Extract ABox Modules for Efficient Ontology Querying

Jia Xu\textsuperscript{a,}\textsuperscript{*}, Ubbo Visser\textsuperscript{b}, Patrick Shironoshita\textsuperscript{a}, Nigel John\textsuperscript{a}, Mansur Kabuka\textsuperscript{a}

\textsuperscript{a}Department of Electrical and Computer Engineering, \\textsuperscript{b}Department of Computer Science, \\
University of Miami, Coral Gables, FL 33146, USA

Abstract
Extraction of logically-independent fragments out of an ontology ABox can be useful for solving the tractability problem of querying ontologies with large ABoxes. In this paper, we propose a formal definition of an ABox module, such that it guarantees complete preservation of facts of a given set of individuals, and thus can be reasoned independently w.r.t. the ontology TBox. With ABox modules of this type, isolated or distributed (parallel) ABox reasoning becomes feasible, and much more efficient data retrieving from ontology ABoxes can be expected. To compute such an ABox module, we present a theoretical approach and an approximation for SHIQ ontologies. Testing the approximation on different types of ontologies shows that our method is efficient and extracted ABox modules are significantly smaller than the entire ABox in average.

Keywords: 
Ontology, Reasoning, ABox Module, SHIQ

1. Introduction

Description logics (DLs), as a decidable fragment of first-order logic, are best known as a family of logic based knowledge representation formalisms (Baader et al., 2007). A DL ontology (or knowledge base) consists of a terminological part (TBox) \( T \) that defines terminologies such as concepts and roles of a given domain, and an assertional part (ABox) \( A \) that describes instances of the conceptual knowledge. Similar to a database, the ontology TBox usually represents the data schema, and the ontology ABox corresponds to the actual data set.

Standard DL reasoning services include subsumption test (i.e. testing if one concept is more general than the other) and instance checking (i.e. checking if an individual is one instance of a given concept). The former is considered TBox reasoning, and the latter is considered ABox reasoning as well as the central reasoning task for information retrieving from ontology ABoxes (Schaerf, 1994).

\textsuperscript{*}Corresponding author.

\textit{Email addresses:} j.xu11@umiami.edu (Jia Xu), visser@cs.miami.edu (Ubbo Visser), patrick@infotechsoft.com (Patrick Shironoshita), nigel.john@miami.edu (Nigel John), m.kabuka@miami.edu (Mansur Kabuka)

Preprint submitted to arXiv July 19, 2013
These reasoning services, however, are highly complex tasks, especially for ontologies with an expressive DL, and can have up to exponential worst-case complexity w.r.t. the size of the ontology (Baader et al., 2007).

Despite highly optimized implementation of reasoning algorithms in existing systems, such as HermiT (Motik et al., 2007), Pellet (Sirin et al., 2007), FaCT++ (Tsarkov and Horrocks, 2006), and Racer (Haarslev and Möller, 2001), they are still confronted with the scalability problem in practical DL applications. This is simply because in most practical DL applications, though the TBox could be relatively small and manageable, the ABox tends to be extremely large, which thereby could lead to severe tractability problems (Horrocks et al., 2000).

In this paper, conceiving that a large ABox may consist of data with great diversity and isolation, we think of exploiting modularity of an ontology ABox and expect ABox reasoning to be optimized by utilizing logical properties of ABox modules. More specifically, when an ABox contains data that is not tightly tangled, it is possible to divide the ABox into logically-separated modules, such that each module can be reasoned independently w.r.t. \( T \) and should be (ideally) much smaller than the entire ABox.

Analogous to modularity defined for the ontology TBox (Cuenca Grau et al., 2007a,b), the notion of modularity for the ABox should also be based on the semantics of ontologies; and an ABox module should be a closure or encapsulation of logical implications for a given set of individuals (that is, it should capture all facts, both explicit and implicit, of the given entities).

For illustration, consider a real-world ontology that models and stores massive biomedical data in an ABox, and a query-answering system that is based on this ontology. A biomedical researcher may want to obtain information of a particular gene instance, say \texttt{CHRNA4}, to see if it is involved in any case of diseases and in what manner. To answer queries submitted to this ontology-based system, a reasoning procedure is normally invoked and applied to the ontology. Note however, given the fact that in this case the entire ontology ABox is extremely large, complete reasoning can be prohibitively expensive (Glimm et al., 2008; Ortiz et al., 2008), and it may take a lengthy period to reach any conclusions. Therefore, reasoning on a subject-related module (in this case, \texttt{CHRNA4}-related) would be preferable.

In particular, this \texttt{CHRNA4}-related ABox module should be a closure of all facts about the individual \texttt{CHRNA4}. Thereby, reasoning on this module w.r.t. to \( T \) can achieve the same conclusions about \texttt{CHRNA4} as if the reasoning were applied to the entire ontology. In addition, the reasoning time should be significantly decreased, provided the module is precise and much smaller than the entire ABox.

We therefore focus on the computation of such a precise ABox module, and our main contributions are:

1. a set of formal definitions about the ABox module so that the notion of ABox modularity can be well captured;
2. a theoretical approach for module extraction based on the idea of module-essential assertions;
3. a simple and tractable syntactic approximation to cope with the complexity for checking module-essential assertions using a DL-reasoner.

Additionally, we show evaluation of our approximation on different ontologies with large ABoxes, including those generated by existing benchmark tools and realistic ones that
are used in some biomedical projects.

2. Preliminaries

2.1. Description Logic \text{SHIQ}

Description Logic \text{SHIQ} is extended from the well-known logic \text{ALC} \cite{Schmidt-Schauß and Smolka1991}, with added supports for role hierarchies, inverse roles, transitive roles, and qualified number restrictions \cite{Horrocks et al.2000}. A \text{SHIQ} ontology defines a set \( \mathbf{R} \) of role names, a set \( \mathbf{I} \) of individual names, and a set \( \mathbf{C} \) of classes (or concepts).

**Definition 2.1** (\text{SHIQ}-Role). A \text{SHIQ}-role can be an atomic (named) one \( R \in \mathbf{R} \), or an inverse role \( R^- \) with \( R \in \mathbf{R} \). The complete role set in a \text{SHIQ} ontology is denoted \( \mathbf{R}^* = \mathbf{R} \cup \{ R^- | R \in \mathbf{R} \} \). To avoid \( R^{--} \), function \( \text{Inv}(\cdot) \) is defined, such that \( \text{Inv}(R) = R^- \) and \( \text{Inv}(R^-) = R \).

A role hierarchy \( H_R \) can be defined in an ontology by a set of role inclusion axioms, each of which is expressed in the form of \( R_1 \sqsubseteq R_2 \), with \( R_1, R_2 \in \mathbf{R}^* \). We call \( R_1 \) is a subrole of \( R_2 \), if \( R_1 \sqsubseteq R_2 \in H_R \) or there exist \( S_1, \ldots, S_n \in \mathbf{R}^* \) with \( R_1 \sqsubseteq S_1, S_1 \sqsubseteq S_2, \ldots, S_n \sqsubseteq R_2 \in H_R \), where \( \sqsubseteq \) is reflexive and transitive.

A role \( R \in \mathbf{R} \) is transitive, denoted \( \text{Trans}(R) \), if \( R \circ R \subseteq R \) or \( \text{Inv}(R) \circ \text{Inv}(R) \subseteq \text{Inv}(R) \). Finally, a role is called simple if it is neither transitive nor has any transitive subroles \cite{Baader et al.2007} \cite{Horrocks et al.2000}.

**Definition 2.2** (\text{SHIQ}-Class). A \text{SHIQ}-class is either an atomic (named) class or a complex one that can be defined using the following constructors recursively:

\[
\begin{align*}
C, D &:= A | \neg C | C \sqcap D | C \sqcup D | \forall R.C | \exists R.C | \leq nS.C | \geq nS.C
\end{align*}
\]

where \( A \) is an atomic class in \( \mathbf{C} \), \( R, S \in \mathbf{R}^* \) with \( S \) being simple, and \( n \) is a non-negative integer. Other constructors, namely \( \top \) (universal concept) and \( \bot \) (bottom concept), can be viewed as \( A \sqcup \neg A \) and \( A \sqcap \neg A \) respectively.

**Definition 2.3** (\text{SHIQ Ontology}). A \text{SHIQ} ontology is a tuple, denoted \( \mathcal{K} = (\mathcal{T}, \mathcal{A}) \), where \( \mathcal{T} \) is the terminology representing general knowledge of a specific domain, and \( \mathcal{A} \) is the assertional knowledge representing a particular state of the terminology.

The terminology \( \mathcal{T} \) of ontology \( \mathcal{K} \) is the disjoint union of a finite set of role inclusion axioms (i.e. \( R_1 \sqsubseteq R_2 \)) and a set of concept inclusion axioms in the form of \( C \equiv D \) and \( C \sqsubseteq D \). Statements \( C \sqsubseteq D \) are called general concept inclusion axioms (GCIs), and \( C \equiv D \) can be trivially converted into two GCIs as \( C \sqsubseteq D \) and \( D \sqsubseteq C \).

The assertional part \( \mathcal{A} \) of \( \mathcal{K} \) is also known as a \text{SHIQ}-ABox, consisting of a set of assertions (facts) about individuals, in the form of

\[
\begin{align*}
C(a) & \quad \text{class assertion} \\
R(a, b) & \quad \text{role or property assertion} \\
a \not\approx b & \quad \text{inequality assertion}
\end{align*}
\]

where \( C \in \mathbf{C}, R \in \mathbf{R}^* \), and \( a, b \in \mathbf{I} \).

Note that, explicit assertion of \( a \not\approx b \) is supported in \text{SHIQ}, while conversely, explicit assertion of equality, i.e. \( a \approx b \), is not allowed, since its realization relies on equivalence between nominals \cite{Baader et al.2007}, i.e. \{a\} \equiv \{b\} , which is illegal in \text{SHIQ}.
Definition 2.4 (SHIQ Semantics). Meaning of an ontology is given by an interpretation denoted \( \mathcal{I} = (\Delta^\mathcal{I}, \mathcal{I}) \), with \( \Delta^\mathcal{I} \) referred to as a world or domain and \( \mathcal{I} \) referred to as an interpretation function. This interpretation function \( \mathcal{I} \) maps:

- every atomic class \( A \in C \) to a set \( A^\mathcal{I} \subseteq \Delta^\mathcal{I} \),
- every individual \( a \in I \) to an element \( a^\mathcal{I} \in \Delta^\mathcal{I} \) and,
- every role \( R \in \mathcal{R} \) to a binary relation on the domain, i.e. \( R^\mathcal{I} \subseteq \Delta^\mathcal{I} \times \Delta^\mathcal{I} \).

Interpretation for other classes and inverse roles are given below:

\[
\begin{align*}
\top^\mathcal{I} & = \Delta^\mathcal{I} \\
\bot^\mathcal{I} & = \emptyset \\
\neg C^\mathcal{I} & = \Delta^\mathcal{I} \setminus C^\mathcal{I} \\
(R^-)^\mathcal{I} & = \{(y,x) \mid (x,y) \in R^\mathcal{I}\} \\
(C \cap D)^\mathcal{I} & = C^\mathcal{I} \cap D^\mathcal{I} \\
(C \cup D)^\mathcal{I} & = C^\mathcal{I} \cup D^\mathcal{I} \\
(\exists R.C)^\mathcal{I} & = \{x \mid \exists y.(x,y) \in R^\mathcal{I} \land y \in C^\mathcal{I}\} \\
(\forall R.C)^\mathcal{I} & = \{x \mid \forall y.(x,y) \in R^\mathcal{I} \land y \in C^\mathcal{I}\} \\
(\leq n R.C)^\mathcal{I} & = \{x \mid \{y \mid (x,y) \in R^\mathcal{I} \land y \in C^\mathcal{I}\} \leq n\} \\
(\geq n R.C)^\mathcal{I} & = \{x \mid \{y \mid (x,y) \in R^\mathcal{I} \land y \in C^\mathcal{I}\} \geq n\}
\end{align*}
\]

where \( |\cdot| \) represents the cardinality of a given set.

An interpretation \( \mathcal{I} \) satisfies an axiom \( \alpha : C \sqsubseteq D \), if \( C^\mathcal{I} \subseteq D^\mathcal{I} \), and it is called the model of axiom \( \alpha \). Interpretation \( \mathcal{I} \) satisfies an arbitrary axiom or assertion \( \alpha \):

\[
\begin{align*}
R_1 \sqsubseteq R_2 & \iff R_1^\mathcal{I} \subseteq R_2^\mathcal{I} \\
C(a) & \iff a^\mathcal{I} \in C^\mathcal{I} \\
R(a,b) & \iff (a^\mathcal{I}, b^\mathcal{I}) \in R^\mathcal{I} \\
a \neq b & \iff a^\mathcal{I} \neq b^\mathcal{I}
\end{align*}
\]

For an ontology \( \mathcal{K} \), if interpretation \( \mathcal{I} \) satisfies every axiom in \( \mathcal{K} \), \( \mathcal{I} \) is a model of \( \mathcal{K} \), written \( \mathcal{I} \models \mathcal{K} \). In turn, ontology \( \mathcal{K} \) is said satisfiable or consistent if it has at least one model; otherwise, it is unsatisfiable or inconsistent, and there exists at least one contradiction in \( \mathcal{K} \).

Definition 2.5 (Logic Entailment). Given ontology \( \mathcal{K} \) and an axiom \( \alpha : C \sqsubseteq D \), \( \alpha \) is called a logic entailment of \( \mathcal{K} \), denoted \( \mathcal{K} \models \alpha \), if \( \alpha \) is satisfied in every model of \( \mathcal{K} \).

Definition 2.6 (Instance checking). Given an ontology \( \mathcal{K} \), a SHIQ-class \( C \) and an individual \( a \in I \), instance checking is to test if \( \mathcal{K} \models C(a) \) holds.

Notice that instance checking is considered the central reasoning task for information retrieving from ontology ABoxes and a basic tool for more complex reasoning services [Schaerf, 1994]. Instance checking can also be viewed as “classification” of an individual that checks if an individual can be classified into some defined DL class.
2.2. Other Definitions

We adopt notations from tableaux (Horrocks et al., 2000) for referring to individuals in \( R(a, b) \), such that \( a \) is called an \( R \)-predecessor of \( b \), and \( b \) is an \( R \)-successor (or \( R^- \)-predecessor) of \( a \). If \( b \) is an \( R \)-successor of \( a \), \( b \) is also an \( R \)-neighbor of \( a \).

**Definition 2.7.** Signature is defined as a set of named individuals. Given an assertion \( \gamma \) in ABox \( \mathcal{A} \), signature function \( \text{Sig}(\gamma) \) is defined to return a set of all individuals occurring in \( \gamma \). This function is trivially extensible to \( \text{Sig}(\mathcal{A}) \) for the set of all individuals in ABox \( \mathcal{A} \).

**Definition 2.8.** We say there is a role path between individual \( a_1 \) and \( a_n \), if for individuals \( a_1, \ldots, a_n \in I \) and \( R_1, \ldots, R_{n-1} \in R \), there exist either \( R_i(a_i, a_{i+1}) \) or \( R^-_i(a_{i+1}, a_i) \) in \( \mathcal{A} \) for all \( i = 1, \ldots, n-1 \).

The role path from \( a_1 \) to \( a_n \) may involve inverse roles. For example, given \( R_1(a_1, a_2), R_2(a_3, a_2) \), and \( R_3(a_3, a_4) \), the role path from \( a_1 \) to \( a_4 \) is \( \{R_1, R^-_2, R_3\} \), while the opposite from \( a_4 \) to \( a_1 \) is \( \{R^-_3, R_2, R^-_1\} \).

**Definition 2.9.** Quantification depth of a class is referred to as the maximum level of nested quantifiers in the concept, and it can be formalized as

\[
\text{Depth}(Q(C_1 \bowtie C_2)) = 1 + \max(\text{Depth}(C_1), \text{Depth}(C_2))
\]

where \( Q \) stands for any role restriction and \( \bowtie \) stands for any boolean operator (Royer and Quantz, 1993).

As an example, quantification depth of an atomic class \( A \) is 0, and depth of \( \exists R_1.\exists R_2.C \) is 2.

**Definition 2.10.** A class is said in simple form, if its quantification depth is less than 2.

**Remark 2.1.** An arbitrary concept can be linearly reduced to the simple form, e.g. \( \exists R_1.\exists R_2.C \) can be converted to \( \exists R_1.D \) by letting \( D \equiv \exists R_2.C \) where \( D \) is a new class name.

3. Definition of an ABox Module

The notion of ABox module here is to be formalized w.r.t. ontology semantics and entailments that are only meaningful when the ontology is consistent. In this study, however, instead of restricting ontologies to be consistent, we aim to discuss the problem in a broader sense such that the theoretical conclusions hold regardless the consistency state of an ontology. For this purpose, we introduce the notion of justifiable entailment as defined below.

**Definition 3.1 (Justifiable Entailment).** Let \( \mathcal{K} \) be an ontology, \( \alpha \) an axiom, and \( \mathcal{K} \models \alpha \). \( \alpha \) is called a justifiable entailment of \( \mathcal{K} \), iff there exists a consistent fragment \( \mathcal{K}' \subseteq \mathcal{K} \) entailing \( \alpha \), i.e. \( \mathcal{K}' \not\models \bot \) and \( \mathcal{K}' \models \alpha \).
It is not difficult to see that justifiable entailments of an ontology simply make up a subset of its logical entailments. More precisely, for a consistent ontology, the set of its justifiable entailments is exactly the set of its logical ones according to the definition above; while for an inconsistent ontology, justifiable entailments are those sound logical ones that have consistent bases (Huang et al., 2005). For example, given an inconsistent ontology $K = \{ C(a), \neg C(a) \}$, both $C(a)$ and $\neg C(a)$ are justifiable entailments of $K$, but $R(a,b)$ is not.

Remark 3.1. Unless otherwise stated, we take every entailment mentioned in this paper, denoted $K \models \alpha$, to mean a justifiable entailment.

Though complete reasoning on a large ABox may cause intractabilities, in realistic applications, a large ABox may consist of data with great diversity and isolation, and there are situations where a complete reasoning may not be necessary. For example, when performing instance checking of a given individual, say instance CHRMA4 in the biomedical ontology example in Section 1, the ABox contains a great portion of other unrelated matters.

An ideal solution to this ABox reasoning problem is thus to extract a subject-related module and to have the reasoning applied to the module instead. Particularly, to fulfill soundness and completeness, this subject-related module should be a closure of entailments about the given entities, which in DL, should include class and property assertions. This leads to our formal definition of an ABox module as follows:

Definition 3.2 (ABox Module). Let $K = (T, A)$ be an ontology, and a set $S$ of individuals be a signature. $M_S$ with $M_S \subseteq A$ is called an ABox module for signature $S$, iff for any assertion $\gamma$ (either a class or a property assertion) with $\text{Sig}(\gamma) \cap S \neq \emptyset$, $(T, M_S) \models \gamma$ iff $K \models \gamma$.

This definition provides a formal criteria in terms of necessary and sufficient conditions for being an ABox module. It guarantees that sound and complete entailments (represented by $\gamma$) of individuals in signature $S$ can be achieved by independent reasoning on the ABox module w.r.t $T$. Class assertions preserved by an ABox module here are limited to atomic classes defined in $T$.

Remark 3.2. Limiting the preserved class assertions to only atomic classes can obviously simplify the problems to deal with in this paper. While on the other hand, it should not be an obstacle in principle to allow this notion of ABox module to be applied for efficient ontology query with arbitrary classes. This is simply because we can always assign a new name $A$ for an arbitrary complex class $C$ by adding axiom $C \equiv A$ into $T$.

Definition 3.2 does not guarantee, however, uniqueness of an ABox module for signature $S$, since any super set of $M_S$ is also a module for $S$, due to the monotonicity of DLs (Baader et al., 2007). For example, given any signature $S \subseteq I$, the whole ABox $A$ is always a module for $S$.

Note however, the objective of this paper is to extract a precise ABox module and to select only module-essential assertions for a signature, so that the resulting module ensures completeness of entailments and meanwhile keeps a relatively small size by excluding unrelated assertions. Intuitively, for assertions to be module-essential for a signature $S$, they must be able to affect logical consequences of any individual in $S$, so
that by having all these assertions included, the resulting ABox module can preserve all facts of the given entities. This criteria for being a module-essential assertion can be formalized based on the notion of justification (Kalyanpur et al., 2007) as given below.

**Definition 3.3 (Justification).** Let $K$ be an ontology, $\alpha$ an axiom, and $K \models \alpha$. We say a fragment $K' \subseteq K$ is a justification for axiom $\alpha$, denoted $\text{Just}(\alpha, K)$, iff $K' \models \alpha$ and $K'' \not\models \alpha$ for any $K'' \subset K'$.

**Definition 3.4 (Module-essentiality).** Let $K$ be an ontology, $a$ be an individual name, and $\gamma$ an ABox assertion. $\gamma$ is called module-essential for $\{a\}$, iff

$$\gamma \in \text{Just}(\alpha, K) \land K \models \alpha,$$

for any axiom $\alpha$ (either $C(a)$ or $R(a, b)$) that is a logical entailment of $a$.

A justification $\text{Just}(\alpha, K)$ for axiom $\alpha$ is in fact a minimum fragment of the knowledge base that implies $\alpha$, and every axiom (or assertion) in $\text{Just}(\alpha, K)$ is thus essential for this implication. Following from this point, an assertion $\gamma$ is considered able to affect logical consequences of some individual in signature $S$, if and only if it appears in some justification for either (i) property assertion or (ii) class assertion of that individual. Satisfying either one of them, $\gamma$ is considered module-essential for $S$ and should be included in $M_S$. With the notion of module-essentiality, we can now impose more restrictions on an ABox module, and the notion of Exact ABox Module is derived as follows:

**Definition 3.5 (Exact ABox Module).** Let $S$ be a signature, $A$ be an ABox, and $M_S \subseteq A$ be an ABox module for $S$. $M_S$ is called an Exact ABox Module for $S$, iff every assertion in $M_S$ is module-essential for $S$, while any of those in $A \setminus M_S$ is not.

In the following sections, a theoretical approach and an approximation are presented for the computation of an exact ABox module. Without loss of generality, we assume all ontology classes are in simple form as defined previously, and class terms in all class assertions are atomic.

We will show how to compute an ABox module in two steps: we begin by showing module extraction in an equality-free SHIQ ontology, and later we show how the basic technique can be extended to deal with equality.

### 4. ABox Modules in Equality-Free Ontologies

An ontology is called equality-free, if it does not entail any equality (i.e. $K \not\models a \approx b$, for any individual $a, b$ in $K$). In this section, we concentrate on a method that computes exact ABox modules in ontologies of this type. To further simplify the problem, we will consider module extraction for a single individual instead of an arbitrary signature $S$, since the union of modules of individuals in $S$ yields a module for $S$, as indicated by the following proposition.

**Proposition 4.1.** Let $S$ be a signature, $M_{\{i\}}$ be an ABox module for each individual $i \in S$, and

$$M_S = \bigcup_{i \in S} M_{\{i\}}.$$


\( \mathcal{M}_S \) is an ABox module for \( S \).

**Proof.** Assume \( \mathcal{M}_S \) is not a module for \( S \), then there exists an assertion \( \gamma \), with \( \text{Sig}(\gamma) \cap S \neq \emptyset \), and either (i) \((T, \mathcal{M}_S) \models \gamma \) and \( K \not\models \gamma \) or, (ii) \( K \models \gamma \) and \((T, \mathcal{M}_S) \not\models \gamma \).

(i) clearly contradicts the monotonicity. For (ii), let individual \( a \in \text{Sig}(\gamma) \cap S \), \( \mathcal{M}_{\{a\}} \subseteq \mathcal{M}_S \) be the module for \( a \). Then, by the module definition, we have \((T, \mathcal{M}_{\{a\}}) \models \gamma \), which again conflicts with the monotonicity of DLs, since \( \mathcal{M}_{\{a\}} \) is subsumed by \( \mathcal{M}_S \). Hence, the proposition holds. \( \blacksquare \)

### 4.1. Strategy

To compute an exact ABox module for a given individual \( a \), basically we have to test every assertion in \( A \) to see if it is module-essential for \( a \), that is, to test every assertion whether it contributes in deducing any property assertion or class assertion of individual \( a \).

As a fact of \( \text{SHIQ} \), deducing class assertions (classification) of an individual usually depends on both of its class and property assertions. Conversely, observation of a tableau-expansion procedure for \( \text{SHIQ} \) tells that, deducing property assertions of an individual should not be affected by its class assertions, except via individual equality \( \) (Horrocks et al., 2000) as discussed in Section 4.3. The rationale behind this observation is shown by Theorem 4.1 and the proof for Proposition 4.3 together.

Therefore, given an equality-free ontology, the above observation allows us to deduce property assertions from the ABox w.r.t. only role hierarchies \( \) defined in \( T \) (Horrocks et al., 2000), and a strategy for extracting an ABox module can thus be devised.

**Proposition 4.2.** Given an equality-free \( \text{SHIQ} \) ontology, computation of an ABox module for individual \( a \) can be divided into two steps:

1. compute a set of assertions (denoted \( \mathcal{M}^P_{\{a\}} \)) that preserves all property assertions of \( a \),
2. compute a set of assertions (denoted \( \mathcal{M}^C_{\{a\}} \)) that preserves all class assertions of \( a \).

For simplicity, we call \( \mathcal{M}^P_{\{a\}} \) a property-preserved module and \( \mathcal{M}^C_{\{a\}} \) a classification-preserved module for \( \{a\} \). The following theorem states that, in \( \text{SHIQ} \), no individual classifications can directly entail any property assertion. Thus, given an equality-free ontology, this theorem ensures that the computation of \( \mathcal{M}^P_{\{a\}} \) is independent from that of \( \mathcal{M}^C_{\{a\}} \).

**Theorem 4.1.** Given a \( \text{SHIQ} \)-TBox \( T \), and \( a, b \) two individual names:

\[ (T, \{C_1(a), C_2(b)\}) \not\models R(a, b), \]

for any \( R \in R^* \) (not universal) and arbitrary \( \text{SHIQ} \)-classes \( C_1, C_2 \).

**Proof.** Let \( A \) be \( \{C_1(a), C_2(b)\} \), and assume \((T, A) \models R(a, b) \). We prove by contradiction.

Since \( R(a, b) \) is justifiable, there always exists a consistent fragment \( T' \subseteq T \), \((T', A) \models R(a, b) \). Thus, there exists an interpretation \( I \) such that \( a^I \in C_1^I, b^I \in C_2^I \), and

\[ ^1 \text{Note that transitive roles can be expressed as } R \circ R \subseteq R. \]
(a^T, b^T) ∈ R^T$, and it simply follows that \( a^T ∈ (∃R.b)^T \), where \((∃R.b)^T \) is interpreted as \( \{ x \mid (x, b^T) ∈ R^T \} \).

Interpretation \( I \) can be easily extended by adding an element \( d \), with \( d ∈ C^T_i \) and \( d ≠ a^T \). Since any pair of individuals of \( C_1 \) and \( C_2 \) are related by \( R \), it follows that \( (d, b^T) ∈ R^T \) and thus \( d ∈ (∃R.b)^T \). Then, \( I \) can be further extended by adding any number of different elements of \( C^T_i \), which will all end up as elements of \((∃R.b)^T\), and this property must be satisfied for any interpretation to be a valid model for \((T', A)\). Thus, (i) \((T', A)⊨C_1 ⊑ ∃R.b\), where we have to borrow the nominal (i.e. \( \{b\} \)) to precisely express the class subsumption; and simply it leads to (ii) \((T', A)⊨C_1 ⊑ ∃R.C_3\), if \( A \) is extended with arbitrary assertion \( C_3(b) \) for any \( SHIQ\)-class \( C_3 \).

The latter (ii) clearly contradicts with the theorem that class subsumption is not affected by assertions if nominals are disallowed in \( T \). Thus, the assumption does not hold. ■

4.2. A Property-Preserved Module

A property-preserved module of individual \( a \) is essentially a set of assertions in ontology \( K \), which affect the deduction of \( a \)'s property assertions, and this set is denoted \( M^P_{\{a\}} \) such that

\[
M^P_{\{a\}} = \{ \gamma \mid \gamma ∈ Just(R(a, b), K) ∧ K ⊨ R(a, b) \},
\]

where \( γ \) is an ABox assertion, and \( Just(R(a, b), K) \) is any justification for \( R(a, b) \).

In an equality-free \( SHIQ \) ontology, since property assertions are deduced from the ABox \( A \) w.r.t. only role hierarchies, the computation for \( M^P_{\{a\}} \) is then straightforward based on the following fact: for any \( R ∈ R^* \), if \( K ⊨ R(a, b) \), there are two possibilities on an equality-free \( SHIQ \) ontology, according to [Horrocks et al. 2000].

1. assertion \( R_0(a, b) \) or \( Inv(R_0)(b, a) ∈ A \) with \( R_0 ⊑ R \),
2. assertions involved in a role path from \( a \) to \( b \), with all roles having a common transitive parent \( R_0 \) and \( R_0 ⊑ R \). e.g. \( R_1(a, a_1), R_2(a_2, a_1), R_3(a_2, b) ∈ A \), with \( R_1, R_2, R_3 ⊑ R_0 \) and \( R_0 \) is transitive.

Abstracting from a particular \( R_0 \), these two possibilities can be generalized into a formal criteria to select assertions to include in \( M^P_{\{a\}} \) for individual \( a \):

C1. property assertions in \( A \) that have \( a \) as either subject or object, or
C2. property assertions in \( A \) that are involved in a role path from \( a \) to some \( b \), with all roles in the path having a common transitive parent.

**Proposition 4.3.** On an equality-free \( SHIQ \) ontology, the set of property assertions satisfying criteria C1 or C2 forms a property-preserved module \( M^P_{\{a\}} \) for individual \( a \).

**Proof.** Correctness of this proposition can be verified by observation of the tableau-construction procedure for \( SHIQ \) presented in [Horrocks et al. 2000]. Let a tableau \( T = (Δ, L, E, \mathcal{I}) \) be an interpretation for \( K \) as defined in [Horrocks et al. 2000], where \( Δ \) is a non-empty set, \( L \) maps each element in \( Δ \) to a set of concepts, \( E \) maps each role to a set of pairs of elements in \( Δ \), and \( \mathcal{I} \) maps individuals in \( A \) to elements in \( Δ \).

They have proven that, for tableau \( T \) to be a model for \( K \), if \( K ⊨ R(a, b) \), there must be either \((a^T, b^T) ∈ E(R)\) or a path \((a^T, s_1), (s_1, s_2), \ldots, (s_n, b^T) ∈ E(R_0)\) with
$R_0 \subseteq R$ and $R_0$ being transitive. The second scenario is consistent with criteria $C_2$; while for the first one, i.e. $(a^T, b^T) \in \mathcal{E}(R)$, there are only two possibilities according to the tableau-constructing procedure: (i) $R_0(a, b)$ or $R_0^-(b, a) \in A$ with $R_0 \subseteq R$ that triggers initialization of $\mathcal{E}(R_0)$; (ii) $R_0(a, b)$ or $R_0^-(b, a)$ is obtained through the $\preceq_r$-rule for identical named individuals (Horrocks et al., 2000) with $R_0 \subseteq R$. The (i) reflects exactly the criteria $C_1$, while the (ii) does not apply here for equality-free ontologies. Therefore, the proposition holds.

4.3. An Exact ABox Module

To compute an exact ABox module $\mathcal{M}_{\{a\}}$, we need to further decide a set of assertions that affect classifications of the individual, and this set is denoted $\mathcal{M}^C_{\{a\}}$ such that

$$\mathcal{M}^C_{\{a\}} = \{ \gamma \mid \gamma \in \text{Just}(A(a), K) \land K \models A(a) \},$$

where $\gamma$ is an ABox assertion, $A$ is an atomic class, and $\text{Just}(A(a), K)$ is any justification for $A(a)$.

As previously stated, in SHIQ an individual is usually classified based on both its class and property assertions in $A$. It is obvious that explicit class assertions of a form an indispensable part of $\mathcal{M}^C_{\{a\}}$. Then, to decide any property assertion of $a$ that affects its classification, we examine each one that is captured in $\mathcal{M}^P_{\{a\}}$.

The decision procedure here is based on the idea that instance checking is reducible (though may not be trivial) to concept subsumption (Donini and Era, 1992; Donini et al., 1994; Nebel, 1990), i.e.

Given an ontology $K = (T, A)$, an individual $a$ and a SHIQ-class $C$, $a$ can be classified into $C$, if $a$’s object description in the ABox is subsumed by $C$ w.r.t. $T$.

This idea automatically lends itself as a guideline, such that to determine any assertion of an individual that contributes to its classification, we have to decide if the class term behind this assertion is subsumed by some class w.r.t. $T$.

For exposition, consider the following example: Let an ontology $K = (T, A)$ be

$$((\exists R_0 . B \sqsubseteq A), \{ R_0(a, b), B(b) \}),$$

and let us ask whether $R_0(a, b)$ is essential for individual $a$’s classification or not. To answer that, we need to decide if the class term behind this role assertion is subsumed by some class w.r.t. $T$, i.e. to test if $T$ entails

$$\exists R . C_1 \sqsubseteq C_2$$

for some named class $C_2$, with $R_0 \subseteq R$ and $b \in C_1$. If (1) is satisfied, the answer is YES, and otherwise NO.

It is easy to see (1) is satisfied in this example by substituting $B$ for $C_1$ and $A$ for $C_2$. We can thus determine $R_0(a, b)$ is one of the causes for the entailment $C_2(a)$ (i.e. it is in some justification for $C_2(a)$), and should be an element of $\mathcal{M}^C_{\{a\}}$. Moreover, assertions in $\text{Just}(C_1(b), K)$ should also be elements of $\mathcal{M}^C_{\{a\}}$, since $C_1(b)$ here is another important factor to the classification $C_2(a)$.
1. Compute a property-preserved module $\mathcal{M}_{\{a\}}^P$ for the given individual $a$, by following the criteria $C1$ and $C2$ given in Section 4.2.

2. Add all explicit class assertions of $a$ into $\mathcal{M}_{\{a\}}^C$.

3. For every $R_0(a,b)$ captured in $\mathcal{M}_{\{a\}}^P$ and any $R$ with $R_0 \sqsubseteq R$, test if the corresponding condition $[3]$ is satisfied. If it is yes, add $R_0(a,b)$, assertions in $\text{Just}(C_1(b), K)$s, and any inequality assertions between individuals in $R^K(a, C_1)$ into $\mathcal{M}_{\{a\}}^C$.

4. Unite the sets, $\mathcal{M}_{\{a\}}^P$ and $\mathcal{M}_{\{a\}}^C$, to form the exact ABox module for $a$.

---

Figure 1: Steps for computation of an exact ABox module for individual $a$.

---

The above example illustrates a simple case on when a single property assertion can affect classification of an individual and which assertions are then essential for it. A more general case, however, could be that classification of an individual is caused by multiple assertions of that entity together. For example, let $K$ be

$$(\exists R_0.B \cap \exists R_1.C \sqsubseteq A), \{R_0(a,b), R_1(a,c), B(b), C(c)\}.$$  

Here, $R_0(a,b)$ is still essential for the deduction $A(a)$. But when testing condition [1] for $R_0(a,b)$, it will be found unsatisfied, which is simply because of its complete disregard for other assertions of the individual, which are also factors to the entailment.

Thus, for comprehensiveness and consideration of other information of the individual, condition [1] should be generalized into:

$$\exists R.C_1 \cap C_3 \sqsubseteq C_2$$

(2)

where all other information of individual $a$ is summarized and incorporated into a class $C_3$ with $C_3 \sqsubseteq C_2$. Moreover, taking the number restrictions in $\text{SHIQ}$ into consideration, condition [2] can be further generalized as:

$$\geq nR.C_1 \cap C_3 \sqsubseteq C_2 \land |R^K(a, C_1)| \geq n.$$  

(3)

Note that, $\exists R.C_1$ is only a special case of $\geq nR.C_1$, and $R^K(a, C_1) = \{b_i \in I \mid K \models R(a, b_i) \land C_1(b_i)\}$ denotes the set of decidable and distinct $R$-neighbors of individual $a$ in $C_1$.

With condition [3] derived, we are now in a position to present a procedure for computation of $\mathcal{M}_{\{a\}}^C$, and also an exact ABox module for individual $a$, which is summarized in Figure 1.

### 4.4. Approximation of an Exact ABox Module

Computation of $\mathcal{M}_{\{a\}}^P$ tightly depends on the complete role hierarchy, which should be computable using a DL-reasoner, since roles in $\text{SHIQ}$ are atomic and most importantly the size of $T$ should be much smaller than $A$ in realistic applications (Motik and Sattler, 2006). On the other hand, it is difficult to compute $\mathcal{M}_{\{a\}}^C$, since it demands computation of both concept subsumption (i.e. condition [3]) and justifications for class assertions (i.e. $\text{Just}(C_1(b), K)$). Simple approximations for both are given in this section as follows.
Definition 4.1 (Approximation of (3)). A syntactic approximation for condition (3) for $R_0(a,b)$ is that: to test if $K$ contains any formula in the form as listed below:

$$\exists R.C \subseteq D \quad \equiv \quad \neg D \subseteq \forall R. \neg C$$

$$\geq nR.C \subseteq D \quad \equiv \quad \neg D \subseteq (n - 1)R.C.$$  \hspace{1cm} (4)

where $R_0 \subseteq R, C_1 \in C, \propto$ is a place holder for $\sqcup$ and $\sqcap$, and concepts are checked in Negation Normal Form (NNF) \cite{Baader et al. 2007}. Please also note the following equivalences:

$$\exists R.C \subseteq D \quad \equiv \quad \neg D \subseteq \forall R. \neg C$$

$$\geq nR.C \subseteq D \quad \equiv \quad \neg D \subseteq (n - 1)R.C.$$  \hspace{1cm} (4)

For assertion $R_0(a,b)$, the approximation here for condition (3) is to check if any formula in $K$ is in the form of any listed axioms in (4). If it is yes, $R_0(a,b)$ may potentially affect some logical entailment of individual $a$, and related assertions will be added into $a$'s ABox module to ensure preservation of this potential entailment. Validity of this approximation is shown by the following proposition.

Proposition 4.4. If condition (3) is satisfied, there must exist some formula in $K$ in the form as listed in (4), given all ontology classes in the simple form.

Proof. Since $\exists R.C$ is a special case of $\geq nR.C, C \subseteq \forall R. \neg D$ is equivalent with $\exists R.C \subseteq D$ \cite{Grosof et al. 2003}, and together with those equivalences mentioned above, every role restriction in $T$ can be converted to the form of $\geq nR.C$ by axiom manipulation. Then, the task here is reducible to proving that if (3) is satisfied, there must be some formula in $K$ in the form of $\geq nR.C \propto C_2 \subseteq C_3$ for some $R$ with $R_0 \subseteq R$.

It is straightforward that, if no $R$ with $R_0 \subseteq R$ is used in concept definition, $\geq nR.C$ is not comparable (w.r.t. subsumption) with any atomic class (except $\top$ and its equivalents). On the other hand, if $\geq nR.C$ is used in concept definition but occurring only in the right-hand side (r.h.s.) of GCIs (or $\neg P$ in l.h.s.), it is unable to indicate any atomic class as its subsumer, which can be confirmed by observation of a tableau-constructing procedure.

Let $P,Q$ be two atomic classes, $Q \neq \top, \neg P$ and $\neg P$ not fillers in any restrictions, and all concepts of $T$ in NNF. Assume $(*)$ $P$ occurs only in r.h.s. of GCIs (or $\neg P$ in l.h.s.), and there is a consistent fragment $T' \subseteq T$ that $T' \models P \sqsubseteq Q$. Then, it simply follows that:

(E1) $T' \cup \{a\} \models \neg P \sqcup Q(a)$ for any individual $a$, since $P \sqsubseteq Q$ implies $\top \sqsubseteq \neg P \sqcup Q$.

(E2) $T' \cup \{\neg Q(a)\} \models \neg P(a)$, because of (E1).

(E3) $T' \cup \{P \sqcap \neg Q(a)\} \models \bot$, the so-called refutation-style proof for $P \sqsubseteq Q$.

(E1) implies that, in any tableau that is a model of $T' \cup \{a\}$, there must be either $\neg P$ or $Q$ in $L(a^2)$ (i.e. the class set of $a^2$ in the tableau), which can be shown by contradiction: suppose $I_1$ is a model for $T' \cup \{a\}$, where neither $\neg P$ nor $Q$ is in $L(a^2)$. Let $I_2$ be another tableau such that $I_2$ coincides with $I_1$ except $L(a^2)$ is extended with $\{P, \neg Q\}$, and $I_2$ should be clash-free since both $P$ and $Q$ are atomic and no tableau rules can be applied. Thus, $I_2$ turns out to be a model for $T' \cup \{P \sqcap \neg Q(a)\}$ that violates (E3).
1. Compute a property-preserved module $\mathcal{M}^P_{\{a\}}$ for the given individual $a$, by following the criteria $C_1$ and $C_2$ given in Section 4.2.

2. Add all explicit class assertions of $a$ into $\mathcal{M}^C_{\{a\}}$.

3. For every $R_0(a,b)$ captured in $\mathcal{M}^P_{\{a\}}$ and any $R$ with $R_0 \sqsubseteq R$, test if $K$ contains any formula in the form as listed in (4). If it is yes, add $R_0(a,b)$, all assertions in $\mathcal{M}_{\{b\}}$, and any inequality assertions between $a$’s $R$-neighbors into $\mathcal{M}^C_{\{a\}}$.

4. Unite the sets, $\mathcal{M}^P_{\{a\}}$ and $\mathcal{M}^C_{\{a\}}$, to form an approximation for the exact ABox module for $a$.

Figure 2: Steps for approximation of an exact ABox module for individual $a$.

Analogously for (E2), there must be $\neg P$ in $L(a^I)$ for any model of $T' \cup \{\neg Q(a)\}$. Nevertheless, if $P$ occurs only in r.h.s. of GCIs in $T$, $\neg P$ can never exist after the NNF transformation of axioms in $T$, and since $P$ and $\neg P$ are not fillers in any restrictions, $L(a^I)$ can never comprise $\neg P$ according to the tableau rules [Horrocks et al., 2000]. Hence, the original assumption (*) does not hold.

The above case essentially tells, if an atomic concept occurs only in r.h.s. of GCIs in $T$, its subsumer is undecidable. And the same general principle applies, if we consider all $\geq nR.C$ for any $R$ with $R_0 \sqsubseteq R$ as a single unit. Thus, there must be some $\geq nR.C$ with $R_0 \sqsubseteq R$ occurring in l.h.s. of GCIs in $T$, if (3) is true. ■

Proposition 4.4 shows the completeness of the approximation (4). We can thus conclude, an ABox module resulted from this approximation is still able to capture complete classifications (w.r.t. $T$) of the given individual, which are derivable from its property assertions. And the following definition is an immediate consequence of this conclusion:

**Definition 4.2.** If $C_1 \neq \top$, an approximation for union of assertions in Just$(C_1(b), K)s$ is $\mathcal{M}_{\{b\}}$, the exact ABox module for $b$, which is approximated using the same strategies here.

A procedure for approximating an exact ABox module is then summarized in Figure 2.

5. Module with Equality

In this section, we show how the outcome from the previous section can be utilized to tackle module extraction with individual equality.

In $\mathcal{SHIQ}$, individual equality is recognized through the $\leq_r$-rule defined in [Horrocks et al., 2000], which is briefly summarized as follows:

$\leq_r$-rule: If $\leq n.R.C \in L(x^I)$, and $x^I$ has more than $n$ $R$-neighbors in $C$, then for two of these $R$-neighbors $y^I$ and $z^I$, where $y, z$ are named individuals, if $y^I \not\approx z^I$ does not hold, then $y^I \approx z^I$ (thus $y \approx z$).
Observation of this rule informs that, determination of individual equality requires computation of both property and class assertions of related individuals, i.e. in the $\leq_r$-rule, in order to derive $y \approx z$ we need to know at least

\[ K \models \leq n.R.C(x) \land R(x,y) \land R(x,z) \land C(y) \land C(z). \]

Hence, with arbitrary equality, it is difficult even with assistance of a DL-reasoner to extract an ABox module, since it may involve a prohibitive ontology reasoning to detect the equality. Besides, the strategy proposed in Proposition 4.2 becomes infeasible here, since assertions can be derived from equalities (e.g. given $y \approx z$, $R(y,w)$ simply implies $R(z,w)$) that rely on classification of individuals. In other words, with equality, the computation of $M_{\{a\}}^P$ may be dependent on that of $M_{\{a\}}^C$, which consequently causes inapplicability of all the techniques discussed so far.

Nevertheless, extraction of ABox modules may still be achieved, if we treat and modularize the ABox as if it were equality-free, and resolve the equality in a post-processing manner.

**Proposition 5.1.** Let individual $x \in \leq nR.C$ has more than $n$ R-neighbors in $C$, two of which $y$ and $z$ can be determinately identified (i.e. $y \approx z$). Let signature $S$ consists of $x$ and all its R-neighbors in $C$, and

\[ M_S = \bigcup_{i \in S} M_{\{i\}} \]

where $M_{\{i\}}$ is an ABox module for each individual $i \in S$ in the “equality-free” ABox. $M_S$ preserves the equality $y \approx z$.

Proposition 5.1 suggests a strategy to retain equality between $y$ and $z$, by combining “modules” of related individuals. With $y \approx z$ preserved, $M_S$ automatically preserves all facts of $y$ and $z$ that are derived from the equality. Subsequently, for neighbors of $y$ and $z$ (i.e. individuals in $\text{Sig}(M_{\{y\}}^P)$ and $\text{Sig}(M_{\{z\}}^P)$), modules of these entities should be combined with the $M_S$ obtained above, so that their facts derivable from $y \approx z$ can also be captured. This strategy to retain equality in ABox modules should be applied recursively for all the identities.

Notice however, the strategy in Proposition 5.1 is based on the condition that individual $y$ and $z$ are identified in the first place, which is a difficult task without ABox reasoning. Nevertheless, conceiving that equality in $\text{SHIQ}$ stems from number restrictions (Horrocks et al., 2000), a simple approximation for it is given in the definition below.

**Definition 5.1.** Let $x,y$ and $z$ be named individuals, $y,z$ be R-neighbors of $x$, and $y \not\approx z$ do not hold explicitly. $y$ and $z$ are considered potential equivalents, if their R-predecessor $x$:

1. has no potential equivalents, and has $m$ R-neighbors with $m \geq n$, or
2. has a set of potential equivalents, denoted $X$ (include $x$), and there exists a set $S \in \left( m'-n'+1 \right.$, such that

\[ \max_y \{ |\{ y_i \mid R(x_i, y_i) \in A \land x_i \in S \} | = m \geq n. \] (5)

2 Either $y \not\approx z$ is not explicitly given, or assertions $C(y)$ and $\neg C(z)$ do not occur simultaneously in $A$. 14
where $R$ is used in number restrictions as in the axiom listed in (4), and $n$ is the minimum of the set $\{k \mid \geq (k+1)R.C$ in l.h.s. of GCIs $\}$. $\{X_{m'-n'+1}\}$ denotes the set of all $(m'-n'+1)$-combinations of set $X$, and variables $m'$ and $n'$ are for identification of $x$ that correspond to above $m$ and $n$ respectively.

In this definition, the possibility for $\mathcal{K} \models \leq nR.C(x)$ requires $\leq nR.C$ or its isoform occurs in r.h.s. (or $\geq (n+1)R.C$ in l.h.s. as in (4)) of GCIs in $\mathcal{T}$, proof for which is similar to the one given for Proposition 4.4.

Moreover, if $x$ itself has potential equivalents, individuals $y$ and $z$ should be elements of $\{y_i \mid R(x_i, y_i) \in \mathcal{A} \wedge x_i \in X\}$. For the counting of potential $R$-neighbors of $x$, instead of taking the entire set $X$, only the $(m'-n'+1)$-combination of $X$ that maximizes the counting is considered (see equation (5)), since given $m'$ $R'$-neighbors, the maximum possible number of decidable identical entities is $(m'-n'+1)$ out of $m'$ according to (Horrocks et al., 2000). Examples for two cases of individual $x$ in Definition 5.1 are illustrated in Figure 3.

**Proposition 5.2.** Applying the strategy in Proposition 5.1 to potential equivalents generates modules that each preserves individual equality if any.

**Proof.** (Sketch) We prove by induction.

**Basis:** Assume initially, $y \approx z$ is the only equality in ABox that is entailed from the $\leq_{r}$-rule, then, there must be $R_1(x, y)$ and $R_2(x, z)$ in $\mathcal{M}_{x}(\mathcal{T})$, for some $x \in \leq nR.C$, $x$ having more than $n$ $R$-neighbors, and $R_1, R_2 \subseteq R$. According to the strategy in Proposition 5.1, modules in “equality-free” ABox of $x$ and its $R$-neighbors are merged, and resulted $\mathcal{M}_{S}$ thereby has preservation of all facts (except those derived from $y \approx z$) of these individuals, including assertion $R(x, i)$ and classification for
each x’s R-neighbor i, which are sufficient to entail $y \approx z$, according to the $\leq_r$-rule (Horrocks et al., 2000).

Inductive step: Assuming ABox $A$ with arbitrary individual equality is modularized, and all extracted modules are in compliance with Definition 3.2. $A$ is then further extended by adding a new assertion $R(x', y')$ that causes fresh equality between individual $y'$ and $z'$. If $x'$ has no equivalent, this simply follows what we have discussed above. Otherwise, we take into account all $x'$’s potential equivalents in set $X'$ for its R-neighbors that comprise both $y'$ and $z'$. There exists a set $S \subseteq X'$ of true positives such that $\{|\{y'_i| R(x'_i, y'_i), x'_i \in S\}|$ is greater than the cardinality restriction as required for individual identity, and $S$ is at most a $(m' - n' + 1)$-combination of $X'$. Modules of $x'$ and all its R-neighbors are then merged, and as discussed above equality between $y'$ and $z'$ is preserved.

6. Related Work

ABox partition has been addressed mainly by (Guo and Heflin, 2006), (Du and Shen, 2007) and (Wandelt and Möller, 2012).

In (Guo and Heflin, 2006), the authors proposed a method to compute independent ABox partitions for $SHIF$ ontologies, such that complete reasoning can be achieved if combining results of independent reasoning on each partition. This method for ABox partitioning is based on a set of inference rules in (Royer and Quantz, 1993), and it encapsulates assertions that are antecedents of a inference rule as an ABox partition.

In (Du and Shen, 2007), the authors proposed an algorithm for ABox partition for $SHIQ$ ontologies. Based on the technique in (Hustadt et al., 2004) that converts DL ontologies to disjunctive datalog programs, their algorithm further converts the disjunctive datalog program into a plain datalog program, to generate rules that can be employed as guidelines for ABox partition.

Both approaches above for ABox partition, however, failed to impose any logical restrictions on a single partition. An immediate consequence is that, every single assertion can turn out to be a partition if the ontology is too simple, i.e. has no transitive roles nor concepts defined upon restrictions. Besides, to get complete entailments of an individual, one still has to reason over all partitions of the ontology.

The work that is mostly related to the subject here is presented in (Wandelt and Möller, 2012), where the authors provided a formal definition of ABox modularization that extends the notion of ABox partition from (Guo and Heflin, 2006). The goal in (Wandelt and Möller, 2012) is to compute ABox modules as small as possible (different from Exact ABox Modules in this paper). And the approximation proposed there requires information from the class hierarchy (including concept subsumption and concept disjointness, where complex classes may be involved), which thus requires invocation of a DL reasoner. Their approach focus on $SHI$ ontologies and requires ontology consistency as the prerequisite.
Table 1: ABox modules in different ontologies

| Ontology | Exp | #Ast | #Ind | Max. #Ast/#Ind | Avg. #Ast/#Ind | Extraction Time |
|----------|-----|------|------|---------------|----------------|----------------|
| LUBM-1   | SHI | 67465| 17175| 2921/593      | 13.1/2.4       | 1.12 ms        |
| LUBM-2   | SHI | 319714| 78579| 2921/593      | 14.4/2.5       | 1.22 ms        |
| VICODI   | ALH | 53653| 16942| 8591/1        | 5.3/1          | 0.28 ms        |
| AT       | SHIN| 117405| 42695| 54561/10870   | 6.9/1.7        | 1.06 ms        |
| CE       | SHIN| 105238| 37162| 49914/9315    | 7.1/1.7        | 0.60 ms        |
| DD       | SHIN| 74299| 26100| 33150/6065    | 7.2/1.8        | 0.50 ms        |

Table 2: Small and simple ABox modules

| Ontology | Total #Modules | #Module with #Ast ≤ 10 (%) | #Module with Signature Size = 1 (%) |
|----------|---------------|---------------------------|------------------------------------|
| LUBM-1   | 7264          | 7210 (99.3)               | 7209 (99.2)                        |
| LUBM-2   | 31361         | 30148 (96.1)              | 31103 (99.2)                       |
| VICODI   | 16942         | 16801 (99.2)              | 16942 (100)                        |
| AT       | 24606         | 23464 (95.4)              | 23114 (93.9)                       |
| CE       | 21305         | 20010 (93.9)              | 19455 (91.3)                       |
| DD       | 14855         | 13815 (93.0)              | 13226 (89.0)                       |

7. Empirical Evaluation

We implemented and tested our approximation on a lab PC with Pentium(R)4 3.20GHz CPU, 1.5GB Java Heap and Windows XP.

For test data, we collected a set of ontologies with large ABoxes:

1. VICODI[^4] is an ontology modeling European history;
2. LUBM-1 and LUBM-2 are benchmark ontologies generated using tools provided by [Guo et al., 2005][^5];
3. Arabidopsis Thaliana (AT), Caenorhabditis Elegans (CE), and Dictyostelium Discoidium (DD) are three biomedical ontologies based on a common TBox that models biological pathways.

Details of these ontologies are summarized in Table 1 in terms of expressiveness (Exp), number of assertions (#Ast), and number of individuals (#Ind).

[^4]: http://www.vicodi.org
[^5]: http://www.reactome.org/download
These collected ontologies are then modularized using the approximation for module extraction for every named individual. As discussed previously, extracting ABox module for a single individual may result in a module, signature of which is constituted by a set of entities because of module combination.

In Table 1 we show the statistics of maximum and average module size in terms of number of assertions (#Ast) and size of signature (#Ind). It can be observed that, in average, modules of all these ontologies are significantly smaller comparing with the entire ABox; while in some ontologies, the maximum module is relatively large, either because there is a great number of assertions of a single individual, or because there are indeed intricate relationships between individuals that may affect classification of each other.

VICODI is a simple ontology that has no property assertion in ABox with condition (4) satisfied, and thus, signature of every ABox module is constituted by only a single
entity. However, its maximum ABox module consists of more than 8000 assertions for a single individual.

For biomedical ontologies, AT, CE, and DD, their maximum ABox modules are large and complex, and it is mainly because: (i) the terminological part of these ontologies is highly complex, with 33 out of 55 object properties either functional/inverse functional or used as restrictions for concept definition; and (ii) these ontologies also have single individuals that each has a great number of property assertions (e.g. AT has one individual with 8520 assertions). Thus, connections between these individuals by any role assertions satisfying condition (4) will result in a large module.

For LUBM-1 and LUBM-2, sizes of their maximum modules are in between of these two categories above, due to moderate complexity of the ontologies and only 4 out of 25 object properties are used for class definition.

As shown in Table 2, most of the ABox modules (greater than 90%) in these ontologies, are small and simple with no greater than ten assertions or with a single individual in signature. For modules with more than one individual in signature, we plot distributions of signature sizes of these ABox modules in Figure 4 for LUBM-1, LUBM-2, AT and CE. The X-axis gives the size range and the Y-axis gives the number of modules in each size range. Because of a large span and ragged distribution of module sizes, we use non-uniformed size ranges, so that we can have a relatively simple while detailed view of distributions of small, medium, and large modules. It can be observed from the figure that:

1. For all these ontologies, the majority of ABox modules are still small with no more than five individuals in signature;
2. LUBMs have more medium modules that have signature size between 50 and 600 due to moderate ontology complexity, and
3. The biomedical ontologies have almost all modules with signature size below 200 but one with very large signature (more than 1,000 individuals).

Those large ABox modules in biomedical ontologies can be caused by complexity of the ontology as discussed above and also the inadequacy of our approximation.

8. Conclusion

In this paper, we have proposed a formal definition of an ABox module that ensures complete preservation of logical entailments for a given set of individuals. We have presented a theoretical approach for extraction of an ABox module. For applicability in realistic ontologies, we have also provided a simple and tractable approximation. This approximation is straightforward and easy to implement, while on the other hand, it is naive and inadequate, as it would cause a module to contain many unrelated assertions when the ontology is complex. Nonetheless, it is obviously amenable to optimization.

For future work, we will concentrate on developing more rigorous approximations for condition (3) and individual equality, such that smaller and precise modules can be extracted for complex ontologies.

9. Acknowledgments

We’d like to thank Saminda Abeyruwan and Hande Küçük for their contributions to the project, and anonymous reviewers for their helpful comments.
References

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (Eds.), 2007. The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press.

Cuenca Grau, B., Horrocks, I., Kazakov, Y., Sattler, U., 2007a. Just the right amount: Extracting modules from ontologies, in: Proceedings of WWW’07, pp. 717–726.

Cuenca Grau, B., Horrocks, I., Kazakov, Y., Sattler, U., 2007b. A logical framework for modularity of ontologies, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), AAAI, pp. 298–304.

Donini, F., Era, A., 1992. Most specific concepts for knowledge bases with incomplete information, in: Proceedings of CIKM, pp. 545–551.

Donini, F., Lenziner, M., Nardi, D., Schaar, A., 1994. Deduction in concept languages: From subsumption to instance checking. Journal of logic and computation 4, 429–452.

Du, J., Shen, Y., 2007. Partitioning aboxes based on converting DL to plain datalog, in: Proceedings of Description Logics (DL-07), pp. 251–258.

Glimm, B., Horrocks, I., Lutz, C., Sattler, U., 2008. Conjunctive query answering for the description logic. Journal Artificial Intelligence Research 31, 157–204.

Gros, B., Horrocks, I., Volz, R., Decker, S., 2003. Description logic programs with description logic, in: Proceedings of WWW, ACM. pp. 48–57.

Guo, Y., Heflin, J., 2006. A scalable approach for partitioning owl knowledge bases, in: International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS), Springer. pp. 636–641.

Guo, Y., Pan, Z., Heflin, J., 2005. LUBM: A benchmark for owl knowledge base systems. Journal of Web Semantics 3, 158–182.

Haarslev, V., Möller, R., 2001. RACER system description, in: Proceedings of the First International Joint Conference on Automated Reasoning, Springer. pp. 701–705.

Horrocks, I., Sattler, U., Tobies, S., 2000. Reasoning with individuals for the description logic SHIQ, in: Proceedings of Conference on Automated Deduction (CADE), Springer. pp. 482–496.

Huang, Z., Van Harmelen, F., Teije, A., 2005. Reasoning with inconsistent ontologies, in: Proceedings of IJCAI, p. 454.

Hustadt, U., Motik, B., Sattler, U., 2004. Reducing SHIQ- description logic to disjunctive datalog programs, in: Proceedings of KR, pp. 152–162.

Kalyanpur, A., Parsia, B., Horridge, M., Sirin, E., 2007. Finding all justifications of OWL DL entailments, in: Proceedings of International Semantic Web Conference (ISWC), pp. 267–280.

Motik, B., Sattler, U., 2006. A comparison of reasoning techniques for querying large description logic aboxes, in: Proceedings of LPAR’06, pp. 227–241.

Motik, B., Shearer, R., Horrocks, I., 2007. Optimized Reasoning in Description Logics using Hyper-tableaux, in: Pfenning, F. (Ed.), Proceedings of Conference on Automated Deduction (CADE), Bremen, Germany, pp. 67–83.

Nebel, B., 1990. Reasoning and revision in hybrid representation systems. volume 422. Springer-Verlag Germany.

Ortiz, M., Calvanese, D., Eiter, T., 2008. Data complexity of query answering in expressive description logics via tableaux. Journal of Automated Reasoning 41, 61–98.

Royer, V., Quantz, J., 1993. Deriving Inference Rules for Description Logics: a Rewriting Approach into Sequent Calculi. Technical Report. Technische Universität Berlin. Germany.

Schaerf, A., 1994. Reasoning with individuals in concept languages. Data and Knowledge Engineering 13, 141–176.

Schmidt-Schauss, M., Smolka, G., 1991. Attributive concept descriptions with complements. Artificial intelligence 48, 1–26.

Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y., 2007. Pellet: A practical owl-dl reasoner. Journal of Web Semantics 5, 51 – 53.

Tsarkov, D., Horrocks, I., 2006. FaCT++ description logic reasoner: System description. Automated Reasoning 4, 292–297.

Wandelt, S., Möller, R., 2012. Towards abox modularization of semi-expressive description logics. Applied Ontology 7, 133–167.