Integrity Constraints in OWL

Jiao Tao
Deptartment of CS, RPI
Troy, NY, USA
taoj2@cs.rpi.edu

Evren Sirin
Clark & Parsia, LLC
Washington, DC, USA
evren@clarkparsia.com

ABSTRACT

In many data-centric applications, it is desirable to use OWL as an expressive schema language with which one expresses constraints that must be satisfied by instance data. However, specific aspects of OWL's standard semantics—i.e., the Open World Assumption (OWA) and the absence of Unique Name Assumption (UNA)—make it difficult to use OWL in this way. In this paper, we present an Integrity Constraint (IC) semantics for OWL axioms, show that IC validation can be reduced to query answering, and present our preliminary results with a prototype implementation using Pellet.

1. MOTIVATION

Standard OWL semantics adopting OWA and non-UNA make it difficult to use OWL for data validation. What triggers a constraint violation in closed world systems leads to new inferences in standard OWL systems. To use OWL both as a knowledge representation language and an integrity constraint language for data validation, we must combine open world reasoning and closed world constraint checking. Tao et al. [5] characterize the typical integrity issues in Semantic Web data using autoepistemic operators, and propose an integrity issue checking solution based on SPARQL queries. Although the semantics of typical integrity constraints are provided in this work, a comprehensive approach describing the semantics of all possible constraints would be preferable. Motik et al. [3] propose a semantics for ICs based on the satisfaction of IC axioms in minimal Herbrand models of the KB. Even though this semantics is reasonable, it has several unsatisfying features: First, constraints can be satisfied by unnamed individuals even if almost all ICs are meant to be satisfied only by named individuals. Second, to define a constraint to be satisfied only by named individuals, a predicate O has to be added into the original constraint axiom, which makes the constraint definition unnecessarily complex. Third, the disjunction axioms and ICs interact in an unintuitive way. Due to these limitations, we need an alternative semantics for OWL axioms used for ICs.

2. IC SEMANTICS

Our approach is inspired by Reiter [4] which argues that ICs are epistemic in nature and are about "what the knowledge base knows". Further investigation is found in [1] where an epistemic extension of DL, $\mathcal{ALCK}$, is proposed: ICs are interpreted with epistemic interpretation ($\mathcal{I}$, $\mathcal{W}$), $(\mathbf{K}C)^{(\mathcal{I},\mathcal{W})}$ (resp. $\mathbf{K}R^{(\mathcal{I},\mathcal{W})}$) represents the individuals (resp. pair of individuals) that are *known* to be instances of C (resp. *known* to be associated by R) in $\mathcal{W}$, and a constraint axiom α is satisfied by $\mathcal{K}$ if for every $\mathcal{I} \in \mathcal{W}$, interpretation ($\mathcal{I}$, $\mathcal{W}$) satisfies α where $\mathcal{W} = Mod(\mathcal{K})$ standing for all models of $\mathcal{K}$. However we find several restrictions of this work making it not suitable to represent the semantics of ICs in expressive DL: (1) It does not address how to interpret $\mathbf{K}C$(resp. $\mathbf{K}R$) when C(resp. R) is not an atomic concept(resp. a simple role). (2) It adopts UNA which is not compatible with OWL. A more resonable solution is to treat two different identifiers as dinstinct if doing this does not cause logical inconsistencies, otherwise they should be interpreted as same. (3) We also find out that allowing $\mathcal{W}$ including all models of $\mathcal{K}$ is not satisfying. It is more intuitive to let $\mathcal{W}$ be the set of models that minimally handle the equality relationship between individuals, i.e., two different individuals are interpreted to be same only when this is necessary for the consistency of KB.

Taking above considerations into account, now we describe an IC semantics for OWL. First we define the notion of IC-interpretation ($\mathcal{I}$, $\mathcal{W}$) where $\mathcal{I}$ is a $\mathcal{SROIQ}$ [2](logic underpinning of OWL2) interpretation and $\mathcal{W}$ is a set of $\mathcal{SROIQ}$ interpretations. Note, for compatibility reasons we use the same representation of epistemic interpretation. Also note, different from epistemic interpretation, IC-interpretation does not adopt UNA. With IC-interpretation, atomic concepts and simple roles are interpreted as follows:

$$C^{(\mathcal{I},\mathcal{W})} = \{a^{\mathcal{I}} \mid a \in N_I \text{ s.t. } \forall \mathcal{J} \in \mathcal{W}, a^{\mathcal{J}} \in C^{\mathcal{J}}\}$$

$$R^{(\mathcal{I},\mathcal{W})} = \{\langle a^{\mathcal{I}}, b^{\mathcal{I}}\rangle \mid a, b \in N_I \text{ s.t. } \forall \mathcal{J} \in \mathcal{W}, \langle a^{\mathcal{J}}, b^{\mathcal{J}}\rangle \in R^{\mathcal{J}}\}$$

where N_I denotes the set of named individuals in KB. It is easy to see that $C^{(\mathcal{I},\mathcal{W})}$ and $R^{(\mathcal{I},\mathcal{W})}$ represent the individuals that are *known* to be instances of C in $\mathcal{W}$, and the pair of individuals that are *known* to be associated by R in $\mathcal{W}$ respectively. The interpretation of complex roles and concept descriptions is described in Figure 1. Note that $(C \sqcup D)^{(\mathcal{I},\mathcal{W})}$ represents the individuals that are known to be instances of C in $\mathcal{W}$ or known to be instances of D in $\mathcal{W}$. We define its semantics this way because it is more intuitive for constraint modeling. Consider $\mathcal{K} = \{C(a), (C_1 \sqcup C_2)(a)\}$ and constraint $C \sqsubseteq C_1 \sqcup C_2$. The more reasonable meaning

$$(R^-)^{(\mathcal{I},\mathcal{W})} = \{\langle x^{\mathcal{I}}, y^{\mathcal{I}}\rangle \mid \langle y^{\mathcal{I}}, x^{\mathcal{I}}\rangle \in R^{(\mathcal{I},\mathcal{W})}\}$$
$$(C \sqcap D)^{(\mathcal{I},\mathcal{W})} = C^{(\mathcal{I},\mathcal{W})} \cap D^{(\mathcal{I},\mathcal{W})}$$
$$(C \sqcup D)^{(\mathcal{I},\mathcal{W})} = C^{(\mathcal{I},\mathcal{W})} \cup D^{(\mathcal{I},\mathcal{W})}$$
$$(\neg C)^{(\mathcal{I},\mathcal{W})} = \Delta \setminus C^{(\mathcal{I},\mathcal{W})}$$
$$(\geq nR.C)^{(\mathcal{I},\mathcal{W})} = \{x^{\mathcal{I}} \mid x \in N_I \text{ s.t. } \#\{y^{\mathcal{I}} \mid \langle x^{\mathcal{I}}, y^{\mathcal{I}}\rangle \in R^{(\mathcal{I},\mathcal{W})} \text{ and } y^{\mathcal{I}} \in C^{(\mathcal{I},\mathcal{W})}\} \geq n\}$$
$$(\leq nR.C)^{(\mathcal{I},\mathcal{W})} = \{x^{\mathcal{I}} \mid x \in N_I \text{ s.t. } \#\{y^{\mathcal{I}} \mid \langle x^{\mathcal{I}}, y^{\mathcal{I}}\rangle \in R^{(\mathcal{I},\mathcal{W})} \text{ and } y^{\mathcal{I}} \in C^{(\mathcal{I},\mathcal{W})}\} \leq n\}$$
$$(\exists R.C)^{(\mathcal{I},\mathcal{W})} = \{x^{\mathcal{I}} \mid x \in N_I \text{ s.t. } \exists y.\langle x^{\mathcal{I}}, y^{\mathcal{I}}\rangle \in R^{(\mathcal{I},\mathcal{W})} \text{ and } y^{\mathcal{I}} \in C^{(\mathcal{I},\mathcal{W})}\}$$
$$(\forall R.C)^{(\mathcal{I},\mathcal{W})} = \{x^{\mathcal{I}} \mid x \in N_I \text{ s.t. } \forall y.\langle x^{\mathcal{I}}, y^{\mathcal{I}}\rangle \in R^{(\mathcal{I},\mathcal{W})} \text{ implies } y^{\mathcal{I}} \in C^{(\mathcal{I},\mathcal{W})}\}$$
$$(\exists R.\text{Self})^{(\mathcal{I},\mathcal{W})} = \{x^{\mathcal{I}} \mid x \in N_I \text{ s.t. } \forall x.\langle x^{\mathcal{I}}, x^{\mathcal{I}}\rangle \in R^{(\mathcal{I},\mathcal{W})}\}$$
$$\{a\}^{(\mathcal{I},\mathcal{W})} = \{a^{\mathcal{I}}\}$$

Figure 1: IC interpretation of complex roles and concepts

of this constraint is "every known instance of C should be a known instance of C_1 or a known instance of C_2". As a result, this constraint is violated by individual `a` since we don't know whether a is an instance of C_1 or C_2. If it is really the intention of the ontology modeler to express a constraint with the meaning of "every known instance of C should be a known instance of C_1 or C_2", we can designate a new name $C' = C_1 \sqcup C_2$ and represent the constraint as $C \sqsubseteq C'$. Also note that $(\neg C)^{(\mathcal{I},\mathcal{W})}$ is interpreted as the individuals which are not known to be instances of C. As a result, $\neg C$ has a closed world meaning with the IC interpretation.

Axiom type	Axiom	Condition on$(\mathcal{I},\mathcal{W})$
Concept inclusion	$C \sqsubseteq D$	$C^{(\mathcal{I},\mathcal{W})} \subseteq D^{(\mathcal{I},\mathcal{W})}$
Role inclusion	$R_1 \sqsubseteq R_2$	$R_1^{(\mathcal{I},\mathcal{W})} \subseteq R_2^{(\mathcal{I},\mathcal{W})}$
R-chain inclusion	$R_1 \ldots R_n \sqsubseteq R$	$R_1^{(\mathcal{I},\mathcal{W})} \circ \ldots \circ R_n^{(\mathcal{I},\mathcal{W})} \subseteq R^{(\mathcal{I},\mathcal{W})}$
Reflexivity	`Ref`(R)	$\forall x \in N_I : \langle x^{\mathcal{I}}, x^{\mathcal{I}}\rangle \in R^{(\mathcal{I},\mathcal{W})}$
Irreflexivity	`Irr`(R)	$\forall x \in N_I : \langle x^{\mathcal{I}}, x^{\mathcal{I}}\rangle \notin R^{(\mathcal{I},\mathcal{W})}$
Role disjointness	`Dis`(R_1, R_2)	$R_1^{(\mathcal{I},\mathcal{W})} \cap R_2^{(\mathcal{I},\mathcal{W})} = \emptyset$
Concept assertion	$C(a)$	$a^{(\mathcal{I},\mathcal{W})} \in C^{(\mathcal{I},\mathcal{W})}$
Role assertion	$R(a,b)$	$\langle a^{(\mathcal{I},\mathcal{W})}, b^{(\mathcal{I},\mathcal{W})}\rangle \in R^{(\mathcal{I},\mathcal{W})}$

The satisfaction of $\mathcal{SROIQ}$ axioms in IC-interpretations is defined in above table where $\mathcal{W} = Mod_{MM_=}(\mathcal{K})$ is defined as follows:

$$\{\mathcal{I} \mid \mathcal{I} \in Mod(\mathcal{K}) \text{ s.t. either } \nexists \mathcal{J} \prec \mathcal{I} \text{ or } \forall \mathcal{J} \prec \mathcal{I}.\mathcal{J} \notin Mod(\mathcal{K})\}$$

where $\mathcal{J} \prec \mathcal{I}$ is true if the following three conditions hold at the same time:

$$\text{For all concepts C}, a^{\mathcal{J}} \in C^{\mathcal{J}} \text{ implies } a^{\mathcal{I}} \in C^{\mathcal{I}};$$
$$\text{For all roles R}, \langle a^{\mathcal{J}}, b^{\mathcal{J}}\rangle \in R^{\mathcal{J}} \text{ implies } \langle a^{\mathcal{I}}, b^{\mathcal{I}}\rangle \in R^{\mathcal{I}};$$
$$\exists a, b \in N_I \text{ s.t. } a^{\mathcal{J}} \neq b^{\mathcal{J}} \text{ and } a^I = b^I.$$

According to above definition, it is easy to see every model in $Mod_{MM_=}(\mathcal{K})$ interprets different identifiers to be same only when it is necessary for the logical consistency of KB. We say that an axiom α is IC-satisfied by a $\mathcal{SROIQ}$ knowledge base $\mathcal{K}$, written as $\mathcal{K} \models_{IC} \alpha$, if for every $\mathcal{I} \in \mathcal{W}$ the IC-interpretation $(\mathcal{I},\mathcal{W})$ satisfies α where $\mathcal{W} = Mod_{MM_=}(\mathcal{K})$. A $\mathcal{SROIQ}$ DL KB $\mathcal{K}$ extended with integrity constraint axioms $\mathcal{C}$ denoted as $\langle \mathcal{K}, \mathcal{C}\rangle$, satisfies $\mathcal{C}$ if $\mathcal{K} \models_{IC} \alpha$ for every $\alpha \in \mathcal{C}$.

3. IMPLEMENTATION OF IC VALIDATION

We find out when the given ontology is a $\mathcal{SHI}$ KB or the constraints are not in the form of cardinality constraints with n>1, IC validation can be reduced to query (with NAF) answering over the KB. In the future, we will provide a formal proof of reducing IC validation to corresponding query answering, and research how to validate ICs in more expressive KBs. We have built a prototype IC validator[1] by extending Pellet[2]. The prototype include a parser, a translator, and a validator for ICs that can read, process and validate ICs written as OWL, OWL2, or SWRL axioms. The IC validator can be accessed via a command-line program or the validation API that validate ICs and output validation results. The prototype is implemented by translating ICs to SPARQL queries and then execute the queries over Pellet reasoner. The performance study shows that our prototype IC validator can be used to efficiently validate relatively large datasets. For instance it only takes 2 seconds to check an IC over a KB containing 100K instances and 800K assertions. To validate arbitrary-sized data, we plan to implement incremental IC validation in the future.

4. REFERENCES

[1] Francesco M. Donini, Maurizio Lenzerini, Daniele Nardi, Werner Nutt, and Andrea Schaerf. An epistemic operator for description logics. *Artificial Intelligence*, 100(1–2):225–274, 1998.

[2] Ian Horrocks, Oliver Kutz, and Ulrike Sattler. The even more irresistible sroiq. In *KR2006*, pages 57–67. AAAI Press, 2006.

[3] Boris Motik, Ian Horrocks, and Ulrike Sattler. Bridging the gap between owl and relational databases. In *WWW2007*, pages 807–816. ACM Press, 2007.

[4] Raymond Reiter. On integrity constraints. In *TARK1988*, pages 97–111, 1988.

[5] Jiao Tao, Li Ding, Jie Bao, and Deborah L. McGuinness. Characterizing and detecting integrity issues in owl instance data. In *OWLED2008*, 2008.

[1] `http://clarkparsia.com/pellet/oicv-0.1.2.zip`
[2] `http://clarkparsia.com/pellet`