Fostering Web Intelligence by Semi-automatic OWL Ontology Refinement*

Johanna Völker and Sebastian Rudolph
Institute AIFB, Universität Karlsruhe (TH), Germany
E-mail: {voelker,rudolph}@aifb.uni-karlsruhe.de

Abstract

In this paper, we propose a systematic, reasoner-aided approach to Web ontology acquisition and refinement. It complements methods for acquiring expressive ontology axioms from textual definitions with methodic knowledge exploration techniques based on formal concept analysis. We demonstrate the practical relevance of our approach by means of a real-world example.

1 Introduction

Ontologies constitute the backbone of Semantic Web technologies and hence provide an essential ingredient for Web Intelligence. The increasing uptake of these technologies by industry intensifies the need for medium- to large-size, yet expressive and high-quality ontologies. However, the modeling and maintenance tasks required to satisfy those needs easily surpass the capabilities of human knowledge engineers if not thoroughly assisted by automatic or semi-automatic methods. The incompleteness of information or even modeling errors often remain undetected until their accidental discovery due to unexpected query results.

Let us illustrate this claim by a small example: Querying our institute's Web portal ontology, SWRC (Semantic Web for Research Communities), for all class members of PhDStudent, we find among the results several post-doctoral researchers like Philipp Cimiano. Since Philipp's PhD thesis is recorded properly in SWRC, the ontology intuitively provides enough evidence for him *not* being a PhD student anymore. However, the necessary background knowledge needed to logically infer this piece of information is obviously missing, and a class Postdoc does not exist in SWRC.

In general, it is not always obvious which kind of knowledge is required, nor to formalize it in a way that it is consistent with the existing knowledge base. In our case, a person's position within the academic staff – e.g., postdoctoral researcher, PhD student or undergraduate – is strongly correlated with his or her authorship properties. Hence, it seems sensible to add more background knowledge specifying which publications and authors can take part in an authorship relation. We will refer to this kind of binary relations as *roles*.

In the following, we propose a systematic methodology based on a reasoner-aided approach to ontology acquisition and refinement, which combines techniques from natural language processing (NLP) with formal concept analysis (FCA). We start by elaborating on our conceptual framework that integrates the core components of our implementation (Section 2). Preliminaries – our ontology model and basics of FCA – are introduced in Section 3, where we also describe the refinement of complex class descriptions and the acquisition of missing subsumption relationships. Section 4 introduces our approach to a semi-automatic specification of complex domain-range restrictions as well as its implementation in the RoLExO application. A detailed example in Section 5 illustrates the integrated use of our method within the overall process of ontology refinement. Section 6 concludes with some related and future work.

As we will see, our methodic approach to ontology enrichment can help to increase the expressive power of any ontology on the Web. By enabling ontology-based applications to draw additional conclusions about previously hidden class memberships while at the same time revealing undesired logical implications, it fosters Web and Intranet Intelligence.

2 Ontology Refinement Process

The overall process of semi-automatic ontology refinement, that we envision for future ontology engineering environments, consists of four steps.

Step 1: *Selection of a role to be refined.* In the first step the user, possibly assisted by automatically computed heuristics, selects a role whose domain and range characteristics are to be refined.

Step 2: *Specification of relevant class descriptions.* The user is asked to name a set of classes that she considers

*Research reported in this paper has been partially financed by the EU under the IST-2006-027595 project NeOn and by the Deutsche Forschungsgemeinschaft (DFG) under the projects ReaSem and Multipla.

relevant for the domain (or range, respectively) of the role. These classes (like PhDStudent or Person) constitute the focus for the subsequent exploration process (see *Step 3* and *Step 4*). If any of these classes is not yet specified in the ontology, the user is prompted to enter a natural language definition for it (e.g. "A *postdoc* is a graduate who has written a doctoral thesis."). An ontology learning component automatically transforms this informal definition into a complex class description, whose parts are mapped to already existing, atomic classes or roles.

Step 3: *Refinement of domain and range.* Once the focus of the exploration (i.e. the classes most relevant for domain or range, respectively) has been defined, the user is recommended to complete the ontology with respect to these classes.[1] For this purpose, she is guided through an interrogation process in the course of which the system asks smart[2] questions with respect to the domain (e.g. "Can an article be a PhD thesis?"). By answering them and occasionally giving counterexamples the user efficiently acquires missing subsumption and disjointness relationships.

Step 4: *Acquisition of role restrictions.* In the final step of the ontology refinement process, the user is supported in specifying complex domain-range restrictions for a given role (e.g. author). Again, she is guided by a methodic interrogation, whose implementation (cf. Section 4) relies on a reasoner for minimizing the user involvement. The result of this phase is a set of axioms that constrain the allowed usage of the initially selected role.

The advantages of our approach are manifold:

Integration of lexical and logical approaches to knowledge acquisition. The NLP-based ontology learning component assists the user in formalizing her knowledge and increasing the expressivity of the ontology by adding complex class descriptions (*Step 2*). The subsequent exploration of domain and range (*Step 3*), helps to clarify previously underspecified logical dependencies, thereby integrating the newly acquired axioms into the ontology. Implementations of additional, automatic experts allow for a seamless integration of supplementary lexical or logical ontology acquisition methods into the engineering process.

Efficient acquisition of logically complete knowledge. Relational exploration of classes and role restrictions guarantees completeness,[3] while at the same time minimizing the human modeling effort. Guiding the user through the ontology refinement process by asking smart questions, the system enforces modeling decisions that might have otherwise been ignored though being important for an appropriate representation of the user's domain knowledge. Since it is most often easier for humans to give concrete examples than to come up with abstract axioms, our approach facilitates the acquisition of even complex subsumption relationships (*Step 3*) and generalized domain-range restrictions (*Step 4*). The reasoner support underlying our implementation reduces the amount of user interaction required in the exploration process even further and ensures logical consistency of the ontology.

Interactive evaluation of learned or manually engineered ontologies. The systematic exploration of class and role extension relationships helps to detect underspecified logical dependencies. At the same time, modeling errors and wrong facts are relentlessly revealed by unforseen questions and automatically retrieved, erroneous counterexamples. By interactively refining the ontology, the user can effectively determine whether a given formalization of domain knowledge matches her conceptualization.

Open-source Implementation. Our implementation is open-source and publicly available under the LGPL license. Sources, binaries and examples can be downloaded from a dedicated Web page.[4]

3 Preliminaries

In what follows, we will introduce the preliminaries of knowledge specification in the description logic $\mathcal{SHOIN}$ as well as the basic notions of formal concept analysis (FCA), a mathematical discipline also dealing with conceptual knowledge specification.

Knowledge Specification in $\mathcal{SHOIN}$. The description logic $\mathcal{SHOIN}$ serves as the theoretical basis for the Web Ontology Language OWL DL as defined in [14]. For a thorough treatise on the rich field of description logics, see [2].

A $\mathcal{SHOIN}$ knowledge base (KB, also: ontology) is based on sets N_R (*role names*) **C** (*atomic classes*) and **I** (*individuals*).[5] In the following, we leave this vocabulary implicit and assume that A, B are atomic classes, a, b, i are individuals, and R, S are roles. Those can be used to define class descriptions employing the constructors from the upper part of Table 1. We use C, D to denote class descriptions. Moreover, a $\mathcal{SHOIN}$ KB consists of two finite sets of axioms that are referred to as *TBox* and *ABox*. The possible axiom types for each are displayed in the lower part of Table 1.[6] We use the common model-theoretic seman-

[1] This step is optional, but it significantly facilitates and shortens the acquisition of domain-range restrictions that follows (*Step 4*).

[2] These questions are smart in a sense that they are non-redundant and will be posed to the user only if they cannot be answered automatically by a reasoner or ontology learning component.

[3] Full completeness is not mandatory as the user may quit the exploration at any point of time, e.g., if she is tired of answering questions or already satisfied with the results. In this case, she can either stay with the partially refined ontology or resume the exploration later on.

[4] `http://relexo.ontoware.org`

[5] In DL settings one usually speaks of *concepts* and *roles*, the synonym terms used in OWL are *classes* and *properties*. For clarity, we will consequently use *classes* but *roles*.

[6] As usual, we require to restrict number restrictions to simple roles, being (roughly speaking and omitting further technical details) roles without

Name	Syntax	Semantics	
inverse role	R^-	$\{(x,y) \mid (y,x) \in R^{\mathcal{I}}\}$	
top	$\top$	Δ	
bottom	$\bot$	$\emptyset$	
nominal	$\{i\}$	$\{i^{\mathcal{I}}\}$	
negation	$\neg C$	$\Delta \setminus C^{\mathcal{I}}$	
conjunction	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$	
disjunction	$C \sqcup D$	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$	
universal restriction	$\forall R.C$	$\{x \mid (x,y) \in R^{\mathcal{I}} \text{ implies } y \in C^{\mathcal{I}}\}$	
existential restriction	$\exists R.C$	$\{x \mid \text{for some } y \in \Delta,\ (x,y) \in R^{\mathcal{I}}, y \in C^{\mathcal{I}}\}$	
(unqualified) number	$\leq n\, R$	$\{x \mid \#\{y \in \Delta \mid (x,y) \in R^{\mathcal{I}}\} \leq n\}$	
restriction	$\geq n\, R$	$\{x \mid \#\{y \in \Delta \mid (x,y) \in R^{\mathcal{I}}\} \geq n\}$	
role inclusion	$S \sqsubseteq R$	$S^{\mathcal{I}} \subseteq R^{\mathcal{I}}$	TBox
transitivity	$\mathsf{Trans}(S)$	$S^{\mathcal{I}}$ is transitive	TBox
general class inclusion	$C \sqsubseteq D$	$C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$	TBox
class membership	$C(a)$	$a^{\mathcal{I}} \in C^{\mathcal{I}}$	ABox
role membership	$R(a,b)$	$(a^{\mathcal{I}}, b^{\mathcal{I}}) \in R^{\mathcal{I}}$	ABox

Table 1. Concept constructors and axioms in $\mathcal{SHOIN}$.

tics for $\mathcal{SHOIN}$: an interpretation $\mathcal{I}$ consists of a set Δ called *domain* together with a function $\cdot^{\mathcal{I}}$ mapping individual names to elements of Δ, class names to subsets of Δ, and role names to subsets of $\Delta \times \Delta$. This function is then extended to complex expressions and axioms (cf. Table 1).

The following tiny examples illustrate how knowledge specification in $\mathcal{SHOIN}$ works: The fact that Sydney is an Australian city could be expressed by $\mathsf{City} \sqcap \mathsf{Australian}(\mathsf{Sydney})$, whereas Nicole being an inhabitant of Sydney can be represented as $\mathsf{inhabitantOf}(\mathsf{Nicole}, \mathsf{Sydney})$. We can further model the proposition that all Australian state capitals are located at the coast ($\mathsf{Australian} \sqcap \mathsf{State\_capital} \sqsubseteq \exists \mathsf{locatedAt}.\mathsf{Coast}$) and make the claim that all Australian cities have only polite and beautiful citizens: $\mathsf{Australian} \sqcap \mathsf{City} \sqsubseteq \forall \mathsf{inhabitantOf}^-.(\mathsf{Polite} \sqcap \mathsf{Beautiful})$. Note the knowledge increase obtainable by this role restriction! Given the above facts about Sydney and Nicole, one can derive the class memberships $\mathsf{Polite}(\mathsf{Nicole})$ and $\mathsf{Beautiful}(\mathsf{Nicole})$.

M

G	east_coast	south_cost	pop3000000+	founded_1800+
Sydney	×		×	
Melbourne		×	×	×
Brisbane	×			×
Perth				×
Adelaide		×		×

Figure 1. Example context.

Formal Concept Analysis (FCA). The primal notion of formal concept analysis (see [12]) is that of a *formal context* $\mathbb{K} := (G, M, I)$, which formally consists of two sets G (objects) and M (attributes) as well as an incidence relation $I \subseteq G \times M$. Intuitively, it can be conceived as a cross table indicating whether an object has an attribute as depicted by Figure 1.

An important means of expressing information in FCA is via *attribute implications*, written as $A \rightarrow B$ where A and B are sets of attributes: $A \rightarrow B$ is valid in a given formal context, if every object that has all attributes from A also has all attributes from B. In our example, $\{\text{founded\_1800+}, \text{pop3000000+}\} \rightarrow \{\text{south\_cost}\}$ would be a valid implication, expressing that all Australian cities established after 1800 with more than 3 million inhabitants are situated at the south cost.

The method of *attribute exploration* [11] tackles the problem of determining all implications valid in a formal context which might be still partially unknown to the computer (but is completely known by a human expert). Essentially, the algorithm enumerates non-redundant, hypothetic implications and asks the human expert for their validity in a domain of interest. The expert then has to decide: if the implication in question is valid, it will be added to an implication set called the *stem base*. If not, the expert has to provide an object (the so-called *counterexample*) that refutes the hypothesis. In the end, after this algorithm has terminated, the acquired knowledge (consisting of implications and counterexamples) is complete in the following sense: every implication is either a logical consequence of the stem base or there is a recorded counterexample for it.

Explorative Refinement of Class Interdependencies. Comparing the knowledge representation approaches of DLs and FCA, one finds considerable similarities: FCA objects can be conceived as individuals of a KB, while FCA attributes resemble DL classes with the formal context displaying class memberships. In particular, for a given set of DL class descriptions C and a DL interpretation $\mathcal{I} = (\Delta, \cdot^{\mathcal{I}})$, one can construct the formal context $\mathbb{K}_{\mathcal{I},C} := (\Delta, C, \models)$ (where $\delta \models \mathsf{C}$ denotes $\delta \in \mathsf{C}^{\mathcal{I}}$). In this context, an arbitrary implication $\mathsf{C}_1, \ldots \mathsf{C}_n \rightarrow \mathsf{D}_1, \ldots \mathsf{D}_m$ is valid exactly if the class subsumption axiom $\mathsf{C}_1 \sqcap \ldots \sqcap \mathsf{C}_n \sqsubseteq \mathsf{D}_1 \sqcap \ldots \sqcap \mathsf{D}_m$ is valid in $\mathcal{I}$ (cf. [18], Theorem 4.2). Hence, attribute exploration techniques can be used for the interactive refinement of an ontology's class interdependencies. Existing approaches [1, 17, 18, 3, 23] mainly differ in the supported logical fragment to be explored and the opportunity of providing only partially specified counterexamples.

This refinement of class hierarchies, more specifically, the acquisition of missing subsumption relationships (*Step 2* and *3* in Section 2), is supported by our previous implementation of reasoner-aided relational exploration with partial contexts [23]. In the following, we will therefore concentrate on the more ambitious goal of acquiring fine-grained role restrictions (*Step 4*) as motivated in the introduction.

4 Acquisition of Role Restrictions

As mentioned earlier, most exploration methods focus on the interdependencies of class descriptions, whereas roles merely occur as building blocks for complex class descriptions. However, as argued in our introductory example, it is worth focusing on specific roles and to model logical

transitive subroles.

interrelationships between its domain and range individuals, since those can entail valuable conclusions about class memberships.

Indeed, the use of the attribute exploration technique from formal concept analysis is not limited to the exploration of the conceptual hierarchy of class descriptions (and their conjunctions). Rather, by using other mappings from interpretations to formal contexts, it is possible to explore different logical fragments. Recently, the fragment of generalized domain-range restrictions (GDRRs) has been proposed as a both interesting and intuitive fragment eligible for exploration [19].

Given a set C of named classes and a role R, a *generalized domain-range restriction* (short: GDRR) is a rule

$$\mathsf{R}(X,Y) \wedge \bigwedge_{\mathsf{A}\in\mathbf{A}} \mathsf{A}(X) \wedge \bigwedge_{\mathsf{B}\in\mathbf{B}} \mathsf{B}(Y) \rightarrow \bigwedge_{\mathsf{C}\in\mathbf{C}} \mathsf{C}(X) \wedge \bigwedge_{\mathsf{D}\in\mathbf{D}} \mathsf{D}(Y) \qquad (*)$$

where $\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D} \subseteq C$ and R is a role name. Note that for $\mathbf{C}\cup\mathbf{D} = \emptyset$, the rule will have an empty head (also denoted by $\square$) and hence, will be interpreted as integrity constraint. So, the GDRR presented in the above definition would mean the following: "*For any two elements X and Y of the domain of interest that are connected by a role* R *and where X fulfills (all of)* $\mathbf{A}$ *as well as Y fulfills (all of)* $\mathbf{B}$, *we know that X additionally fulfills* $\mathbf{C}$ *and Y additionally fulfills* $\mathbf{D}$."

It has been proven that every GDRR can be equivalently expressed by a DL axiom. The rule scheme (*) corresponds to the axiom:

$$\bigsqcap_{\mathsf{A}\in\mathbf{A}} \mathsf{A} \sqcap \exists\mathsf{R}.\Big(\bigsqcap_{\mathsf{B}\in\mathbf{B}} \mathsf{B}\Big) \sqsubseteq \bigsqcap_{\mathsf{C}\in\mathbf{C}} \mathsf{C} \sqcap \forall\mathsf{R}.\Big(\Big(\bigsqcup_{\mathsf{B}\in\mathbf{B}} \neg\mathsf{B}\Big) \sqcup \Big(\bigsqcap_{\mathsf{D}\in\mathbf{D}} \mathsf{D}\Big)\Big).$$

GDRRs allow for the specification of semantic ramifications caused by the presence of a specific role between two individuals (some of whose class memberships are also known). Exploring the fragment of GDRRs therefore enables a novel role-focused way of ontology refinement.

Technically this is done – as opposed to all other current exploration methods for ontologies – by exploring a formal context whose objects are *not* single individuals from the knowledge base (like Sydney, see Figure 1), but *pairs* of individuals which instantiate the role R. Consequently, also the counterexamples are individual pairs.

For a given interpretation $\mathcal{I}$ together with two sets C_{domain} and C_{range} of named classes and a role R, the *role context* $\mathbb{K}_{\mathsf{R}}$ is defined as formal context (G, M, I) with

- $G := \mathsf{R}^{\mathcal{I}} = \{(\delta_1,\delta_2) \mid \delta_1,\delta_2 \in \Delta, (\delta_1,\delta_2) \in \mathsf{R}^{\mathcal{I}}\}$
- $M := \{\mathsf{C}_1 \mid \mathsf{C} \in C_{\text{domain}}\} \cup \{\mathsf{C}_2 \mid \mathsf{C} \in C_{\text{range}}\} \cup \{\perp\}$
- $I \subseteq G \times M$ with $(\delta_1,\delta_2) \not{I} \perp$ and $(\delta_1,\delta_2) I \mathsf{C}_1 \iff \delta_1 \in \mathsf{C}^{\mathcal{I}}$ as well as $(\delta_1,\delta_2) I \mathsf{C}_2 \iff \delta_2 \in \mathsf{C}^{\mathcal{I}}$

The following theorem shows how the validity of a GDRR in an interpretation can be read from a corresponding role context.

Theorem 1 An interpretation $\mathcal{I}$ satisfies a GDRR of the shape (*) if and only if the corresponding role context $\mathbb{K}_{\mathsf{R}}$ satisfies the implication

$$\{\mathsf{A}_{\mathrm{d}} \mid \mathsf{A} \in \mathbf{A}\} \cup \{\mathsf{B}_{\mathrm{r}} \mid \mathsf{B} \in \mathbf{B}\} \rightarrow \begin{cases} \perp & \text{if } \mathbf{C} \cup \mathbf{D} = \emptyset, \text{otherwise:} \\ \{\mathsf{C}_{\mathrm{d}} \mid \mathsf{C} \in \mathbf{C}\} \cup \{\mathsf{D}_{\mathrm{r}} \mid \mathsf{D} \in \mathbf{D}\}. & \end{cases}$$

This theorem enables us to "translate" any implication in a role context into an equivalent GDRR and via the above-mentioned correspondence further into a DL axiom.

Now, the basic idea for the knowledge acquisition method we are going to propose is to carry out attribute exploration on the context $\mathbb{K}_{\mathsf{R}}$. Thereby, our basic assumption is that there exists a distinguished interpretation $\mathcal{I}'$ entirely (but implicitly) known by the human expert that we want to specify in terms of GDRRs.

After role exploration, the updated knowledge base is *complete* in the following sense: any GDRR (referring to R, C_{domain} and C_{range}) can either be deduced from the updated knowledge base or there is a pair of individuals in the ontology witnessing that this GDRR does not hold.

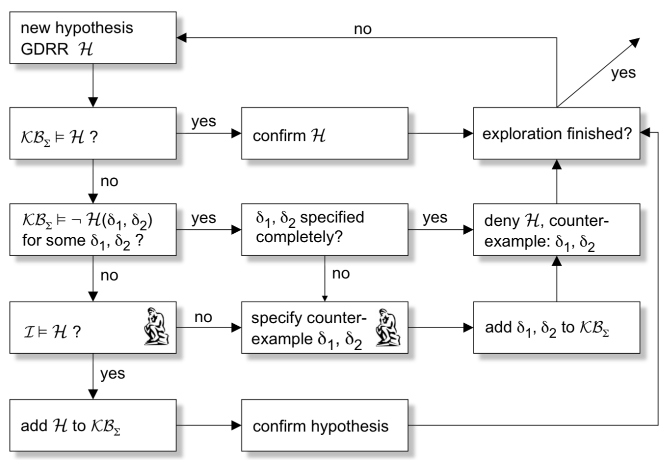

Figure 2. RoLExO Implementation. "The Thinker" indicates user involvement.

We instantiated our approach by implementing RoLExO[7] (*Role Exploration for Learning Expressive Ontologies*), an interactive application supporting the acquisition of complex role restrictions (cf. *Step 4* in Section 2). The architecture of RoLExO relies upon KAON2[8] as an ontology management back-end and features a simple graphical user interface. The ontology refinement process, depicted by Figure 2, is handled by a role exploration component which manages a partial context and an implication set. Both are updated based on answers obtained from the "expert team" constituted by a KAON2 reasoner and the human knowledge engineer.

[7] http://relexo.ontoware.org
[8] http://kaon2.semanticweb.org

5 Example Scenario

We now illustrate the practical relevance of our approach to ontology refinement by means of a real-world example.

Ontology. The SWRC (Semantic Web for Research Communities)[9] [21] ontology is a well-known ontology modeling the domain of Semantic Web research. Version 0.3 containing 55 classes as well as 41 roles serves as a basis for the AIFB Web portal[10] which manages information about 2,982 persons, projects, and publications. For our experiment, we exported all the instance data stored in the AIFB portal into one single OWL file (more than 5 MB in RDF syntax), and merged it with the corresponding TBox, i.e. the imported version of SWRC. This way, we obtained an ontology with 33,426 axioms.

Refinement Process. As motivated in Section 1, we first decide to refine the author role of the SWRC ontology (*Step 1*, Section 2). Subsequently, we select several classes that we assumed to be relevant for the domain and range of this role (*Step 2*), thereby noticing that an equivalent to "postdoctoral researcher" is missing in the ontology. After some discussion about our understanding of this class, we enter the following definition into LExO [22].

"*A postdoc is a graduate who has written a doctoral thesis.*"

The result is a set of axioms representing a formal description of the class a_postdoc.[11]

$\mathsf{a\_postdoc} \equiv \mathsf{a\_graduate\_who\_has\_written\_a\_doctoral\_thesis}$

$\mathsf{a\_graduate\_who\_has\_written\_a\_doctoral\_thesis}$
$\quad \equiv \mathsf{a\_graduate} \sqcap \mathsf{has\_written\_a\_doctoral\_thesis}$

$\mathsf{has\_written\_a\_doctoral\_thesis} \equiv \exists \mathsf{has\_written.a\_doctoral\_thesis}$

$\mathsf{a\_doctoral\_thesis} \equiv \mathsf{a\_thesis} \sqcap \mathsf{doctoral}$

Since some of the atomic classes obviously map to existing classes in SWRC, we further add the following mapping axioms: $\mathsf{swrc{:}Thesis} \equiv \mathsf{lexo{:}thesis}$, $\mathsf{swrc{:}Graduate} \equiv \mathsf{lexo{:}a\_graduate}$, $\mathsf{swrc{:}PhDThesis} \equiv \mathsf{lexo{:}a\_thesis} \sqcap \mathsf{lexo{:}doctoral}$ and $\mathsf{lexo{:}has\_written} \sqsubseteq \mathsf{swrc{:}author}^-$. The resulting, extended version of SWRC is input to the exploration process described in the sequel.

Refinement of Domain and Range. We only briefly display the results of the two relational exploration processes dedicated to the domain and range of the considered author role (cf. Section 2, *Step 3*). For a more detailed description, we refer the reader to [23]. Altogether, the following 7 new axioms were acquired during the process:

$\mathsf{PhDThesis} \sqsubseteq \mathsf{Book}$	$\mathsf{Article} \sqcap \mathsf{Thesis} \sqsubseteq \bot$
$\mathsf{Book} \sqcap \mathsf{Thesis} \sqsubseteq \mathsf{PhDThesis}$	$\mathsf{Book} \sqcap \mathsf{MasterThesis} \sqsubseteq \bot$
$\mathsf{FullProfessor} \sqsubseteq \mathsf{a\_postdoc}$	$\mathsf{PhDStudent} \sqsubseteq \mathsf{Graduate}$
$\mathsf{has\_written\_a\_doctoral\_thesis} \sqsubseteq \mathsf{Graduate}$	

[9] http://ontoware.org/projects/swrc/
[10] http://www.aifb.uni-karlsruhe.de
[11] $\mathsf{a\_postdoc} \equiv \mathsf{a\_graduate} \sqcap \exists \mathsf{has\_written.(a\_thesis} \sqcap \mathsf{doctoral)}$

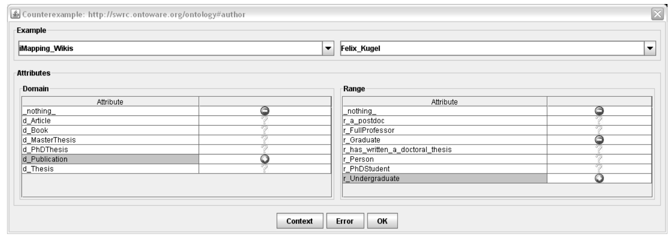

Figure 3. Entering a counterexample, i.e. a pair of individuals being in the authorship relation.

Acquisition of Role Restrictions. Subsequent to these preprocessing steps, we are now ready to carry out the role exploration focusing on the author role as well as the classes Article, Book, MasterThesis, PhDThesis, Publication, Thesis on the domain side and the classes a_postdoc, FullProfessor, Graduate, has_written_a_doctoral_thesis, Person, PhDStudent, Undergraduate on the range side (see *Step 4* in Section 2).

After a few trivial questions, that can be denied by automatically retrieved counterexamples, the first question requiring human interaction is posed. It asks for the validity of the rule $\mathsf{author}(X, Y) \to \mathsf{Publication}(X) \wedge \mathsf{Graduate}(Y) \wedge \mathsf{Person}(Y)$, investigating whether, if some X is authored by some Y, X must be a publication and Y must be both a graduate and a person. As there are no records in SWRC of any authors not being graduates, no counterexample can be brought up automatically. Instead, after denying this hypothesis, the human expert is asked to add a publication-author pair disproving it. We do so by adding information about an undergraduate's authorship (Fig. 3).

The next hypothesis brought up by the system, $\mathsf{author}(X, Y) \to \mathsf{Publication}(X) \wedge \mathsf{Person}(Y)$, can be clearly confirmed by the human expert as it gives straightforward domain and range restrictions for the author role: every author is a person and everything authored is a publication.

In the subsequent execution, the algorithm poses the GDRR $\mathsf{author}(X, Y) \wedge \mathsf{Book}(X) \to \mathsf{Graduate}(Y)$ expressing that everybody publishing a book must be a graduate. Though quite arguable in general, in the scientific area considered by us, we judge that this axiom constitutes a sensible restriction and therefore confirm it.

As opposed to the previous, the next presented potential axiom, $\mathsf{author}(X, Y) \wedge \mathsf{Thesis}(X) \to \mathsf{Graduate}(Y)$ can be accepted without much argument. Clearly everybody having written a thesis qualifies as a graduate – be it a master or a PhD thesis.

The last axiom in question, displayed in Figure 4, encodes that any thesis that has been written by somebody being now a PhD student must be a master thesis. We confirm this (as we assume any thesis to be either a PhD or a master thesis and furthermore, somebody having written a PhD thesis must not not be classified as a PhD student).

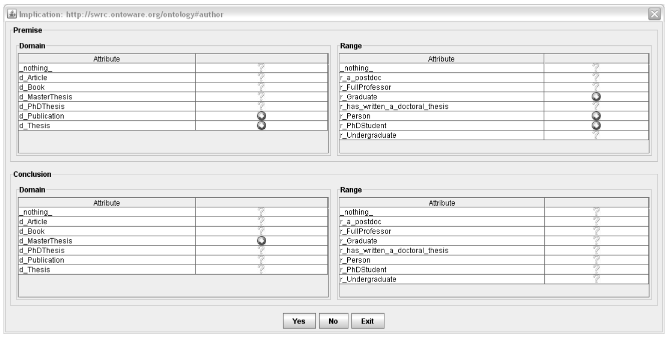

Figure 4. How a hypothetic GDRR is displayed.

In the end, we have acquired the following four new axioms that characterize how class memberships of domain and range individuals "influence" each other:

$$
\begin{aligned}
\exists\mathsf{author}.\top &\sqsubseteq \mathsf{Publication} \sqcap \forall\mathsf{author}.\mathsf{Person}\\
\mathsf{Book} \sqcap \exists\mathsf{author}.\top &\sqsubseteq \forall\mathsf{author}.\mathsf{Graduate}\\
\mathsf{Thesis} \sqcap \exists\mathsf{author}.\top &\sqsubseteq \forall\mathsf{author}.\mathsf{Graduate}\\
\mathsf{Thesis} \sqcap \exists\mathsf{author}.\mathsf{PhDStudent} &\sqsubseteq \mathsf{MasterThesis}.
\end{aligned}
$$

The fact that this constitutes a rather small set of axioms is thanks to the preprocessing via domain and range exploration. Otherwise, numerous additional questions actually relating to domain- or range-inherent dependencies would have been brought up. By proceeding as we propose, the knowledge acquired in this exploration step only incorporates logical ramifications where both domain and range are involved. A statistical summary of the entire exploration process is given in Table 2.

	domain	range	RoLExO	$\sum$
Reasoner-answered questions	9	8	19	36
Questions decided by human	6	5	13	24
New TBox axioms	5	3	4	12
New individuals	1	2	14	17
New class memberships $C(a)$	18	27	36	81
New role memberships $R(a, b)$	0	0	7	7

Table 2. Summary of the exploration process.

Revisiting our initially described problems, we find them solved. It is now possible to correctly classify our colleague, Philipp Cimiano, (and others) based on ontological knowledge: The fact that Philipp is author of an individual classified as a PhD thesis was originally stated in SWRC. Now, from the knowledge acquired via LExO from natural language resources defining the notion of a postdoc (together with the introduced mappings), we can infer that Philipp is indeed a postdoc. Likewise, thanks to the GDRR $\mathsf{Thesis} \sqcap \exists\mathsf{author}.\top \sqsubseteq \forall\mathsf{author}.\mathsf{Graduate}$, every person specified as being the author of a master thesis will from now on be inferred to be a graduate. Hence, by employing a combination of ontology learning and logical exploration techniques we have considerably increased the knowledge inferrable from the SWRC ontology (cf. Table 3).

6 Related Work and Conclusion

The incremental refinement of ontologies has become an important issue that is addressed by several ontology learning tools and methodologies such as Text2Onto [8] or the work by Navigli and Velardi [15] (for an overview, see also [4]). However, only few of them support the acquisition of logically complex axioms, and most lexically inspired techniques lack a proper integration with reasoning-based methods for ontology evaluation or debugging. This also holds for approaches to learning domain-range restrictions of binary relations [6, 5]. Parallel to these more or less NLP-based ontology learning methods, approaches relying on Inductive Logic Programming (ILP) [13, 10, 9] and Formal Concept Analysis (FCA) [18] have been developed in the logics community. But although there are a few approaches, similar to ours, aiming to reconcile the two worlds of lexical and logical ontology acquisition by either FCA [20, 7] or ILP [16], none of them has been designed specifically for refining OWL DL ontologies.

	SWRC original	SWRC refined
Article	189	190
Book	36	95
MasterThesis	0	4
PhDThesis	58	59
Publication	1499	1507
Thesis	58	63
a_postdoc	0	67
FullProfessor	6	9
Graduate	52	139
has_written_a... _doctoral_thesis	0	67
Person	1213	1222
PhDStudent[12]	50	47
Undergraduate	6	9
$\sum$	3167	3478

Table 3. Inferred class memberships.

In this paper, we have therefore sketched a way to combine two complementary approaches to the acquisition and refinement of OWL DL ontologies: the more intensional approach of distilling conceptual information from lexical resources and the extensional method of extracting hypothetical axioms from manually specified or automatically retrieved domain entities. We have instantiated our approach by designing and implementing a framework that integrates Relational Exploration and Role Exploration – two techniques for the systematic refinement of OWL ontologies based on FCA. To the best of our knowledge, RoLExO is the first publicly available implementation of an exploration-based approach to ontology refinement. In an example using the well-known SWRC ontology we have demonstrated the feasibility of semi-automatic ontology refinement, and its applicability to real-world ontology engineering tasks.

After all, we identify several issues for future research.

[12] Exploring the range of author for the first time, we noticed a modeling error in SWRC. It was the second question, automatically answered by the reasoner, which brought up the counterexample Peter Haase. A simple query to SWRC revealed that he was explicitly stated to be a PhDStudent and inferred to be a_postdoc. Since we assumed both classes to be disjoint, we changed the explicit classification of Peter and two other postdocs to Person and restarted the exploration with the corrected ontology.

Firstly, we will further extend RoLExO by an additional automatic expert which uses ontology learning techniques or Web resources, including other ontologies, for confirming hypotheses and suggesting counterexamples. Moreover, FCA-based exploration can be applied to a set of classes from multiple distributed knowledge bases, in order to clarify logical dependencies across ontologies. Hence our approach can be extended to support Web ontology alignment and integration. Finally, we will integrate RoLExO into an ontology engineering environment such as the NeOn Toolkit[13] and improve its usability by adding a natural language generation component for translating hypotheses, i.e. logical implications, into natural language questions. In the end, we are confident that our planned user studies will demonstrate the advantages of interactive ontology refinement and we hope that our approach will initiate the development of new tools for enhancing Web Intelligence.

References

[1] F. Baader. Computing a minimal representation of the subsumption lattice of all conjunctions of concepts defined in a terminology. In *Proc. Int. Symposium on Knowledge Retrieval, Use, and Storage for Efficiency (KRUSE)*, pages 168–178, Santa Cruz, USA, 1995.

[2] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[3] F. Baader, B. Ganter, B. Sertkaya, and U. Sattler. Completing description logic knowledge bases using formal concept analysis. In M. M. Veloso, editor, *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 230–235, 2007.

[4] P. Buitelaar and P. Cimiano, editors. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, volume 167 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, Amsterdam, JAN 2008.

[5] M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric, and I. Rojas. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In L. P. Kaelbling and A. Saffiotti, editors, *IJCAI*, pages 659–664. Professional Book Center, 2005.

[6] P. Cimiano, M. Hartung, and E. Ratsch. Finding the appropriate generalization level for binary relations extracted from the genia corpus. In *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, pages 161–169. ELRA, MAY 2006.

[7] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *JAIR*, 24:305–339, AUG 2005.

[8] P. Cimiano and J. Völker. Text2Onto - a framework for ontology learning and data-driven change discovery. In *Proc. 10th Int. Conf. on Applications of Natural Language to Information Systems*, pages 227–238. Springer, JUN 2005.

[13]`http://www.neon-toolkit.org`

[9] W. W. Cohen and H. Hirsh. Learning the classic description logic: Theoretical and experimental results. In J. Doyle, E. Sandewall, and P. Torasso, editors, *Proc. of the 4th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR)*, pages 121–133. Morgan Kaufmann, 1994.

[10] N. Fanizzi, L. Iannone, I. Palmisano, and G. Semeraro. Concept formation in expressive description logics. In *Proc. 15th European Conf. on Machine Learning (ECML)*. Springer Verlag, 2004.

[11] B. Ganter. Two basic algorithms in concept analysis. Technical Report 831, FB4, TH Darmstadt, 1984.

[12] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1997.

[13] F. A. Lisi and F. Esposito. Ilp meets knowledge engineering: A case study. In S. Kramer and B. Pfahringer, editors, *ILP*, volume 3625 of *LNCS*, pages 209–226. Springer, 2005.

[14] D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview. `CTAN:http://www.w3.org/TR/2004/REC-owl-features-20040210/`, FEB 2004.

[15] R. Navigli and P. Velardi. Ontology enrichment through automatic semantic annotation of on-line glossaries. In S. Staab and V. Svátek, editors, *Proc. 15th Int. Conf. on Knowledge Engineering and Knowledge Management (EKAW)*, volume 4248 of *LNCS*, pages 126–140. Springer, 2006.

[16] C. Nedellec. Corpus-based learning of semantic relations by the ILP system, Asium. In J. Cussens and S. Dzeroski, editors, *Learning Language in Logic*, volume 1925 of *LNCS*, pages 259–278. Springer, 1999.

[17] S. Rudolph. Exploring relational structures via FLE. In K. E. Wolff, H. D. Pfeiffer, and H. S. Delugach, editors, *Proc. 12th Int. Conf. on Conceptual Structures (ICCS)*, volume 3127 of *LNCS*, pages 196 – 212. Springer, JUL 2004.

[18] S. Rudolph. *Relational Exploration - Combining Description Logics and Formal Concept Analysis for Knowledge Specification*. Universitätsverlag Karlsruhe, 2006. Dissertation.

[19] S. Rudolph. Acquiring generalized domain-range restrictions. In R. Medina and S. Obiedkov, editors, *Proc. 6th Int. Conf. on Formal Concept Analysis (ICFCA)*, volume 4933 of *LNAI*, pages 32–45. Springer, FEB 2008.

[20] G. Stumme and A. Maedche. FCA-merge: Bottom-up merging of ontologies. In *Proc. 17th Int. Conf. on Artificial Intelligence (IJCAI)*, pages 225–230, 2001.

[21] Y. Sure, S. Bloehdorn, P. Haase, J. Hartmann, and D. Oberle. The SWRC ontology - semantic web for research communities. In C. Bento, A. Cardoso, and G. Dias, editors, *Proc. 12th Portuguese Conf. on Artificial Intelligence*, pages 218 – 231. Springer, DEC 2005.

[22] J. Völker, P. Hitzler, and P. Cimiano. Acquisition of OWL DL axioms from lexical resources. In *Proc. 4th European Semantic Web Conf. (ESWC)*. Springer, JUN 2007.

[23] J. Völker and S. Rudolph. Lexico-logical acquisition of OWL DL axioms - an integrated approach to ontology refinement. In R. Medina and S. Obiedkov, editors, *Proceedings of the 6th Int. Conf. on Formal Concept Analysis (ICFCA)*, volume 4933 of *LNAI*, pages 62–77. Springer, FEB 2008.