y \text{ is labeled with } r \text{ in } G \,\}.$$

It is clear that the interpretation $\mathcal{I}$ is just another notation for the graph G. Therefore, if G indeed has been constructed from a set of RDF Triples R, then the interpretation $\mathcal{I}$ is only another syntactical representation of R. However, using the interpretation $\mathcal{I}$ we are now able to apply the methods developed by Baader and Distel.

For the following experiments, we have considered the DBpedia data set version 3.5, which extracted its data from the Wikipedia at late March 2010.

For our experiment, we apply the above mentioned construction to obtain an interpretation $\mathcal{I}_{\text{DBpedia}} = (\Delta^{\mathcal{I}_{\text{DBpedia}}}, \cdot^{\mathcal{I}_{\text{DBpedia}}})$ that represents the child-relation in DBpedia. For this, we collect the set of all role triples whose second component is

```
http://dbpedia.org/ontology/child
```

Additionally, we collect all instance triples where the first or third entry also occurs in such a role triple. From the resulting set of RDF Triples we then construct

the interpretation $\mathcal{I}_{\text{DBpedia}}$. This interpretation then contains 5624 elements, i. e. $|\Delta^{\mathcal{I}_{\text{DBpedia}}}| = 5624$, and 60 concept names. Since we only considered the child relation during our construction, it is the only role that appears in $\mathcal{I}_{\text{DBpedia}}$. To get a base for this interpretation, we apply the algorithm by Baader and Distel and obtain a set $\mathcal{B}_{\mathcal{I}_{\text{DBpedia}}}$ of GCIs with 1252 elements. This set of GCIs now compactly represents all valid GCIs of $\mathcal{I}_{\text{DBpedia}}$, i. e. a GCI is valid in $\mathcal{I}_{\text{DBpedia}}$ if and only if it follows from $\mathcal{B}_{\mathcal{I}_{\text{DBpedia}}}$.

While carefully examining the GCIs thus obtained, one notes that some of these GCIs are a bit artificial in the sense that one would expect more general GCIs to hold. For example, the following GCI has been obtained during the algorithm:

$$\exists \mathsf{child}.\exists \mathsf{child}.\top \sqsubseteq \exists \mathsf{child}.(\mathsf{Person} \sqcap \exists \mathsf{child}.\top)$$

Here, Person is an extracted concept name denoting persons.[2] This GCI roughly states that everything (everyone) which has a grandchild also has a child which is a person having a child. Albeit the child-relation in DBpedia is more general as expected (for example, it also denotes works of artists as children of these artists), one would expect a more general GCI than this to be true. In particular, one can expect that everything that has a child is already a person, even in this more general setting of the child-relation in DBpedia. Therefore, one would expect the GCI

$$\exists \mathsf{child}.\top \sqsubseteq \mathsf{Person} \tag{1}$$

to be true in $\mathcal{I}_{\text{DBpedia}}$. However, this is not the case. A closer look at the data set reveals that there are four counterexamples in $\mathcal{I}_{\text{DBpedia}}$ for this GCI, i. e. there are four elements $x \in \Delta^{\mathcal{I}_{\text{DBpedia}}}$ such that that $x \in (\exists \mathsf{child}.\top)^{\mathcal{I}_{\text{DBpedia}}} \setminus \mathsf{Person}^{\mathcal{I}_{\text{DBpedia}}}$. These four counterexamples are the individuals `Teresa_Carpio`, `Charles_Heung`, `Adam_Cheng` and `Lydia_Shum`. All these individuals represent artists from Hong Kong, which are certainly persons and should therefore be labeled as Person. Therefore, all counterexamples to (1) are caused by errors, and the GCI (1) can indeed be regarded as valid.

This observation reveals a drawback in the approach of Baader and Distel. When considering only valid GCIs, even singleton errors can turn otherwise valid GCIs into invalid ones, which are then not extracted anymore by the algorithm. However, those GCIs may very well be of interest for an ontology engineer.

As first approach to circumvent this undesired behavior is to consider GCIs which are "almost true" in addition to valid GCIs. To make this more precise, we introduce the notion of *confidence* for GCIs as follows.

Definition 1. *Let $\mathcal{I}$ be a finite interpretation and let $C \sqsubseteq D$ be a GCI. Then define the* confidence *of $C \sqsubseteq D$ in $\mathcal{I}$ to be*

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

[2] For readability, we omit the prefix `http://dbpedia.org/ontology/` from now on

For a given GCI $C \sqsubseteq D$, the value $\mathrm{conf}_{\mathcal{I}}(C \sqsubseteq D)$ is just the empirical probability that an element $x \in \Delta^{\mathcal{I}}$ satisfying C also satisfies D. Of course, if there are no elements in $\mathcal{I}$ that satisfy C, then this GCI is vacuously true and its confidence is 1.

Note that we can equally well define the confidence for implications $A \to B$ in a formal context $\mathbb{K}$ in a very similar way.

We now want to use this notion of confidence to generalize the approach by Baader and Distel. Let us denote with $\mathrm{Th}(\mathcal{I})$ the set of all GCIs valid in $\mathcal{I}$. The approach of Baader and Distel can be understood as finding a finite set $\mathcal{B} \subseteq \mathrm{Th}(\mathcal{I})$ of GCIs such that every GCI in $\mathrm{Th}(\mathcal{I})$ is already entailed by $\mathcal{B}$. Note that $\mathcal{B}$ is a base of $\mathcal{I}$ if and only $\mathcal{B}$ entails all GCIs from $\mathrm{Th}(\mathcal{I})$ and vice versa. In the following definition, we shall lift this understanding to the setting of GCIs with high confidence.

Definition 2. *Let $\mathcal{I}$ be a finite interpretation and let $c \in [0,1]$. Let us denote with $\mathrm{Th}_c(\mathcal{I})$ the set of all GCIs whose confidence in $\mathcal{I}$ is at least c.*

A set $\mathcal{B}$ of GCIs is called base *for $\mathrm{Th}_c(\mathcal{I})$ if and only if $\mathcal{B}$ entails all GCIs from $\mathrm{Th}_c(\mathcal{I})$ and vice versa. The set $\mathcal{B}$ is called a* confident base *if and only if $\mathcal{B} \subseteq \mathrm{Th}_c(\mathcal{I})$ and $\mathcal{B}$ is a base for $\mathrm{Th}_c(\mathcal{I})$.*

Our goal now is to find a finite, confident base of $\mathrm{Th}_c(\mathcal{I})$. The hope is that this base will be much less sensitive to small sets of errors as bases of $\mathcal{I}$ are, and may therefore contain additional information for the construction of an ontology from $\mathcal{I}$.

4 Bases of Confident General Concept Inclusions

For the following discussions let us fix an arbitrary but finite interpretation $\mathcal{I}$ and a number $c \in [0,1]$. The purpose of this section is to effectively describe a confident base of $\mathrm{Th}_c(\mathcal{I})$. For this, we shall make use of ideas from the approach of Baader and Distel [2], which we shall introduce in the next subsection. Thereafter, we shall introduce a first base of $\mathrm{Th}_c(\mathcal{I})$ by applying ideas from the theory of *partial implications* to our setting. These ideas go back to work of Luxenburger [11,10]. Finally, we shall exploit another idea from Luxenburger to describe a base of $\mathrm{Th}_c(\mathcal{I})$ that is potentially smaller than the one discussed before.

4.1 Model-Based Most-Specific Concept Descriptions

One of the main achievements of the approach by Baader and Distel is to reveal a tight connection between formal concept analysis and the description logic $\mathcal{EL}^{\perp}$ (or, more precisely, $\mathcal{EL}^{\perp}_{\mathrm{gfp}}$.) The key notion necessary for this result is the one of a model-based most-specific concept description.

Let $X \subseteq \Delta^{\mathcal{I}}$. Then a *model-based most-specific concept description* is a concept description C such that

i. $X \subseteq C^{\mathcal{I}}$ and
ii. for each concept description D satisfying $X \subseteq D^{\mathcal{I}}$, it is true that $C \sqsubseteq D$.

Intuitively, a model-based most-specific concept description for X is a most-specific concept description such that all elements of X satisfy it. Obviously, if such a concept description exists, it is unique up to equivalence. We shall denote it by $X^{\mathcal{I}}$, to remind the similarities with the derivation operators from formal concept analysis. Indeed, provided that model-based most-specific concept descriptions exists, it is true for all $X \subseteq \Delta^{\mathcal{I}}$ and concept descriptions C that

$$X \subseteq C^{\mathcal{I}} \iff X^{\mathcal{I}} \sqsubseteq C. \tag{2}$$

Thus, the interpretation function of $\mathcal{I}$ and model-based most-specific concept descriptions satisfy the main condition of a Galois connection. Note, however, that $\sqsubseteq$ does not constitute an order relation on the set of all concept descriptions.

An easy consequence of (2), and indeed of the very definition of model-based most-specific concept description, is that $(C^{\mathcal{I}})^{\mathcal{I}} \sqsubseteq C$ is true for all concept descriptions C. We shall exploit this fact repeatedly in our further discussions. Conversely, note that $C \sqsubseteq C^{\mathcal{I}\mathcal{I}}$ is a valid GCI of $\mathcal{I}$ for all concept descriptions C.

One drawback of the notion of model-based most-specific concept description is that they do not necessarily need to exist in the description logic $\mathcal{EL}^{\bot}$. In other words, if X is given in the above definition, it may occur that there is no $\mathcal{EL}^{\bot}$-concept description that is a model-based most-specific concept description for X. See [8] for examples.

This shortcoming can be circumvented by considering $\mathcal{EL}^{\bot}_{\text{gfp}}$-concept descriptions instead of $\mathcal{EL}^{\bot}$-concept description. This, however, requires a lot more technical work to do, which is not possible in the available amount of space. Luckily, it can be shown that the bases we are going to discuss in the next subsections can effectively be turned into equivalent sets of GCIs only containing $\mathcal{EL}^{\bot}$-concept description. Restricting our attention to $\mathcal{EL}^{\bot}$-concept descriptions is therefore no loss of generality. See [6] for further details on this.

With the help of model-based most-specific concept descriptions, we can effectively find bases for $\mathcal{I}$. For this, we define the following set of *essential concept descriptions*:

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

We can view $M_{\mathcal{I}}$ as a set of *attributes* for a suitable formal context $\mathbb{K}_{\mathcal{I}}$, which we shall call the *induced formal context* of $\mathcal{I}$. This formal context is defined as $\mathbb{K}_{\mathcal{I}} = (\Delta^{\mathcal{I}}, M_{\mathcal{I}}, \nabla)$, where $x \nabla C \iff x \in C^{\mathcal{I}}$ for all $x \in \Delta^{\mathcal{I}}$ and $C \in M_{\mathcal{I}}$.

Now consider a subset $U \subseteq M_{\mathcal{I}}$. Then U is a set of concept descriptions, but also a set of attributes of $\mathbb{K}_{\mathcal{I}}$. If $x \in \Delta^{\mathcal{I}}$ is such that $x \in U'$, then x satisfies every concept description in U. Therefore, x also satisfies the conjunction of all concept descriptions in U, i. e. $x \in (\bigsqcap_{V \in U} V)^{\mathcal{I}}$. To be able to write this more briefly, let us define

$$\bigsqcap U := \begin{cases} \top & \text{if } U = \varnothing, \\ \bigsqcap_{V \in U} V & \text{otherwise.} \end{cases}$$

From [8], we now obtain the following result.

Theorem 1. *The set*

$$\mathcal{B}_2 := \{ \bigsqcap U \sqsubseteq (\bigsqcap U)^{\mathcal{I}\mathcal{I}} \mid U \subseteq M_{\mathcal{I}} \}$$

is a finite base for $\mathcal{I}$.

Indeed, this result can be generalized in the following way: for every base $\mathcal{B}$ of $\mathbb{K}_{\mathcal{I}}$, where the implications in $\mathcal{B}$ are of the form $U \to U''$, it is true that the set

$$\{ \bigsqcap U \sqsubseteq (\bigsqcap U)^{\mathcal{I}\mathcal{I}} \mid (U \to U'') \in \mathcal{B} \}$$

is a base of $\mathcal{I}$. See also [8] for more details on this.

4.2 A First Base

In this subsection we want to effectively describe a finite base of $\mathrm{Th}_c(\mathcal{I})$. To achieve this, we shall make use ideas from the theory of partial implications, developed by Luxenburger. The work of Luxenburger was concerned, among others, with finding *bases of partial implications* of a formal context $\mathbb{K}$. Due to space restrictions, we shall only give a very brief overview of the relevant parts here.

Partial implications can be understood as implications where their confidence in $\mathbb{K}$ is attached to them, i. e. partial implications are of the form $A \to^c B$, where c is the confidence of $A \to B$ in $\mathbb{K}$. The two main observations of Luxenburger's study which we want to utilize are the following: firstly, partial implications with confidence 1 correspond bijectively with the valid implications of $\mathbb{K}$. Thus when searching for bases of partial implications, it is enough to consider only those whose confidence is not 1, since for those with confidence 1 we can simply use bases of $\mathbb{K}$. Secondly, we can observe that the confidence of $A \to B$ and $A'' \to B''$ are the same, and it is sufficient to only consider the latter when searching for bases, since it already entails the former.

We shall make these ideas more precisely by translating them to our setting, and using them to find confident bases of $\mathrm{Th}_c(\mathcal{I})$. To this end, we use the first idea and consider the partition $\mathrm{Th}_c(\mathcal{I}) = \mathrm{Th}(\mathcal{I}) \cup (\mathrm{Th}_c(\mathcal{I}) \setminus \mathrm{Th}(\mathcal{I}))$ and try to separately find a base for $\mathrm{Th}(\mathcal{I})$ and a subset of $\mathrm{Th}_c(\mathcal{I}) \setminus \mathrm{Th}(\mathcal{I})$ which already entails all GCIs of this set. Of course, a base $\mathcal{B}$ of $\mathrm{Th}(\mathcal{I})$ is already given in Theorem 1, so it remains to find a complete subset of $\mathrm{Th}_c(\mathcal{I}) \setminus \mathrm{Th}(\mathcal{I})$.

To achieve this, we use the second idea as follows: if $(C \sqsubseteq D) \in \mathrm{Th}_c(\mathcal{I}) \setminus \mathrm{Th}(\mathcal{I})$, it is true that

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

because $\mathcal{B} \models (C \sqsubseteq C^{\mathcal{I}\mathcal{I}})$, and $D^{\mathcal{I}\mathcal{I}} \sqsubseteq D$ holds anyway. Therefore, it suffices to consider only GCIs of the form $C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}}$. So, let us define

$$\mathrm{Conf}(\mathcal{I}, c) := \{ X^{\mathcal{I}} \sqsubseteq Y^{\mathcal{I}} \mid Y \subseteq X \subseteq \Delta^{\mathcal{I}} \text{ and } \mathrm{conf}_{\mathcal{I}}(X^{\mathcal{I}} \sqsubseteq Y^{\mathcal{I}}) \in [c, 1) \}.$$

Note that each GCI in $(X^{\mathcal{I}} \sqsubseteq Y^{\mathcal{I}}) \in \mathrm{Conf}(\mathcal{I}, c)$ is of the form $C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}}$: just define $C := X^{\mathcal{I}}, D := Y^{\mathcal{I}}$ and note that $C^{\mathcal{I}\mathcal{I}} = X^{\mathcal{I}\mathcal{I}\mathcal{I}} \equiv X^{\mathcal{I}}$ and likewise for D.

Theorem 2. *Let $\mathcal{I}$ be a finite interpretation, let $c \in [0,1]$ and let $\mathcal{B}$ 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. Clearly $\mathcal{B} \cup \mathrm{Conf}(\mathcal{I}, c) \subseteq \mathrm{Th}_c(\mathcal{I})$ and it only remains to show that $\mathcal{B} \cup \mathrm{Conf}(\mathcal{I}, c)$ entails all GCIs with confidence at least c in $\mathcal{I}$.

Let $C \sqsubseteq D$ be an GCI with $\mathrm{conf}_{\mathcal{I}}(C \sqsubseteq D) \geqslant c$. We have to show that $\mathcal{B} \cup \mathrm{Conf}(\mathcal{I}, c) \models (C \sqsubseteq D)$. If $C \sqsubseteq D$ is already valid in $\mathcal{I}$, then $\mathcal{B} \models (C \sqsubseteq D)$ and nothing remains to be shown. We therefore assume that $\mathrm{conf}_{\mathcal{I}}(C \sqsubseteq D) \neq 1$.

As $C \sqsubseteq C^{\mathcal{I}\mathcal{I}}$ is valid in $\mathcal{I}$, $\mathcal{B} \models (C \sqsubseteq C^{\mathcal{I}\mathcal{I}})$. Furthermore, $\mathrm{conf}_{\mathcal{I}}(C \sqsubseteq D) = \mathrm{conf}_{\mathcal{I}}(C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}})$ and hence $(C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}}) \in \mathrm{Conf}(\mathcal{I}, c)$. Additionally, $D^{\mathcal{I}\mathcal{I}} \sqsubseteq D$ holds. We therefore obtain

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

and hence $\mathcal{B} \cup \mathrm{Conf}(\mathcal{I}, c) \models (C \sqsubseteq D)$ as required. □

It is not hard to see that the prerequisites of the previous theorem can be weakened in the following way: instead of considering the whole set $\mathrm{Conf}(\mathcal{I}, c)$, it is sufficient to choose a subset $\mathcal{C} \subseteq \mathrm{Conf}(\mathcal{I}, c)$ of $\mathrm{Conf}(\mathcal{I}, c)$ that already entails all GCIs in $\mathrm{Conf}(\mathcal{I}, c)$ (i. e. is complete for it), since then

$$\mathcal{B} \cup \mathcal{C} \models \mathcal{B} \cup \mathrm{Conf}(\mathcal{I}, c).$$

Furthermore, it is not necessary for $\mathcal{B}$ to be a base of $\mathcal{I}$. Instead, one can choose a set $\hat{\mathcal{B}}$ of valid GCIs such that $\hat{\mathcal{B}} \cup \mathcal{C}$ is complete for $\mathcal{I}$, because then

$$\hat{\mathcal{B}} \cup \mathcal{C} \models \mathcal{B} \cup \mathcal{C}.$$

Corollary 1. *Let $\mathcal{I}$ be a finite interpretation, $c \in [0,1]$. Let $\mathcal{C} \subseteq \mathrm{Conf}(\mathcal{I}, c)$ be complete for $\mathrm{Conf}(\mathcal{I}, c)$ and let $\mathcal{B} \subseteq \mathrm{Th}(\mathcal{I})$ such that $\mathcal{B} \cup \mathcal{C}$ is complete for $\mathcal{I}$. Then $\mathcal{B} \cup \mathcal{C}$ is a confident base of $\mathrm{Th}_c(\mathcal{I})$.*

4.3 A Smaller Base

With the previous result, we are able to effectively describe a finite base of $\mathrm{Th}_c(\mathcal{I})$. However, we can make the set $\mathrm{Conf}(\mathcal{I}, c)$ potentially smaller by using another idea of Luxenburger, which is based on the following observation: let C_1, C_2, C_3 be concept descriptions such that $C_1^{\mathcal{I}} \supseteq C_2^{\mathcal{I}} \supseteq C_3^{\mathcal{I}}$. Then it is true that

$$\mathrm{conf}_{\mathcal{I}}(C_1 \sqsubseteq C_3) = \mathrm{conf}_{\mathcal{I}}(C_1 \sqsubseteq C_2) \cdot \mathrm{conf}_{\mathcal{I}}(C_2 \sqsubseteq C_3). \tag{3}$$

We can make use of (3) to find a subset of $\mathrm{Conf}(\mathcal{I}, c)$ that is complete for it in the following way. Suppose that $(C_1 \sqsubseteq C_3) \in \mathrm{Conf}(\mathcal{I}, c)$. Then by (3), $\mathrm{conf}_{\mathcal{I}}(C_1 \sqsubseteq C_2) \geqslant c$ and $\mathrm{conf}_{\mathcal{I}}(C_2 \sqsubseteq C_3) \geqslant c$ and hence $(C_1 \sqsubseteq C_2), (C_2 \sqsubseteq C_3) \in \mathrm{Conf}(\mathcal{I}, c)$. But the latter GCIs already entail $C_1 \sqsubseteq C_3$, therefore it is not needed. Generalizing this idea, we can say that each GCI $(C \sqsubseteq D) \in \mathrm{Conf}(\mathcal{I}, c)$ is redundant whenever there exists a concept description E such that $C^{\mathcal{I}} \supseteq E^{\mathcal{I}} \supseteq D^{\mathcal{I}}$ and E not equivalent to both C and D.

We shall now give proofs for this argumentation. The line of argumentation has been inspired by proofs from [14].

Lemma 1. *Let $\mathcal{I}$ be a finite interpretation and let $(C_i \mid i = 0, \dots, n), n \in \mathbb{N}$, be a finite sequence of concept descriptions such that $C_{i+1}^{\mathcal{I}} \subseteq C_i^{\mathcal{I}}$ for all $i = 1, \dots, n-1$. Then*

$$\mathrm{conf}_{\mathcal{I}}(C_0 \sqsubseteq C_n) = \prod_{i=0}^{n-1} \mathrm{conf}_{\mathcal{I}}(C_i \sqsubseteq C_{i+1}).$$

Proof. Let us first assume that the set $\{\, i \mid C_i^{\mathcal{I}} = \varnothing \,\}$ is not empty and let

$$i_0 := \min \{\, i \mid C_i^{\mathcal{I}} = \varnothing \,\}.$$

If $i_0 = 0$, then $C_j^{\mathcal{I}} = \varnothing$ for all $j \in \{\, 0, \dots, n \,\}$, hence $\mathrm{conf}_{\mathcal{I}}(C_0 \sqsubseteq C_n) = 1$ and $\mathrm{conf}_{\mathcal{I}}(C_j \sqsubseteq C_{j+1}) = 1$ for all $j \in \{\, 0, \dots, n \,\}$.

Otherwise, $0 < i_0 \leqslant n$. But then $C_n^{\mathcal{I}} = \varnothing$ and hence $\mathrm{conf}_{\mathcal{I}}(C_0 \sqsubseteq C_n) = 0$. Furthermore, $\mathrm{conf}_{\mathcal{I}}(C_{i-1} \sqsubseteq C_i) = 0$ since $C_{i-1}^{\mathcal{I}} \neq \varnothing$ and $C_i^{\mathcal{I}} = \varnothing$. Therefore,

$$\prod_{i=0}^{n-1} \mathrm{conf}_{\mathcal{I}}(C_i \sqsubseteq C_{i+1}) = 0 = \mathrm{conf}_{\mathcal{I}}(C_0 \sqsubseteq C_n).$$

Finally, let us consider the case when $\{\, i \mid C_i^{\mathcal{I}} = \varnothing \,\}$ is empty. Then we can calculate

$$\begin{aligned}
\prod_{i=1}^{n-1} \mathrm{conf}_{\mathcal{I}}(C_i \sqsubseteq C_{i+1}) &= \prod_{i=1}^{n-1} \frac{|C_i^{\mathcal{I}} \cap C_{i+1}^{\mathcal{I}}|}{|C_i^{\mathcal{I}}|} \\
&= \prod_{i=1}^{n-1} \frac{|C_{i+1}^{\mathcal{I}}|}{|C_i^{\mathcal{I}}|} = \frac{|C_n^{\mathcal{I}}|}{|C_0^{\mathcal{I}}|} \\
&= \frac{|C_0^{\mathcal{I}} \cap C_n^{\mathcal{I}}|}{|C_0^{\mathcal{I}}|} = \mathrm{conf}_{\mathcal{I}}(C_0 \sqsubseteq C_n).
\end{aligned}$$

□

Theorem 3. *Let $\mathcal{I}$ be a finite interpretation and let $c \in [0, 1]$. Define the set*

$$\begin{aligned}
\mathrm{Lux}(\mathcal{I}, c) := \{\, X^{\mathcal{I}} \sqsubseteq Y^{\mathcal{I}} \mid Y \subseteq X \subseteq \Delta_{\mathcal{I}}, 1 > \mathrm{conf}_{\mathcal{I}}(X^{\mathcal{I}} \sqsubseteq Y^{\mathcal{I}}) \geqslant c, \\
\nexists Z \subseteq \Delta_{\mathcal{I}} \colon Y \subseteq Z \subseteq X \text{ and } Y^{\mathcal{I}} \neq Z^{\mathcal{I}} \neq X^{\mathcal{I}} \,\}.
\end{aligned}$$

Then $\mathrm{Lux}(\mathcal{I}, c) \subseteq \mathrm{Conf}(\mathcal{I}, c)$ and $\mathrm{Lux}(\mathcal{I}, c)$ is complete for $\mathrm{Conf}(\mathcal{I}, c)$. In particular, if $\mathcal{B}$ is a finite base of $\mathcal{I}$, then $\mathcal{B} \cup \mathrm{Lux}(\mathcal{I}, c)$ is a finite base of $\mathrm{Th}_c(\mathcal{I})$.

Proof. Let $C^{\mathcal{II}} \sqsubseteq D^{\mathcal{II}} \in \mathrm{Conf}(\mathcal{I}, c)$. As $(C \sqcap D)^{\mathcal{II}} \sqsubseteq D^{\mathcal{II}}$ always holds, $C^{\mathcal{II}} \sqsubseteq D^{\mathcal{II}}$ follows from $C^{\mathcal{II}} \sqsubseteq (C \sqcap D)^{\mathcal{II}}$. Furthermore, since $(C^{\mathcal{II}} \sqcap D^{\mathcal{II}})^{\mathcal{I}} = C^{\mathcal{III}} \cap D^{\mathcal{III}} = C^{\mathcal{I}} \cap D^{\mathcal{I}} = (C \cap D)^{\mathcal{I}}$, we obtain

$$\begin{aligned}
\mathrm{conf}_{\mathcal{I}}(C^{\mathcal{II}} \sqsubseteq D^{\mathcal{II}}) &= \frac{|(C^{\mathcal{II}} \sqcap D^{\mathcal{II}})^{\mathcal{I}}|}{|C^{\mathcal{III}}|} \\
&= \frac{|(C \sqcap D)^{\mathcal{I}}|}{|C^{\mathcal{III}}|}
\end{aligned}$$

$$= \frac{|(C \sqcap D)^{\mathcal{I}\mathcal{I}\mathcal{I}}|}{|C^{\mathcal{I}\mathcal{I}\mathcal{I}}|}$$
$$= \mathrm{conf}_{\mathcal{I}}(C^{\mathcal{I}\mathcal{I}} \sqsubseteq (C \sqcap D)^{\mathcal{I}\mathcal{I}})$$

since $|C^{\mathcal{I}\mathcal{I}\mathcal{I}}| \neq 0$, as otherwise $\mathrm{conf}_{\mathcal{I}}(C^{\mathcal{I}\mathcal{I}} \sqsubseteq D^{\mathcal{I}\mathcal{I}}) = 1$. Therefore, $C^{\mathcal{I}\mathcal{I}} \sqsubseteq (C \sqcap D)^{\mathcal{I}\mathcal{I}} \in \mathrm{Conf}(\mathcal{I}, c)$ and we shall show now that $\mathrm{Lux}(\mathcal{I}, c) \models (C^{\mathcal{I}\mathcal{I}} \sqsubseteq (C \sqcap D)^{\mathcal{I}\mathcal{I}})$.

Let us define $X := C^{\mathcal{I}}$ and $Y := (C \sqcap D)^{\mathcal{I}}$. Then $Y \subseteq X$. As $\Delta_{\mathcal{I}}$ is finite, the set

$$\{\, Z^{\mathcal{I}} \mid Y \subseteq Z \subseteq X, Y^{\mathcal{I}} \not\equiv Z^{\mathcal{I}} \not\equiv X^{\mathcal{I}} \,\}$$

is finite as well. Hence we can find a finite sequence $(C_i \mid 0 \leqslant i \leqslant n)$ for some $n \in \mathbb{N}$ of sets $C_i \subseteq \Delta_{\mathcal{I}}$ such that

i. $Y := C_n, X := C_0$,
ii. $C_{i+1} \subsetneq C_i$ for $0 \leqslant i < n$,
iii. $C_i^{\mathcal{I}} \not\equiv C_{i+1}^{\mathcal{I}}$ for $0 \leqslant i < n$,
iv. $C_i^{\mathcal{I}\mathcal{I}} = C_i$ for $0 \leqslant i \leqslant n$,
v. $C_{i+1} \subseteq Z \subseteq C_i$ implies $C_i^{\mathcal{I}} \equiv Z^{\mathcal{I}}$ or $C_{i+1}^{\mathcal{I}} \equiv Z^{\mathcal{I}}$ for $0 \leqslant i < n$.

Then by Lemma 1

$$\mathrm{conf}_{\mathcal{I}}(X^{\mathcal{I}} \sqsubseteq Y^{\mathcal{I}}) = \prod_{i=0}^{n-1} \mathrm{conf}_{\mathcal{I}}(C_i^{\mathcal{I}} \sqsubseteq C_{i+1}^{\mathcal{I}})$$

and therefore $\mathrm{conf}_{\mathcal{I}}(C_i^{\mathcal{I}} \sqsubseteq C_{i+1}^{\mathcal{I}}) \in [c, 1]$. As $C_i^{\mathcal{I}\mathcal{I}} \subseteq C_{i+1}^{\mathcal{I}\mathcal{I}}$ would imply $C_i \subseteq C_{i+1}$ and so $C_i \subseteq C_i$, we obtain $\mathrm{conf}_{\mathcal{I}}(C_i^{\mathcal{I}} \sqsubseteq C_{i+1}^{\mathcal{I}}) \neq 1$. Hence, $C_i^{\mathcal{I}} \sqsubseteq C_{i+1}^{\mathcal{I}\mathcal{I}} \in \mathrm{Lux}(\mathcal{I}, c)$ for $0 \leqslant i < n$. Thus

$$\mathcal{D} \models C_i^{\mathcal{I}} \sqsubseteq C_{i+1}^{\mathcal{I}}, \quad (0 \leqslant i < n)$$

and therefore $\mathrm{Lux}(\mathcal{I}, c) \models (X^{\mathcal{I}} \sqsubseteq Y^{\mathcal{I}}) = (C^{\mathcal{I}\mathcal{I}} \sqsubseteq (C \sqcap D)^{\mathcal{I}\mathcal{I}})$ as required. □

5 Experiments with the DBpedia Data Set

High confidence of a certain GCI does not necessarily imply that the GCI itself is correct. Instead, one could have the case that for this particular GCI only very few (correct) counterexamples exist. To make good use of GCIs with high confidence, ideally each of them has to be checked manually for correctness before one can include them in the final ontology.

To see how much extra work this requires and how many such GCIs have to be considered using our results from Section 4, we consider again the interpretation $\mathcal{I}_{\mathrm{DBpedia}}$. For this interpretation, we want to conduct two experiments. Firstly, we consider as minimal confidence the value $c = 0.95$ and have a closer look at all the GCIs thus obtained. We can see from this in how far our approach is helpful in finding small sets of errors in $\mathcal{I}_{\mathrm{DBpedia}}$.

Secondly, we consider the number of GCIs obtained as $\mathrm{Conf}(\mathcal{I}_{\mathrm{DBpedia}}, c)$ and $\mathrm{Lux}(\mathcal{I}_{\mathrm{DBpedia}}, c)$ for varying values of c. The rationale behind this experiment is to see how many such GCIs have to be considered by an ontology engineer.

5.1 Examining $\mathbf{Conf}(\mathcal{I}_{\mathbf{DBpedia}}, \mathbf{0.95})$

As already mentioned, we are going to investigate the set $\mathrm{Conf}(\mathcal{I}_{\mathrm{DBpedia}}, 0.95)$. For this, we shall discuss whether the GCIs contained in this set are actually valid GCIs, by manually checking whether all counterexamples are only due to errors:[3]

$$\begin{aligned}\mathrm{Conf}(\mathcal{I}_{\mathrm{DBpedia}}, 0.95) = \{\, & \mathsf{Place} \sqsubseteq \mathsf{PopulatedPlace},\\ & \exists \mathsf{child}.\top \sqsubseteq \mathsf{Person},\\ & \exists \mathsf{child}.\exists \mathsf{child}.\top \sqcap \exists \mathsf{child}.\mathsf{OfficeHolder}\\ & \quad \sqsubseteq \exists \mathsf{child}.(\mathsf{OfficeHolder} \sqcap \exists \mathsf{child}.\top)\}\end{aligned}$$

It is quite surprising that this set turns out to have only three elements. Moreover, the set $\mathrm{Conf}(\mathcal{I}_{\mathrm{DBpedia}}, 0.95)$ contains the GCI $\exists \mathsf{child}.\top \sqsubseteq \mathsf{Person}$, for which we have argued in Section 3 that it should be regarded as a valid GCI. It is also convincing that the GCI $\mathsf{Place} \sqsubseteq \mathsf{PopulatedPlace}$ is reasonable as well (places named in DBpedia appear because famous people have been born or lived there), and the only counterexample to this GCI is $\mathsf{Greenwich\_Village}$, denoting a district of New York which certainly is populated.

So, it only remains to consider the GCI

$$\exists \mathsf{child}.\exists \mathsf{child}.\top \sqcap \exists \mathsf{child}.\mathsf{OfficeHolder} \sqsubseteq \exists \mathsf{child}.(\mathsf{OfficeHolder} \sqcap \exists \mathsf{child}.\top).$$

At first sight, this GCI appears to be too specific to be considered as a valid GCI. The only counterexample is the individual $\mathsf{Pierre\_Samuel\_du\_Pont\_de\_Nemours}$, denoting a French government official who had the sons Victor Marie du Pont and Eleuthère Irénée du Pont. The first had become a French diplomat and is therefore listed in $\mathcal{I}_{\mathrm{DBpedia}}$ as an instance of $\mathsf{OfficeHolder}$. Although he had four children, none of them got famous enough to be named in the Wikipedia infobox of the corresponding Wikipedia article[4]. On the other hand, his brother Eleuthère Irénée du Pont became a famous American industrial and had a lot of famous children, which are listed in the Wikipedia infobox and therefore appear in $\mathcal{I}_{\mathrm{DBpedia}}$.

One could see this counterexample as a false one, as both sons had children. If, however, one adopts the interpretation of the child-relation of DBpedia as having *famous* children, one may be inclined to accept this counterexample. The final decision has to be made by a domain expert.

5.2 Size Behavior of $\mathbf{Conf}(\mathcal{I}_{\mathbf{DBpedia}}, c)$ and $\mathbf{Lux}(\mathcal{I}_{\mathbf{DBpedia}}, c)$

As we have seen in the previous section, the examination of the extra GCIs in $\mathrm{Conf}(\mathcal{I}_{\mathrm{DBpedia}}, c)$ may be difficult task. It is therefore interesting to know how many such GCIs an ontology engineer would have to examine for varying values of minimal confidence c.

[3] We have removed some redundancies in the concept descriptions to make them more readable. The GCIs extracted by the algorithm are actually much longer, but equivalent to those shown here.

[4] as of 13. November 2012

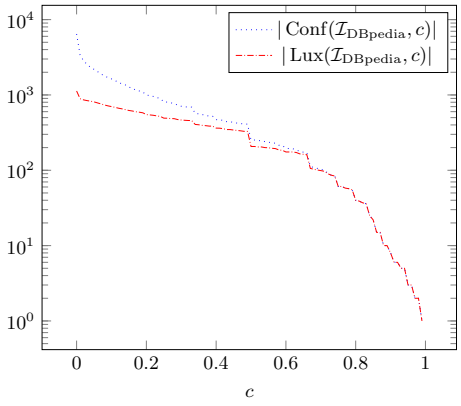

Fig. 1. Size of $\operatorname{Conf}(\mathcal{I}_{\mathrm{DBpedia}}, c)$ and $\operatorname{Lux}(\mathcal{I}_{\mathrm{DBpedia}}, c)$ for all $c \in V$

To see how the number of extra GCIs behaves for varying values of c, we consider the sizes of the sets $\operatorname{Conf}(\mathcal{I}_{\mathrm{DBpedia}}, c)$ and $\operatorname{Lux}(\mathcal{I}_{\mathrm{DBpedia}}, c)$ for all $c = 0, 0.01, 0.02, \ldots, 0.99$. The results are shown graphically in Figure 1. Note that the y-axis is scaled logarithmically.

The results given in this picture show that the number of confident GCIs the ontology engineer has to check manually declines exponentially as the minimal confidence grows. Even for $c = 0.86$, there are only 15 extra GCIs to investigate. Given the fact that a base of $\mathcal{I}_{\mathrm{DBpedia}}$ has 1252 elements, this extra effort seems negligible. Of course, it is not clear whether this behavior is typical or just particular to our data set. However, it indicates that considering confident GCIs for data, where the quality is good enough (i. e. where only few errors have been made), is not a noteworthy overhead.

Another observation is that the sets $\operatorname{Conf}(\mathcal{I}_{\mathrm{DBpedia}}, c)$ and $\operatorname{Lux}(\mathcal{I}_{\mathrm{DBpedia}}, c)$ differ only noticeably for values of c below around 0.7. For higher values of c, the idea of exploiting the multiplicativity of $\operatorname{conf}_{\mathcal{I}_{\mathrm{DBpedia}}}$ does not yield any reduction in the size of the base.

6 Conclusions

Starting from the experimental examination of the approach of Baader and Distel, we have motivated and introduced the notion of confidence for general concept inclusions. Afterwards, we have explicitly (and thus effectively) described bases of $\operatorname{Th}_c(\mathcal{I})$ using ideas from formal concept analysis. Finally, we have applied the results thus obtained to our initial experiment and have shown that the approach provides reasonable results.

However, our approach of considering confident GCIs is highly heuristic, and nothing tells us that the extracted GCIs are really valid in our domain of discourse. To make our approach more reasonable, more investigation has to be done to provide better *validation procedures*. For example, a process of validating confident

GCIs could effectively be combined with the process of *attribute exploration* to reduce the number of expert interactions needed.

Acknowledgments The author has been supported by the DFG Research Training Group 1763 (QuantLA). The author would also like to thank the anonymous reviewers for their useful comments.

References

1. Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.
2. Franz Baader and Felix Distel. A Finite Basis for the Set of $\mathcal{EL}$-Implications Holding in a Finite Model. In Raoul Medina and Sergei A. Obiedkov, editors, *Proceedings of the 6th International Conference on Formal Concept Analysis*, volume 4933 of *Lecture Notes in Computer Science*, pages 46–61. Springer, February 2008.
3. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, March 2009.
4. Christian Bizer, Jens Lehmann, Gergi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A Crystallization Point of the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, 9 2009.
5. Daniel Borchmann. Axiomatizing Confident $\mathcal{EL}^{\perp}_{\mathrm{gfp}}$-GCIs of Finite Interpretations. Report MATH-AL-08-2012, Chair of Algebraic Structure Theory, Institute of Algebra, Technische Universität Dresden, Dresden, Germany, September 2012.
6. Daniel Borchmann. 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`.
7. Daniel Borchmann and Felix Distel. Mining of $\mathcal{EL}$-GCIs. In Myra Spiliopoulou, Haixun Wang, Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu, editors, *ICDM Workshops*, pages 1083–1090. IEEE, 2011.
8. Felix Distel. *Learning Description Logic Knowledge Bases from Data Using Methods from Formal Concept Analysis*. PhD thesis, TU Dresden, 2011.
9. Bernhard Ganter and Rudolph Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin-Heidelberg, 1999.
10. M. Luxenburger. Partial implications. FB4-Preprint, TH Darmstadt, 1994.
11. Michael Luxenburger. *Implikationen, Abhängigkeiten und Galois-Abbildungen*. PhD thesis, TH Darmstadt, 1993.
12. Bernhard Nebel. Terminological Cycles: Semantics and Computational Properties. In *Principles of Semantic Networks*, pages 331–362. Morgan Kaufmann, 1991.
13. Cathy Price and Kent Spackman. SNOMED Clinical Terms. *British Journal of Healtcare Computing and Information Management*, 17:27–31, 2000.
14. Gerd Stumme, Rafik Taouil, Yves Bastide, Nicolas Pasquier, and Lotfi Lakhal. Intelligent structuring and reducing of association rules with formal concept analysis. In Franz Baader, Gerhard Brewka, and Thomas Eiter, editors, *KIÖGAI*, volume 2174 of *Lecture Notes in Computer Science*, pages 335–350. Springer, 2001.