First-order Predicate Logic (FOPL) as an Ontology Language

Syntax of FOPL: signature

A signature for FOPL consists of

- a set $\mathcal{V}$ of **variables** denoted by $x, x_0, x_1, \ldots, y, y_0, y_1, \ldots, z, z_0, z_1 \ldots$;
- a set $\mathcal{P}$ of **predicate symbols** each of which comes with a positive number as its arity. Predicates of arity one are called unary predicates and are often denoted by $A, A_0, A_1 \ldots, B, B_0, B_1, \ldots$ In description logic, they correspond to **concept names**, so we sometimes call them concept names. Predicates of arity two are called binary predicates and are often denoted by $r, r_0, r_1, \ldots, s, s_0, s_1, \ldots$. In description logic, they correspond to **roles**, so we sometimes call them roles.
- a "special" binary relation symbols "$=$".

Syntax of FOPL: formulas

Formulas of FOPL are defined by induction:

- If P is a predicate of arity k and $x_1, \ldots, x_k$ are terms, then $P(x_1, \ldots, x_k)$ is a formula;
- in particular, if x_1, x_2 are terms, then $x_1 = x_2$ is a formula;
- If F is a formula, then $\neg F$ is a formula;
- If F and G are formulas, then $F \wedge G$ and $F \vee G$ are formulas;
- If F is a formula and x a variable, then $\forall x.F$ and $\exists x.F$ are formulas.

We abbreviate

- $F_1 \rightarrow F_2 = \neg F_1 \vee F_2$;
- $F_1 \leftrightarrow F_2 = (F_1 \rightarrow F_2) \wedge (F_2 \rightarrow F_1)$.

Translation of concept inclusions into FOPL: idea

- $\textsf{\textbf{Father}} \equiv \textsf{\textbf{Person}} \sqcap \textsf{\textbf{Male}} \sqcap \exists \textsf{\textbf{hasChild}}.\top$.
- $\textsf{\textbf{Student}} \equiv \textsf{\textbf{Person}} \sqcap \exists \textsf{\textbf{is\_registered\_at}}.\textsf{\textbf{University}}$.
- $\textsf{\textbf{Father}} \sqsubseteq \textsf{\textbf{Person}}$.
- $\textsf{\textbf{Father}} \sqsubseteq \exists \textsf{\textbf{hasChild}}.\top$.

are translated to

- $\forall x.(\textsf{\textbf{Father}}(x) \leftrightarrow (\textsf{\textbf{Person}}(x) \land \textsf{\textbf{Male}}(x) \land \exists y.\textsf{\textbf{hasChild}}(x, y)))$.
- $\forall x.(\textsf{\textbf{Student}}(x) \leftrightarrow (\textsf{\textbf{Person}}(x) \land \exists y.(\textsf{\textbf{is\_registered\_at}}(x, y) \land \textsf{\textbf{University}}(y))))$.
- $\forall x.(\textsf{\textbf{Father}}(x) \rightarrow \textsf{\textbf{Person}}(x))$.
- $\forall x.(\textsf{\textbf{Father}}(x) \rightarrow \exists y.\textsf{\textbf{hasChild}}(x, y))$.

Semantics of FOPL: interpretation

An interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ for FOPL consists of

- a non-empty set $\Delta^{\mathcal{I}}$ (the domain);
- an interpretation function that maps
  - every k-ary predicate symbol P to a k-ary relation over $\Delta^{\mathcal{I}}$. In particular,
    - every unary predicate symbol A is mapped to a subset $A^{\mathcal{I}}$ of $\Delta^{\mathcal{I}}$ and
    - every binary predicate symbol r is a mapped to a subset of $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$.

This is almost the same definition as for description logics. The only difference is that we have predicate symbols of arity larger than 2.

Example

Consider

- the concept names **University**, **BritishUniversity**, **Student**,
- the role name **student_at**.

An interpretation $\mathcal{I}_U$ is given by

- $\Delta^{\mathcal{I}_U} = \{\textbf{LU}, \textbf{MU}, \textbf{CMU}, \textbf{Tim}, \textbf{Tom}, \textbf{Rob}, a, b, c\}$.
- $\textbf{University}^{\mathcal{I}_U} = \{\textbf{LU}, \textbf{MU}, \textbf{CMU}, a, b\}$;
- $\textbf{BritishUniversity}^{\mathcal{I}_U} = \{\textbf{LU}, \textbf{MU}\}$;
- $\textbf{Student}^{\mathcal{I}_U} = \{\textbf{Tim}, \textbf{Tom}, c\}$;
- $\textbf{student\_at}^{\mathcal{I}_U} = \{(\textbf{Tim}, \textbf{LU}), (\textbf{Rob}, \textbf{CMU})\}$.

Semantics of FOPL: Satisfaction

A **variable assignment** α for an interpretation $\mathcal{I}$ is a function that maps every variable $x \in \mathcal{V}$ to an element $\alpha(x) \in \Delta^{\mathcal{I}}$.

We define the **satisfaction relation** "$\mathcal{I}, \alpha \models F$" between a formula F and an interpretation $\mathcal{I}$ w.r.t. an assignment α inductively as follows:

- $\mathcal{I}, \alpha \models P(x_1, \ldots, x_k)$ if, and only if, $(\alpha(x_1), \ldots, \alpha(x_k)) \in P^{\mathcal{I}}$;
- $\mathcal{I}, \alpha \models x_1 = x_2$ if, and only if, $\alpha(x_1) = \alpha(x_2)$;
- $\mathcal{I}, \alpha \models \neg F$ if, and only if, not $\mathcal{I}, \alpha \models F$;
- $\mathcal{I}, \alpha \models F \wedge G$ if, and only if, $\mathcal{I}, \alpha \models F$ and $\mathcal{I}, \alpha \models G$;
- $\mathcal{I}, \alpha \models F \vee G$ if, and only if, $\mathcal{I}, \alpha \models F$ or $\mathcal{I}, \alpha \models G$;

Satisfaction continued

- $\mathcal{I}, \alpha \models \exists x.F$ if, and only if, there exists a $d \in \Delta^{\mathcal{I}}$ such that $\mathcal{I}, \alpha[x \mapsto d] \models F$
- $\mathcal{I}, \alpha \models \forall x.F$ if, and only if, for all $d \in \Delta^{\mathcal{I}}$ we have $\mathcal{I}, \alpha[x \mapsto d] \models F$.

where $\alpha[x \mapsto d]$ stands for the assignment that coincides with α except that x is mapped to d.

A variable x is called **free** in a formula F if x has an occurrence in F such that no occurrence of $\exists x$ or $\forall x$ is in front of it. To indicate that the free variables in a formula F are among $x_1, \ldots, x_k$, we sometimes write $F(x_1, \ldots, x_k)$ instead of F.

To test whether $\mathcal{I}, \alpha \models F(x_1, \ldots, x_k)$ it is enough to know the $\alpha(x_1), \ldots, \alpha(x_k)$. Thus, we write

$$\mathcal{I} \models F(a_1, \ldots, a_k)$$

if $\mathcal{I}, \alpha \models F(x_1, \ldots, x_k)$ holds for some assignment (equivalently, all assignments) α with $\alpha(x_1) = a_1, \ldots, \alpha(x_k) = a_k$.

Example $\mathcal{I}_U$

- Let $\alpha(x) = \textbf{LU}$. Then $\mathcal{I}_U, \alpha \models \textbf{University}(x)$. We sometimes denote this by $\mathcal{I}_U \models \textbf{University}(\textbf{LU})$.
- Let $\alpha(x) = \textbf{Tim}$. Then $\mathcal{I}_U, \alpha \not\models \textbf{University}(x)$. We sometimes denote this by $\mathcal{I}_U \not\models \textbf{University}(\textbf{Tim})$.
- Let $\alpha(x) = \textbf{Tim}$ and $\alpha(y) = \textbf{LU}$. Then $\mathcal{I}_U, \alpha \models \textbf{student\_at}(x, y)$. We sometimes denote this by $\mathcal{I}_U \models \textbf{student\_at}(\textbf{Tim}, \textbf{LU})$.
- Let α be arbitrary. Then $\mathcal{I}_U, \alpha \models \exists x.\textbf{University}(x)$; since this does not depend on α, we denote this by $\mathcal{I}_U \models \exists x.\textbf{University}(x)$.
- We have $\mathcal{I}_U \models \exists x.\neg\textbf{University}(x)$.
- We have $\mathcal{I}_U \models \neg\forall x.(\textbf{Student}(x) \rightarrow \exists y.\textbf{student\_at}(x, y))$.

FOPL Ontologies and Reasoning

A closed FOPL formula (also called a sentence) is a FOPL formula without free variables.

An **FOPL ontology** T is a finite set of closed FOPL formulas.

For example, the following translation **MED**$^\sharp$ of the $\mathcal{EL}$ TBoxes MED is an FOPL ontology:

$$\begin{aligned}
\forall x.(\mathbf{Pericardium}(x) &\rightarrow (\mathbf{Tissue}(x) \wedge \exists y.(\mathbf{cont\_in}(x,y) \wedge \mathbf{Heart}(y))))\\
\forall x.(\mathbf{Pericarditis}(x) &\rightarrow (\mathbf{Inflammation}(x) \wedge \exists y.(\mathbf{has\_loc}(x,y) \wedge \mathbf{Pericardium}(y))))\\
\forall x.(\mathbf{Inflammation}(x) &\rightarrow (\mathbf{Disease}(x) \wedge \exists y.(\mathbf{acts\_on}(x,y) \wedge \mathbf{Tissue}(y))))\\
\forall x.(F(x) &\rightarrow (\mathbf{Heartdisease}(x) \wedge \mathbf{NeedTreatment}(x)))
\end{aligned}$$

where

$$F(x) = \mathbf{Disease}(x) \wedge \exists y.(\mathbf{has\_loc}(x,y) \wedge \exists z.(\mathbf{cont\_in}(y,z) \wedge \mathbf{Heart}(z)))$$

FOPL Ontologies and Reasoning

For an interpretation $\mathcal{I}$, we say that $\mathcal{I}$ is a model of T (or, equivalently, that $\mathcal{I}$ satisfies T) and write

$$\mathcal{I} \models T,$$

if, and only if, $\mathcal{I} \models F$ for all $F \in T$.

An FOPL sentence F follows from a FOPL ontology T, in symbols

$$T \models F,$$

if, and only if, for all interpretations $\mathcal{I}$ the following holds: If $\mathcal{I} \models T$, then $\mathcal{I} \models F$.

For example, "Pericarditis needs treatment" follows from **MED** if, and only if,

$$\mathbf{MED}^{\sharp} \models \forall x(\mathbf{Percarditis}(x) \rightarrow \mathbf{NeedsTreatment}(x)).$$

As mentioned above already, for FOPL ontologies T and FOPL sentences F the problem "$T \models F$" is undecidable!

First-order Predicate Logic (FOPL) as a Query Language

Data

A **ground sentence** F is of the form $P(c_1, \ldots, c_n)$, where P is an n-ary predicate and $c_1, \ldots, c_n$ are individual symbols.

A **database instance** is a finite set $\mathcal{D}$ of ground sentences. The set $\mathbf{Ind}(\mathcal{D})$ denote the set of individual symbols in $\mathcal{D}$.

For example, the following set $\mathcal{D}_{\mathsf{uni}}$ is a database instance:

- $\mathbf{University}(\mathbf{LU}), \quad \mathbf{University}(\mathbf{MU}), \quad \mathbf{University}(\mathbf{CMU});$
- $\mathbf{BritishUniversity}(\mathbf{LU})$
- $\mathbf{Student}(\mathbf{Tim}), \quad \mathbf{Student}(\mathbf{Tom});$
- $\mathbf{student\_at}(\mathbf{Tim}, \mathbf{LU}), \quad \mathbf{student\_at}(\mathbf{Rob}, \mathbf{CMU}).$

Here we have

$$\mathbf{Ind}(\mathcal{D}_{\mathsf{uni}}) = \{\mathbf{LU}, \mathbf{MU}, \mathbf{CMU}, \mathbf{Tim}, \mathbf{Tom}, \mathbf{Rob}\}$$

Querying Database Instances

An **n-ary query** Q is a mapping that takes as input a database instance $\mathcal{D}$ and outputs a set $Q(\mathcal{D})$ of n-tuples from $\mathbf{Ind}(\mathcal{D})$.

If $n = 0$, then Q is called a **Boolean query** and outputs either 'Yes' or 'No' or 'Don't know':

$$Q(\mathcal{D}) \in \{ \text{Yes, No, Don't know} \}.$$

Consider the database instance $\mathcal{D}_{\mathsf{uni}}$ with the following queries:

- Boolean Query: Is every British university a university?
- Boolean Query: Does every university have a student?
- Unary query: Output all students that are not students at any university.
- Unary query: Output all universities that are not British universities.

Assumptions

The answers to the queries above depend on our assumptions about the way in which $\mathcal{D}_{\mathsf{uni}}$ represents the world. The main options are:

- Closed World Assumption: if something is not stated explicitly in $\mathcal{D}_{\mathsf{uni}}$ it is false. We assume complete knowledge about the domain: if some ground sentence is true, it is included in $\mathcal{D}_{\mathsf{uni}}$.
- Open World Assumption: if something is not stated in $\mathcal{D}_{\mathsf{uni}}$, then we do not know whether it is true or false. We assume incomplete knowledge about the domain: if some ground sentence is not in $\mathcal{D}_{\mathsf{uni}}$, it can still be true.

In relational databases we make the closed world assumption!

In Schema.org Markup we make the Open World Assumption

Consider the following Schema.org Markup:

```
<div vocab="http://schema.org/" typeof="Movie">
  <h1 property="name">Avatar</h1>
  <div property="director" typeof="Person">
  Director: <span property="name">James Cameron</span>
 (born <time property="birthDate" datetime="1954-08-16">August 16, 1954</time>
  </div>
  <span property="genre">Science fiction</span>
</div>
```

Corresponds to the database instance:

$$\textbf{Movie}(m), \quad \textbf{name}(o, \text{Avatar}), \quad \textbf{director}(m, p)$$

$$\textbf{name}(p, \text{Cameron}), \quad \textbf{birthDay}(p, 1954), \quad \textbf{genre}(m, \text{ScienceFiction})$$

Clearly, we cannot be sure that by collecting markup of webpages we collect all movies!

FOPL Relational Database Querying

We formulate queries as FOPL formulas. The query defined by F is n-ary if F has n free variables. A query is Boolean if it is defined by a closed formula (has no free variable).

We realize the closed world assumption by regarding a database instance as an interpretation.

The interpretation $\mathcal{I}_{\mathcal{D}}$ defined by a database instance $\mathcal{D}$ is given by setting:

- $\Delta^{\mathcal{I}_{\mathcal{D}}} = \textbf{Ind}(\mathcal{D})$;
- For any n-ary predicate P and any $(a_1, \ldots, a_n)$ in $\textbf{Ind}(\mathcal{D})$:

$$(a_1, \ldots, a_n) \in P^{\mathcal{I}_{\mathcal{D}}} \quad \text{if and only if} \quad P(a_1, \ldots, a_n) \in \mathcal{D}$$

Example

The interpretation $\mathcal{J} = \mathcal{I}_{\mathcal{D}_{\textsf{uni}}}$ defined by $\mathcal{D}_{\textsf{uni}}$ is given as follows:

- $\Delta^{\mathcal{J}} = \{\textsf{LU}, \textsf{MU}, \textsf{CMU}, \textsf{Tim}, \textsf{Tom}, \textsf{Rob}\}$;
- $\textsf{University}^{\mathcal{J}} = \{\textsf{LU}, \textsf{MU}, \textsf{CMU}\}$;
- $\textsf{BritishUniversity}^{\mathcal{J}} = \{\textsf{LU}\}$;
- $\textsf{Student}^{\mathcal{J}} = \{\textsf{Tim}, \textsf{Tom}\}$;
- $\textsf{student\_at}^{\mathcal{J}} = \{(\textsf{Tim}, \textsf{LU}), (\textsf{Rob}, \textsf{CMU})\}$.

FOPL Relational Database Queries (Closed World Semantics)

A tuple $(a_1, \ldots, a_n)$ in $\mathbf{Ind}(\mathcal{D})$ is an **answer** to an FOPL query $F(x_1, \ldots, x_n)$ in database instance $\mathcal{D}$ if

$$\mathcal{I}_{\mathcal{D}} \models F(a_1, \ldots, a_n)$$

We set

$$\mathbf{answer}(F(x_1, \ldots, x_n), \mathcal{D}) = \{(a_1, \ldots, a_n) \mid \mathcal{I}_{\mathcal{D}} \models F(a_1, \ldots, a_n)\}$$

The answer to a Boolean FOPL query F in $\mathcal{D}$ is 'Yes' if

$$\mathcal{I}_{\mathcal{D}} \models F$$

The answer to a Boolean FOPL query F in $\mathcal{D}$ is 'No' if

$$\mathcal{I}_{\mathcal{D}} \not\models F$$

Note that the answer 'Don't know' is never returned!

Examples

We query $\mathcal{D}_{\textsf{uni}}$ under closed world semantics:

- Output all universities. The corresponding query is

  $$F(x) = \textbf{University}(x).$$

  The set of answers is $\textbf{answer}(F(x), \mathcal{D}_{\textsf{uni}}) = \{\textbf{LU}, \textbf{MU}, \textbf{CMU}\}$

- Is every British university a university? The corresponding query is

  $$F = \forall x(\textbf{BritishUniversity}(x) \rightarrow \textbf{University}(x))$$

  The answer is 'Yes'.

Examples

- Does every university have a student? The corresponding query is

  $$F = \forall x(\textbf{University}(x) \rightarrow \exists y \textbf{student\_at}(y, x))$$

  The answer is 'No'.

- Output all students that are not students at any university. The corresponding query is

  $$F(x) = \textbf{Student}(x) \wedge \neg\exists y \textbf{student\_at}(x, y)$$

  The set of answers is $\textbf{answer}(F(x), \mathcal{D}_{\textsf{uni}}) = \{\textbf{Tom}\}$

- Output all universities that are not British universities. The corresponding query is

  $$F(x) = \textbf{University}(x) \wedge \neg\textbf{BritishUniversity}(x)$$

  The set of answers is $\textbf{answer}(F(x), \mathcal{D}_{\textsf{uni}}) = \{\textbf{MU}, \textbf{CMU}\}$.

Complexity of querying relational databases

Consider, for simplicity, Boolean queries. There are two different ways of measuring the complexity of querying relational databases (RDBs):

- **Data complexity**: This measures the time/space needed (in the worst case) to evaluate a fixed query F in an instance $\mathcal{I}$ of a RDB (i.e., to decide whether $\mathcal{I} \models F$). Thus, when measuring data complexity, the input variable is the size of the instance and the query is assumed to be fixed.

- **Combined Complexity**: This measures the time/space needed (in the worst case) to evaluate a query F in an instance $\mathcal{I}$ of a RDB. Thus, when measuring combined complexity, both the size of the instance and the query are input variables.

Data complexity is regarded as being more relevant as queries are typically rather small, but database instances are very large compared to the query.

FOPL: Data complexity is in LogSpace. So it is in polytime and, therefore, tractable. Combined complexity is PSpace-complete, and, therefore, not tractable.

FOPL Query Answering (Open World Semantics)

Let $\mathcal{D}$ be a database instance. An interpretation $\mathcal{I}$ is a model of $\mathcal{D}$ if

- $\mathbf{Ind}(\mathcal{D}) \subseteq \Delta^{\mathcal{I}}$;
- If $P(a_1, \ldots, a_n) \in \mathcal{D}$, then $(a_1, \ldots, a_n) \in P^{\mathcal{I}}$.

The set of models of $\mathcal{D}$ is denoted by $\mathbf{Mod}(\mathcal{D})$.

Let $F(x_1, \ldots, x_n)$ be an FOPL query. Then $(a_1, \ldots, a_n)$ in $\mathbf{Ind}(\mathcal{D})$ is a **certain answer** to $F(x_1, \ldots, x_n)$ in $\mathcal{D}$, in symbols

$$\mathcal{D} \models F(a_1, \ldots, a_n),$$

if $\mathcal{I} \models F(a_1, \ldots, a_n)$ for all $\mathcal{I} \in \mathbf{Mod}(\mathcal{D})$.

The set of certain answers to $F(x_1, \ldots, x_n)$ in $\mathcal{D}$ is

$$\mathbf{certanswer}(F(x_1, \ldots, x_n), \mathcal{D}) = \{(a_1, \ldots, a_n) \mid \mathcal{D} \models F(a_1, \ldots, a_n)\}$$

FOPL Query Answering (Open World Semantics)

- 'Yes' is the certain answer to a Boolean query $\boldsymbol{F}$ if $\mathcal{I} \models \boldsymbol{F}$ for all $\mathcal{I} \in \mathbf{Mod}(\mathcal{D})$.
- 'No' is the certain answer to a Boolean query $\boldsymbol{F}$ if $\mathcal{I} \not\models \boldsymbol{F}$ for all $\mathcal{I} \in \mathbf{Mod}(\mathcal{D})$
- If neither 'Yes' nor 'No' is a certain answer, then we say that the certain answer is 'Don't know'.

Note that under closed world semantics the answer to a Boolean query is always either 'Yes' or 'No'.

Example

We consider again the database instance $\mathcal{D}_{\mathsf{uni}}$:

- $\mathbf{University}(\mathbf{LU}), \quad \mathbf{University}(\mathbf{MU}), \quad \mathbf{University}(\mathbf{CMU});$
- $\mathbf{BritishUniversity}(\mathbf{LU})$
- $\mathbf{Student}(\mathbf{Tim}), \quad \mathbf{Student}(\mathbf{Tom});$
- $\mathbf{student\_at}(\mathbf{Tim}, \mathbf{LU}), \quad \mathbf{student\_at}(\mathbf{Rob}, \mathbf{CMU}).$

Example

Output all universities. The corresponding query is

$$F(x) = \mathbf{University}(x).$$

The set of certain answers is

$$\mathbf{certanswer}(F(x), \mathcal{D}_{\mathsf{uni}}) = \{\mathbf{LU}, \mathbf{MU}, \mathbf{CMU}\}$$

This coincides with $\mathbf{answer}(F(x), \mathcal{D}_{\mathsf{uni}})$. To see this, we have to show:

- for all $\mathcal{I} \in \mathbf{Mod}(\mathcal{D}_{\mathsf{uni}})$: $\mathcal{I} \models F(\mathbf{LU})$, $\mathcal{I} \models F(\mathbf{MU})$, and $\mathcal{I} \models F(\mathbf{CMU})$.
- there exists $\mathcal{I} \in \mathbf{Mod}(\mathcal{D}_{\mathsf{uni}})$ with $\mathcal{I} \not\models F(\mathbf{Tim})$;
- there exists $\mathcal{I} \in \mathbf{Mod}(\mathcal{D}_{\mathsf{uni}})$ with $\mathcal{I} \not\models F(\mathbf{Tom})$.
- there exists $\mathcal{I} \in \mathbf{Mod}(\mathcal{D}_{\mathsf{uni}})$ with $\mathcal{I} \not\models F(\mathbf{Rob})$.

For Point 2 to 4 we just take $\mathcal{I}_{\mathcal{D}_{\mathsf{uni}}}$!

Example

Is every British university a university? The corresponding query is

$$F = \forall x(\textbf{BritishUniversity}(x) \rightarrow \textbf{University}(x))$$

The certain answer to F in $\mathcal{D}_{\text{uni}}$ is 'Don't know'. To see this we have to construct two interpretations $\mathcal{I} \in \textbf{Mod}(\mathcal{D}_{\text{uni}})$ and $\mathcal{J} \in \textbf{Mod}(\mathcal{D}_{\text{uni}})$ such that

- $\mathcal{I} \models F$;
- $\mathcal{J} \not\models F$.

For Point 1 we can take $\mathcal{I}_{\mathcal{D}_{\text{uni}}}$. For Point 2, we expand $\mathcal{I}_{\mathcal{D}_{\text{uni}}}$ by adding a new individual a to its domain $\Delta^{\mathcal{J}}$ and let $a \in \textbf{BritishUniversity}^{\mathcal{J}}$ and $a \notin \textbf{University}^{\mathcal{J}}$.

Example

Does every university have a student? The corresponding query is

$$F = \forall x(\textbf{University}(x) \rightarrow \exists y \textbf{student\_at}(y, x))$$

The certain answer to F in $\mathcal{D}_{\textsf{uni}}$ is ‘Don't know’. To see this we have to construct two interpretations $\mathcal{I} \in \textbf{Mod}(\mathcal{D}_{\textsf{uni}})$ and $\mathcal{J} \in \textbf{Mod}(\mathcal{D}_{\textsf{uni}})$ such that

- $\mathcal{I} \models F$;
- $\mathcal{J} \not\models F$.

For Point 2 we can take $\mathcal{J} = \mathcal{I}_{\mathcal{D}_{\textsf{uni}}}$ (since **MU** does not have a student in $\mathcal{I}_{\mathcal{D}_{\textsf{uni}}}$). For Point 1, we expand $\mathcal{I}_{\mathcal{D}_{\textsf{uni}}}$ to an interpretation $\mathcal{I}$ by adding a new individual a to its domain $\Delta^{\mathcal{I}}$ and let $(a, \textbf{MU}) \in \textbf{student\_at}^{\mathcal{I}}$.

Example

Output all students that are not students at any university. The corresponding query is

$$F(x) = \textbf{Student}(x) \land \neg\exists y \textbf{student\_at}(x, y)$$

The set of certain answers is $\textbf{certanswer}(F(x), \mathcal{D}_{\textsf{uni}}) = \emptyset$.

To see this, observe that $\mathcal{I}_{\mathcal{D}_{\textsf{uni}}}$ shows already that only **Tom** could be a certain answer. But **Tom** is not a certain answer since the interpretation $\mathcal{I}$ that expands $\mathcal{I}_{\mathcal{D}_{\textsf{uni}}}$ by adding $(\textbf{Tom}, \textbf{MU})$ to $\textbf{student\_at}^{\mathcal{I}}$ shows that there is a model of $\mathcal{D}_{\textsf{uni}}$ in which **Tom** is a student at some university.

Example

Output all universities that are not British universities. The corresponding query is

$$F(x) = \textbf{University}(x) \land \neg \textbf{BritishUniversity}(x)$$

The set of answers is $\textbf{certanswer}(F(x), \mathcal{D}_{\textsf{uni}}) = \emptyset$.

To see this, expand $\mathcal{I}_{\mathcal{D}_{\textsf{uni}}}$ by adding **MU** and **CMU** to $\textbf{BritishUniversity}^{\mathcal{I}}$. Then $\mathcal{I}$ is a model of $\mathcal{D}_{\textsf{uni}}$ in which there is no university that is not a British university.

The Relationship between Relational Databases and Ontologies

- FOPL database querying (closed world assumption) means checking

$$\mathcal{I} \models F$$

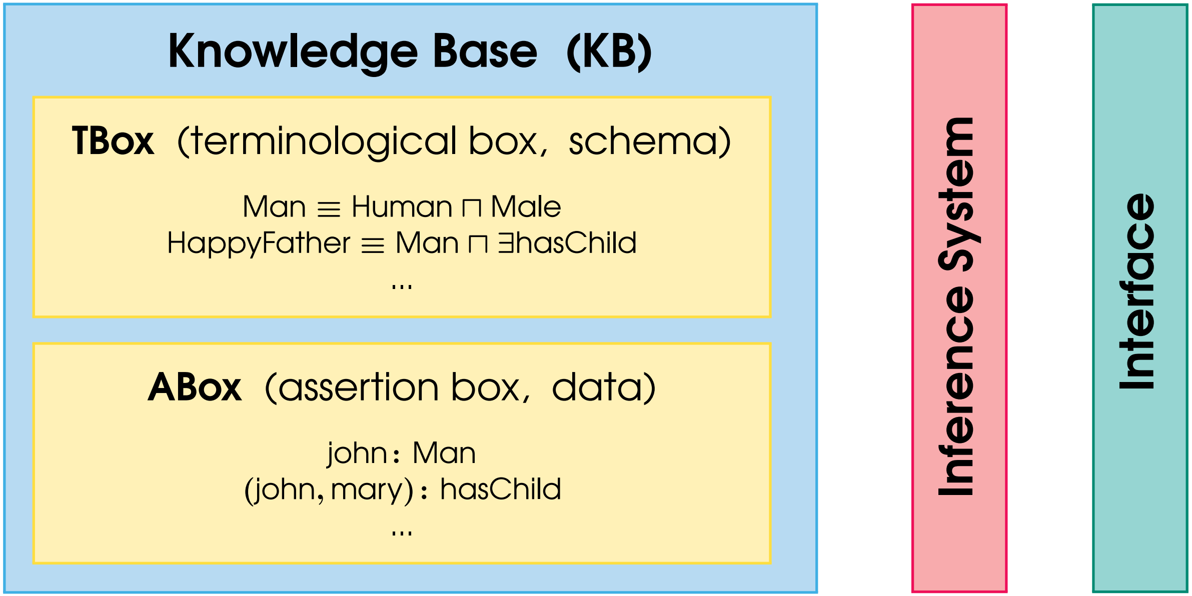
- FOPL and Description Logic ontology querying means checking

$$T \models F$$

- Typically, an FOPL or Description Logic ontology T has many different interpretations (many different $\mathcal{I}$ such that $\mathcal{I} \models T$). Data instances $\mathcal{D}$ have only one interpretation (under closed world assumption).
- FOPL and Description Logic ontologies make the open world assumption. For example, often individuals are not mentioned at all.

Using ABoxes to store Data

ABoxes (Assertion Boxes)



Assertion Box (ABox)

Let $\mathcal{L}$ be a description logic. A **$\mathcal{L}$-ABox** is a finite set $\mathcal{A}$ of assertions of the form

$$C(a), \quad r(a,b),$$

where C is an $\mathcal{L}$-concept, r a role name, and a, b are individual names.

- $C(a)$ says that a is an instance of C;
- $r(a,b)$ says that (a,b) is an instance of r.

ABoxes generalize database instances in which only ground sentences

$$A(a), \quad r(a,b)$$

with A a concept name and r a role name are allowed. We sometimes call ABoxes that are database instances **simple ABoxes**.

Semantics for ABoxes (Open World Assumption)

Let $\mathcal{A}$ be an ABox. By **Ind**$(\mathcal{A})$ we denote the set of individual names in $\mathcal{A}$. An interpretation $\mathcal{I}$ is a model of $\mathcal{A}$, in symbols $\mathcal{I} \models \mathcal{A}$, if

- **Ind**$(\mathcal{A}) \subseteq \Delta^{\mathcal{I}}$;
- If $C(a) \in \mathcal{A}$, then $a \in C^{\mathcal{I}}$;
- If $r(a, b) \in \mathcal{A}$, then $(a, b) \in r^{\mathcal{I}}$.

The set of models of $\mathcal{A}$ is denoted by **Mod**$(\mathcal{A})$.

Let $F(x_1, \ldots, x_n)$ be an FOPL query. Then $(a_1, \ldots, a_n)$ in **Ind**$(\mathcal{A})$ is a **certain answer** to $F(x_1, \ldots, x_n)$ in $\mathcal{A}$, in symbols

$$\mathcal{A} \models F(a_1, \ldots, a_n),$$

if $\mathcal{I} \models F(a_1, \ldots, a_n)$ for all $\mathcal{I} \in$ **Mod**$(\mathcal{A})$.

The set of certain answers to $F(x_1, \ldots, x_n)$ in $\mathcal{A}$ is

$$\textbf{certanswer}(F(x_1, \ldots, x_n), \mathcal{A}) = \{(a_1, \ldots, a_n) \mid \mathcal{A} \models F(a_1, \ldots, a_n)\}$$

FOPL Query Answering (Open World Semantics)

- 'Yes' is the certain answer to a Boolean query $\boldsymbol{F}$ if $\mathcal{I} \models \boldsymbol{F}$ for all $\mathcal{I} \in \mathbf{Mod}(\mathcal{A})$.
- 'No' is the certain answer to a Boolean query $\boldsymbol{F}$ if $\mathcal{I} \not\models \boldsymbol{F}$ for all $\mathcal{I} \in \mathbf{Mod}(\mathcal{A})$
- If neither 'Yes' nor 'No' is a certain answer, then we say that the certain answer is 'Don't know'.

What is the answer to this query?

Consider the ABox $\mathcal{A}$:

1. friend(john, susan)
2. friend(john, andrea)
3. loves(susan, andrea)
4. loves(andrea, bill)
5. Female(susan)
6. ¬Female(bill)

Does John have a female friend who is in love with a not female person?

The corresponding Boolean FOPL query is

$$F = \exists x.(\text{friend}(\text{john}, x) \wedge \text{Female}(x) \wedge \exists y.(\text{loves}(x, y) \wedge \neg\text{Female}(y)))$$

or, in description logic notation:

$$\exists\text{friend}.(\text{Female} \sqcap \exists\text{loves}.\neg\text{Female})(\text{john})$$

Answers: Example

Let

$$\mathcal{A} = \{ \text{Male}(\text{harry}), \text{hasChild}(\text{peter}, \text{harry}) \}$$

The answer to the query "Are all children of Peter male?", in symbols

$$F = \forall x.(\text{hasChild}(\text{peter}, x) \to \text{Male}(x)),$$

given by $\mathcal{A}$ is "don't know".

In order to prevent this, we could add

- $\forall$hasChild.Male(peter)
- or (≤ 1 hasChild .$\top$)(peter)

to the ABox $\mathcal{A}$.

3-Colorability

A graph G is a pair (W, E) consisting of a set W and a symmetric relation E on W.

G is 3-colorable if there exist subsets **blue**, **red**, and **green** of W such that

- the sets **blue**, **green**, and **red** are mutually disjoint;
- **blue** $\cup$ **red** $\cup$ **green** $= W$;
- if $(a, b) \in E$, then a and b do not have the same color.

3-colorability of graphs is an NP-complete problem.

3-Colorability as a Query Answering Problem

Assume $G = (W, E)$ is given. Construct the ABox $\mathcal{A}$ by taking a role name r and concept names **Blue**, **Green**, and **Red** and setting

- $r(a, b) \in \mathcal{A}$ for all $a, b \in W$ with $(a, b) \in E$.
- **Blue** $\sqcup$ **Green** $\sqcup$ **Red**$(a) \in \mathcal{A}$ for all $a \in W$.
- $($**Blue** $\rightarrow \forall r.($**Red** $\sqcup$ **Green**$))(a) \in \mathcal{A}$, for all $a \in W$;
- $($**Red** $\rightarrow \forall r.($**Blue** $\sqcup$ **Green**$))(a) \in \mathcal{A}$, for all $a \in W$;
- $($**Green** $\rightarrow \forall r.($**Red** $\sqcup$ **Blue**$))(a) \in \mathcal{A}$, for all $a \in W$.

Define query F by setting

$$F = \exists x((\mathbf{Blue}(x) \wedge \mathbf{Red}(x)) \vee (\mathbf{Blue}(x) \wedge \mathbf{Green}(x)) \vee (\mathbf{Red}(x) \wedge \mathbf{Green}(x))$$

Then G is not 3-colorable if, and only if, the certain answer to F in $\mathcal{A}$ is 'Yes'.

Thus, query answering is coNP-hard (the complement of NP) in data complexity!

Using the $\mathcal{ALC}$ Tableau to Answer Queries

Consider an $\mathcal{ALC}$ ABox $\mathcal{A}$ and a query of the form $C(x)$, where C is an $\mathcal{ALC}$ concept. Assume $a \in \mathbf{Ind}(\mathcal{A})$ is given. We want to know whether

$$a \in \mathbf{certanswer}(C(x), \mathcal{A}),$$

in other words, we want to know whether $a \in C^{\mathcal{I}}$ for all interpretations $\mathcal{I} \in \mathbf{Mod}(\mathcal{A})$.

We can reformulate this problem as follows: Let $\mathcal{A}' = \mathcal{A} \cup \{\neg C(a)\}$. Then $a \in \mathbf{certanswer}(C(x), \mathcal{A})$ if there does not exist any model of $\mathcal{A}'$.

Tableau Algorithm Deciding whether $\mathcal{A}$ has a Model

Consider $\mathcal{ALC}$ ABox $\mathcal{A}$. We may assume that each concept D in $\mathcal{A}$ is in negation normal form and obtain the constraint system $\mathcal{A}^*$ as the set of constraints

- $a : C$ for all $C(a) \in \mathcal{A}$;
- $(a, b) : r$ for all $r(a, b) \in \mathcal{A}$.

Then $\mathcal{A}$ has a model if, and only if, starting from $\mathcal{A}^*$ there is a sequence of completion rule applications that terminates with a set of constraints containing no clash.

Example

Consider again the ABox $\mathcal{A}$:

1. $\mathsf{friend}(\mathsf{john}, \mathsf{susan})$
2. $\mathsf{friend}(\mathsf{john}, \mathsf{andrea})$
3. $\mathsf{loves}(\mathsf{susan}, \mathsf{andrea})$
4. $\mathsf{loves}(\mathsf{andrea}, \mathsf{bill})$
5. $\mathsf{Female}(\mathsf{susan})$
6. $\neg\mathsf{Female}(\mathsf{bill})$

Does John have a female friend who is in love with a not female person?

Thus, we want to know whether 'Yes' is the certain answer to the query:

$$\exists \mathsf{friend}.(\mathsf{Female} \sqcap \exists \mathsf{loves}.\neg\mathsf{Female})(\mathsf{john})$$

Example

To this end we check whether

$$\mathcal{A} \cup \{\neg\exists \text{friend}.(\text{Female} \sqcap \exists \text{loves}.\neg\text{Female})(\text{john})\}$$

has a model. If not, then 'Yes' is indeed the certain answer to

$$\exists \text{friend}.(\text{Female} \sqcap \exists \text{loves}.\neg\text{Female})(\text{john})$$

Transformation into negation normal form gives:

$$\forall \text{friend}.(\neg\text{Female} \sqcup \forall \text{loves}.\text{Female})(\text{john})$$

Example

Thus, we apply the tableau to the constraint system

$$\mathcal{A}^* \cup \{\mathsf{john} : \forall \mathsf{friend}.(\neg \mathsf{Female} \sqcup \forall \mathsf{loves}.\mathsf{Female})\}$$

given by

1. $(\mathsf{john}, \mathsf{susan}) : \mathsf{friend}$
2. $(\mathsf{john}, \mathsf{andrea}) : \mathsf{friend}$
3. $(\mathsf{susan}, \mathsf{andrea}) : \mathsf{loves}$
4. $(\mathsf{andrea}, \mathsf{bill}) : \mathsf{loves}$
5. $\mathsf{susan} : \mathsf{Female}$
6. $\mathsf{bill} : \neg \mathsf{Female}$
7. $\mathsf{john} : \forall \mathsf{friend}.(\neg \mathsf{Female} \sqcup \forall \mathsf{loves}.\mathsf{Female})$

Example

Two applications of the rule $\rightarrow_{\forall}$ give the additional constraints:

$$\text{susan} : (\neg\text{Female} \sqcup \forall\text{loves}.\text{Female})$$

and

$$\text{andrea} : (\neg\text{Female} \sqcup \forall\text{loves}.\text{Female})$$

We now apply the rule $\rightarrow_{\vee}$ to the first constraint:

- Adding the constraint susan : ¬Female results in a clash since we have already susan : Female $\in \mathcal{A}^*$.
- Thus we add the constraint susan : ∀loves.Female to the constraint system.

Example

We now apply $\rightarrow_\forall$ to

$$\mathsf{susan} : \forall \mathsf{loves}.\mathsf{Female}, \quad (\mathsf{susan}, \mathsf{andrea}) : \mathsf{loves}$$

and add

$$\mathsf{andrea} : \mathsf{Female}$$

to the constraint system. We apply $\rightarrow_\vee$ to

$$\mathsf{andrea} : (\neg \mathsf{Female} \sqcup \forall \mathsf{loves}.\mathsf{Female})$$

- Adding $\mathsf{andrea} : \neg \mathsf{Female}$ to the constraint systems results in a clash since $\mathsf{andrea} : \mathsf{Female}$ is in the constraint system.

- Thus we add the constraint $\mathsf{andrea} : \forall \mathsf{loves}.\mathsf{Female}$ to the constraint system.

Example

Now we apply $\rightarrow_\forall$ to

$$\mathsf{andrea} : \forall \mathsf{loves}.\mathsf{Female}, \quad (\mathsf{andrea}, \mathsf{bill}) : \mathsf{loves}$$

and add

$$\mathsf{bill} : \mathsf{Female}$$

to the constraint system. But this results in a clash since $\mathsf{bill} : \neg\mathsf{Female}$ is already in the constraint system.

It follows that every sequence of completion rule application results in a clash.

Ontology Based Data Access

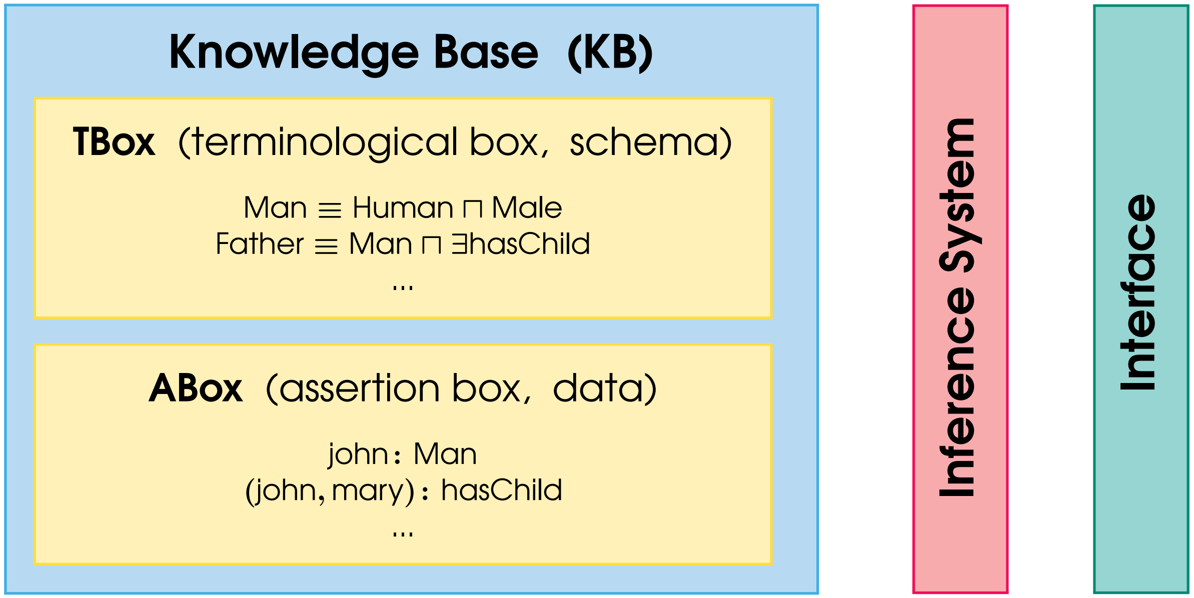
Vision: Ontologies at the Core of Information Systems

- Usage of all system resources (data and services) is done through a domain conceptualization.
- Cooperation between systems is done at the level of the conceptualizations.
- This implies:
  - Hide to the user where and how data and services are stored or implemented;
  - Present to the user a conceptual view of the data and services.

Ontology based Data Access

- An ontology provides meta-information about the data and the vocabulary used to query the data. It can impose constraints on the data.
- Actual data can be incomplete w.r.t. such meta-information and constraints. So data should be stored using open world semantics rather than closed world semantics: use ABoxes instead of relational database instances.
- During query answering, the system has to take into account the ontology.

We discuss ontology based data access in the framework of description logic knowledge bases.


Knowledge Base (KB)
TBox (terminological box, schema)
$\text{Man} \equiv \text{Human} \sqcap \text{Male}$
$\text{Father} \equiv \text{Man} \sqcap \exists\text{hasChild}$
...
ABox (assertion box, data)
$\text{john} : \text{Man}$
$(\text{john}, \text{mary}) : \text{hasChild}$
...
Inference System
Interface

Knowledge Base (= Ontology with database instance)

A **knowledge base** $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ consists of a TBox $\mathcal{T}$ and a simple ABox $\mathcal{A}$ (or, equivalently, a database instance).

We combine the open world semantics for TBoxes and ABoxes in the obvious manner, and obtain an **open world semantics** for knowledge bases.

An interpretation $\mathcal{I}$ **satisfies** a knowledge base $(\mathcal{T}, \mathcal{A})$, in symbols

$$\mathcal{I} \models (\mathcal{T}, \mathcal{A}),$$

if it satisfies both $\mathcal{T}$ and $\mathcal{A}$. In this case we also say that $\mathcal{I}$ is a **model** of $(\mathcal{T}, \mathcal{A})$. The set of models of $(\mathcal{T}, \mathcal{A})$ is denoted by $\mathbf{Mod}(\mathcal{T}, \mathcal{A})$.

Certain Answers

Given a knowledge base $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ and an FOPL query $F(x_1, \ldots, x_k)$, we say that $(a_1, \ldots, a_k)$ is a **certain answer** to $F(x_1, \ldots, x_k)$ by $\mathcal{K}$, in symbols

$$\mathcal{K} \models F(a_1, \ldots, a_k),$$

if

- $a_1, \ldots, a_k$ are individual names in $\mathcal{A}$;
- for all interpretations $\mathcal{I}$:

$$\mathcal{I} \models \mathcal{K} \quad \Rightarrow \quad \mathcal{I} \models F(a_1, \ldots, a_k).$$

The set of certain answers given to F by $\mathcal{K}$ is defined as:

$$\textbf{certanswer}(F, \mathcal{K}) = \{(a_1, \ldots, a_k) \mid \mathcal{K} \models F(a_1, \ldots, a_k)\}$$

Boolean Queries

Let $\mathcal{K}$ be a knowledge base. For a query $\boldsymbol{F}$ without variables (Boolean query), we say that

- the certain answer given by $\mathcal{K}$ is "yes" if $\mathcal{I} \models \boldsymbol{F}$, for all interpretations $\mathcal{I}$ satisfying $\mathcal{K}$;
- the certain answer given by $\mathcal{K}$ is "no" if $\mathcal{I} \not\models \boldsymbol{F}$, for all interpretations $\mathcal{I}$ satisfying $\mathcal{K}$.
- Otherwise the certain answer is: "Don't know".

Example

Consider the TBox $\mathcal{T}_U$:

- $\textbf{BritishUniversity} \sqsubseteq \textbf{University}$;
- $\textbf{University} \sqcap \textbf{Student} \sqsubseteq \bot$;
- $\top \sqsubseteq \forall \textbf{registered\_at.University}$;
- $\top \sqsubseteq \forall \textbf{student\_at.University}$;
- $\exists \textbf{student\_at.}\top \sqsubseteq \textbf{Student}$;
- $\textbf{Student} \sqsubseteq \exists \textbf{student\_at.}\top$;
- $\textbf{NonBritishUni} \equiv \textbf{University} \sqcap \neg \textbf{BritishUniversity}$.

Example (continued)

and the simple ABox (equivalently, database instance) $\mathcal{A}$:

- $\mathsf{NonBritishUni}(\mathsf{CMU})$
- $\mathsf{Institution}(\mathsf{Harvard}), \quad \mathsf{Institution}(\mathsf{FUBerlin})$
- $\mathsf{BritishUniversity}(\mathsf{LU}), \quad \mathsf{BritishUniversity}(\mathsf{MU})$
- $\mathsf{Student}(\mathsf{Tim})$
- $\mathsf{registered}(\mathsf{Tim}, \mathsf{LU}), \quad \mathsf{registered}(\mathsf{Bob}, \mathsf{MU})$
- $\mathsf{student\_at}(\mathsf{Tom}, \mathsf{Harvard})$

Example (continued)

Denote by $\mathcal{I}_{\mathcal{A}}$ the interpretation corresponding to the database instance $\mathcal{A}$:

- $\Delta^{\mathcal{I}_{\mathcal{A}}} = \{\textsf{CMU}, \textsf{Harvard}, \textsf{FUBerlin}, \textsf{Tim}, \textsf{Tom}, \textsf{Bob}, \textsf{MU}, \textsf{LU}\}$;
- $\textsf{NonBritishUni}^{\mathcal{I}_{\mathcal{A}}} = \{\textsf{CMU}\}$;
- $\textsf{Institution}^{\mathcal{I}_{\mathcal{A}}} = \{\textsf{Harvard}, \textsf{FUBerlin}\}$;
- $\textsf{BritishUniversity}^{\mathcal{I}_{\mathcal{A}}} = \{\textsf{LU}, \textsf{MU}\}$;
- $\textsf{Student}^{\mathcal{I}_{\mathcal{A}}} = \{\textsf{Tim}\}$;
- $\textsf{registered\_at}^{\mathcal{I}_{\mathcal{A}}} = \{(\textsf{Tim}, \textsf{LU}), (\textsf{Bob}, \textsf{MU})\}$;
- $\textsf{student\_at}^{\mathcal{I}_{\mathcal{A}}} = \{(\textsf{Tom}, \textsf{Harvard})\}$.

(Certain) Answers

In the table below, we consider Boolean queries $C(a)$ (in description logic notation!) and give the (certain) answer to $C(a)$ of the database instance $\mathcal{I}_{\mathcal{A}}$, the ABox $\mathcal{A}$, and the knowledge base $\mathcal{K}_U = (\mathcal{T}_U, \mathcal{A})$.

Boolean Query	$\mathcal{I}_{\mathcal{A}}$	Abox $\mathcal{A}$	KB $\mathcal{K}_U$
University(CMU)	No	Don't know	Yes
University(Harvard)	No	Don't know	Yes
NonBritishUni(CMU)	Yes	Yes	Yes
Student(Tim)	Yes	Yes	Yes
Student(Tom)	No	Don't know	Yes
$\exists$**student_at.**$\top$**(Tom)**	Yes	Yes	Yes
$\exists$**student_at.**$\top$**(Tim)**	No	Don't know	Yes
(Student $\sqcap \neg$**University)(Tim)**	Yes	Don't know	Yes
(Institution $\sqcap \neg$**University)(FUBerlin)**	Yes	Don't know	Don't know

Example

Let $\mathcal{S} = (\mathcal{O}, \mathcal{B})$ be a knowledge base with simple ABox $\mathcal{B}$ given by

$$\mathbf{Person(john)}, \quad \mathbf{Person(nick)}, \quad \mathbf{Person(toni)}$$

$$\mathbf{hasFather(john, nick)}, \quad \mathbf{hasFather(nick, toni)}$$

and TBox $\mathcal{O}$ defined as

$$\mathcal{O} = \{\mathbf{Person} \sqsubseteq \exists \mathbf{has\_Father.Person}\}$$

For the FOPL query

$$F(x, y) = \mathbf{hasFather}(x, y)$$

we obtain

$$\mathbf{certanswer}(F, \mathcal{S}) = \{(\mathbf{john, nick}), (\mathbf{nick, toni})\}.$$

Example

- For the query

$$F(x) = \exists y.\mathsf{hasFather}(x, y)$$

  we obtain

$$\mathsf{certanswer}(F(x), \mathcal{S}) = \{\mathsf{john}, \mathsf{nick}, \mathsf{toni}\}$$

- For the query

$$F(x) = \exists y_1 \exists y_2 \exists y_3.(\mathsf{hasFather}(x, y_1) \wedge \mathsf{hasFather}(y_1, y_2) \wedge \mathsf{hasFather}(y_2, y_3))$$

  we obtain

$$\mathsf{certanswer}(F(x), \mathcal{S}) = \{\mathsf{john}, \mathsf{nick}, \mathsf{toni}\}$$

- For the query

$$F(x, y_3) = \exists y_1 \exists y_2.(\mathsf{hasFather}(x, y_1) \wedge \mathsf{hasFather}(y_1, y_2) \wedge \mathsf{hasFather}(y_2, y_3))$$

  we obtain

$$\mathsf{certanswer}(F(x, y_3), \mathcal{S}) = \emptyset$$

Complexity of querying $(\mathcal{T}, \mathcal{A})$

Consider, for simplicity, Boolean queries. There are two different ways of measuring the complexity of querying:

- Data complexity: Measures the time/space needed to evaluate a fixed query $\boldsymbol{F}$ for a fixed TBox $\mathcal{T}$ in $(\mathcal{T}, \mathcal{A})$ (i.e., check $\mathcal{T}, \mathcal{A}) \models \boldsymbol{F}$). The only input variable is the size of $\mathcal{A}$.

- Combined complexity: Measure the time/space needed to evaluate a query in $(\mathcal{T}, \mathcal{A})$. The input variables are the size of the query, the size of $\mathcal{T}$, and the size of $\mathcal{A}$.

Data complexity is relevant if $\mathcal{T}$ and the query are very small compared to $\mathcal{A}$. This is the case in most applications.

Non-Tractability of Query answering in $\mathcal{ALC}$ in Data Complexity

A graph G is a pair (W, E) consisting of a set W and a symmetric relation E on W.

G is 3-colorable if there exist subsets **blue**, **red**, and **green** of W such that

- the sets **blue**, **green**, and **red** are mutually disjoint;
- **blue** $\cup$ **red** $\cup$ **green** $= W$;
- if $(a, b) \in E$, then a and b do not have the same color.

3-colorability of graphs is an NP-complete problem.

3-Colorability as a Query Answering Problem

Assume $G = (W, E)$ is given. Construct the ABox $\mathcal{A}_G$ by taking a role name r and setting

- $r(a, b) \in \mathcal{A}$ for all $a, b \in W$ with $(a, b) \in E$.

Construct the TBox $\mathcal{ALC}$ TBox $\mathcal{T}_C$ by taking concept names **Blue**, **Green**, and **Red** and taking the inclusions:

- $\top \sqsubseteq$ **Blue** $\sqcup$ **Green** $\sqcup$ **Red**
- **Blue** $\sqcap \exists r.$**Blue** $\sqsubseteq$ **Clash**
- **Red** $\sqcap \exists r.$**Red** $\sqsubseteq$ **Clash**
- **Green** $\sqcap \exists r.$**Green** $\sqsubseteq$ **Clash**

Let $F = \exists x$ **Clash**(x). Then $(\mathcal{T}_C, \mathcal{A}_G) \models F$ if, and only if, G is not 3-colorable.

Restricting the Description Logic and the Query Language

- FOPL is too expressive as a query language for knowledge bases. The combined complexity of querying even DL-Lite or $\mathcal{EL}$ knowledge bases with FOPL queries is undecidable.

- For $\mathcal{ALC}$ knowledge bases and basic Boolean queries of the form $\exists x A(x)$, (A a concept name) query answering is still non-tractable. The best algorithms for query answering in this case are extensions of the $\mathcal{ALC}$ tableaux algorithms discussed above.

- We consider
  - knowledge bases in $\mathcal{EL}$, restricted Schema.org, and DL-Lite only;
  - queries in $\mathcal{EL}$ and conjunctive queries only.

Answering $\mathcal{EL}$-Queries in $\mathcal{EL}$ Knowledge Bases

$\mathcal{EL}$ Concept Queries

An **$\mathcal{EL}$ concept query** is a Boolean query of the form

$$C(a)$$

where C is an $\mathcal{EL}$-concept and a an individual name. We develop a method for answering $\mathcal{EL}$ concept queries in knowledge bases

$$(\mathcal{T}, \mathcal{A}),$$

where $\mathcal{T}$ is a $\mathcal{EL}$-TBox and $\mathcal{A}$ a simple ABox.

Note: Then we also have a method for computing

$$\textbf{certanswer}(C(x), (\mathcal{T}, \mathcal{A})) = \{a \mid (\mathcal{T}, \mathcal{A}) \models C(a)\}$$

Fundamental Idea: reduce knowledge base querying to relational database querying

To answer the question whether

$$(\mathcal{T}, \mathcal{A}) \models C(a)$$

we construct from $(\mathcal{T}, \mathcal{A})$ a finite interpretation $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ such that

$$(\mathcal{T}, \mathcal{A}) \models C(a) \quad \Leftrightarrow \quad \mathcal{I}_{\mathcal{T},\mathcal{A}} \models C(a).$$

Thus, we reduce ontology based reasoning to database querying. After this construction database technology can be used to process queries.

Note: Such a reduction works only for a very limited number of ontology and query languages!

From $(\mathcal{T}, \mathcal{A})$ to $\mathcal{I}_{\mathcal{T},\mathcal{A}}$

The algorithm constructing $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ is a rather simple extension of the algorithm deciding concept subsumption $A \sqsubseteq_{\mathcal{T}} B$ for $\mathcal{EL}$.

Firstly, we assume again that $\mathcal{T}$ is in normal form: it consists of inclusions of the form

- $A \sqsubseteq B$, where A and B are concept names;
- $A_1 \sqcap A_2 \sqsubseteq B$, where A_1, A_2, B are concept names;
- $A \sqsubseteq \exists r.B$, where A, B are concept names;
- $\exists r.A \sqsubseteq B$, where A, B are concept names.

General Description

The domain $\Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}}$ of $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ consists of

- all individual names a that occur in $\mathcal{A}$;
- objects d_A, for every concept name A in $\mathcal{T}$. (In the description of the subsumption algorithm d_A is denoted by A!)

It remains to compute

- $r^{\mathcal{I}_{\mathcal{T},\mathcal{A}}}$, for all role names r;
- $A^{\mathcal{I}_{\mathcal{T},\mathcal{A}}}$, for all concept names A.

This is done by computing functions S and R that are very similar to the functions introduced in the subsumption algorithm.

Algorithm Computing $\mathcal{I}_{\mathcal{T},\mathcal{A}}$

Given $\mathcal{T}$ in normal form and ABox $\mathcal{A}$, we compute functions S and R:

- S maps every $d \in \Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}}$ to a set $S(d)$ of concept names.

  We then set $d \in A^{\mathcal{I}_{\mathcal{T},\mathcal{A}}}$ if $A \in S(d)$;

- R maps every role name r to a set $R(r)$ of pairs (d_1, d_2) in $\Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}}$.

  We then set $(d_1, d_2) \in r^{\mathcal{I}_{\mathcal{T},\mathcal{A}}}$ if $(d_1, d_2) \in R(r)$.

We initialise S and R as follows:

- $S(a) = \{B \mid B(a) \in \mathcal{A}\}$;
- $S(d_A) = \{A\}$ (as in the subumption algorithm, where we had $d_A = A$!)
- $R(r) = \{(a, b) \mid r(a, b) \in \mathcal{A}\}$.

Algorithm

Apply the following four rules to S and R exhaustively:

(simpleR) If $A \in S(d)$ and $A \sqsubseteq B \in \mathcal{T}$ and $B \notin S(d)$, then

$$S(d) := S(d) \cup \{B\}.$$

(conjR) If $A_1, A_2 \in S(d)$ and $A_1 \sqcap A_2 \sqsubseteq B \in \mathcal{T}$ and $B \notin S(d)$, then

$$S(d) := S(d) \cup \{B\}.$$

(rightR) If $A \in S(d)$ and $A \sqsubseteq \exists r.B \in \mathcal{T}$ and $(d, d_B) \notin R(r)$, then

$$R(r) := R(r) \cup \{(d, d_B)\}.$$

(leftR) If $(d_1, d_2) \in R(r)$ and $B \in S(d_2)$ and $\exists r.B \sqsubseteq A \in \mathcal{T}$ and $A \notin S(d_1)$, then

$$S(d_1) := S(d_1) \cup \{A\}.$$

Example

Let $\mathcal{T}$ be defined as:

$$\begin{aligned}
\mathbf{BasketballClub} &\sqsubseteq \mathbf{Club} \\
\mathbf{BasketballPlayer} &\sqsubseteq \exists \mathbf{plays\_for.BasketballClub} \\
\exists \mathbf{plays\_for.Club} &\sqsubseteq \mathbf{Player} \\
\mathbf{Player} &\sqsubseteq \mathbf{Human\_being}
\end{aligned}$$

Let $\mathcal{A}$ be defined as:

$$\mathbf{Basketballplayer(bob)}, \quad \mathbf{Player(jim)}$$

$$\mathbf{Basketballclub(tigers)}, \quad \mathbf{Club(lions)}$$

$$\mathbf{plays\_for(rob, tigers)}, \quad \mathbf{plays\_for(bob, lions)}$$

Construction of $\mathcal{I}_{\mathcal{T},\mathcal{A}}$

The initial assignment (with obvious abbreviations) is given by

$$
\begin{aligned}
S(d_{\textsf{Basketclub}}) &= \{\textsf{Basketclub}\}\\
S(d_{\textsf{Basketplayer}}) &= \{\textsf{Basketplayer}\}\\
S(d_{\textsf{Club}}) &= \{\textsf{Club}\}\\
S(d_{\textsf{Player}}) &= \{\textsf{Player}\}\\
S(d_{\textsf{Human}}) &= \{\textsf{Human}\}\\
R(\textsf{plays\_for}) &= \{(\textsf{rob}, \textsf{tigers}), (\textsf{bob}, \textsf{lion})\}\\
S(\textsf{bob}) &= \{\textsf{Baskplayer}\}\\
S(\textsf{jim}) &= \{\textsf{Player}\}\\
S(\textsf{tigers}) &= \{\textsf{Baskclub}\}\\
S(\textsf{lions}) &= \{\textsf{Club}\}\\
S(\textsf{rob}) &= \emptyset
\end{aligned}
$$

Rule Applications

Now applications of (simpleR), (rightR), (leftR) are step-by-step as follows:

- Update S using (simpleR):
$$S(d_{\textbf{BaskClub}}) = \{\textbf{BaskClub}, \textbf{Club}\}.$$
- Update R using (rightR):
$$R(\textbf{plays\_for}) = \{(d_{\textbf{Baskplayer}}, d_{\textbf{BaskClub}})\}.$$
- Update S using (simpleR):
$$S(d_{\textbf{Player}}) = \{\textbf{Player}, \textbf{Human}\}.$$
- Update S using (leftR):
$$S(d_{\textbf{Baskplayer}}) = \{\textbf{Baskplayer}, \textbf{Player}\}.$$
- Update S using (simpleR):
$$S(d_{\textbf{Baskplayer}}) = \{\textbf{Baskplayer}, \textbf{Player}, \textbf{Human}\}.$$

Rule applications continued

- Update S using (simpleR):
$$S(\textbf{tigers}) = \{\textbf{BaskClub}, \textbf{Club}\}.$$
- Update S using (simpleR):
$$S(\textbf{jim}) = \{\textbf{Player}, \textbf{Human}\}.$$
- Update R using (rightR):
$$R(\textbf{plays\_for}) = \{(d_{\textbf{Baskplayer}}, d_{\textbf{BaskClub}}), (\textbf{bob}, d_{\textbf{BaskClub}})\}.$$
- Since $S(\textbf{bob})$ contains **Baskplayer**, we obtain using rules:
$$S(\textbf{bob}) = \{\textbf{Baskplayer}, \textbf{Player}, \textbf{Human}\}.$$
- Update S using (leftR):
$$S(\textbf{rob}) = \{\textbf{Player}\}.$$
- Update S using (leftR):
$$S(\textbf{rob}) = \{\textbf{Player}, \textbf{Human}\}.$$

The final assignment

$$\begin{aligned}
S(d_{\textsf{Baskclub}}) &= \{\textsf{Baskclub}, \textsf{Club}\} \\
S(d_{\textsf{Baskplayer}}) &= \{\textsf{Baskplayer}, \textsf{Player}, \textsf{Human}\} \\
S(d_{\textsf{Club}}) &= \{\textsf{Club}\} \\
S(d_{\textsf{Player}}) &= \{\textsf{Player}, \textsf{Human}\} \\
S(d_{\textsf{Human}}) &= \{\textsf{Human}\} \\
R(\textsf{plays\_for}) &= \{(d_{\textsf{Baskplayer}}, d_{\textsf{BaskClub}}), (\textsf{rob}, \textsf{tigers}), (\textsf{bob}, \textsf{lion}), (\textsf{bob}, d_{\textsf{BaskClub}})\} \\
S(\textsf{bob}) &= \{\textsf{Baskplayer}, \textsf{Player}, \textsf{Human}\} \\
S(\textsf{jim}) &= \{\textsf{Player}\} \\
S(\textsf{tigers}) &= \{\textsf{Baskclub}\} \\
S(\textsf{lions}) &= \{\textsf{Club}\} \\
S(\textsf{rob}) &= \{\textsf{Player}, \textsf{Human}\}
\end{aligned}$$

The interpretation $\mathcal{I}_{\mathcal{T},\mathcal{A}}$

- $\Delta^{\mathcal{I}}_{\mathcal{T},\mathcal{A}} = \{d_{\mathsf{Baskclub}}, d_{\mathsf{Baskplayer}}, d_{\mathsf{Club}}, d_{\mathsf{Player}}, d_{\mathsf{Human}}, \mathbf{bob}, \mathbf{jim}, \mathbf{tigers}, \mathbf{lions}, \mathbf{rob}\}$;
- $\mathbf{Baskclub}^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} = \{d_{\mathsf{Baskclub}}, \mathbf{tigers}\}$;
- $\mathbf{Club}^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} = \{d_{\mathsf{Club}}, d_{\mathsf{Baskclub}}, \mathbf{tigers}\}$;
- $\mathbf{Baskplayer}^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} = \{d_{\mathsf{Baskplayer}}, \mathbf{bob}\}$;
- $\mathbf{Player}^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} = \{d_{\mathsf{Player}}, d_{\mathsf{Baskplayer}}, \mathbf{bob}, \mathbf{jim}, \mathbf{rob}\}$;
- $\mathbf{Human}^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} = \{d_{\mathsf{Human}}, d_{\mathsf{Player}}, d_{\mathsf{Baskplayer}}, \mathbf{bob}, \mathbf{jim}, \mathbf{rob}\}$;
- $\mathbf{plays\_for}^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} = \{(d_{\mathsf{Baskplayer}}, d_{\mathsf{BaskClub}}), (\mathbf{rob}, \mathbf{tigers}), (\mathbf{bob}, \mathbf{lion}), (\mathbf{bob}, d_{\mathsf{BaskClub}})\}$.

Now

$$(\mathcal{T}, \mathcal{A}) \models C(a) \quad \Leftrightarrow \quad \mathcal{I}_{\mathcal{T},\mathcal{A}} \models C(a)$$

for all $\mathcal{EL}$ concepts C and a in $\mathcal{A}$. For example,

$$\mathcal{I}_{\mathcal{T},\mathcal{A}} \models \exists \mathbf{plays\_for}.\mathbf{Baskclub}(\mathbf{bob}), \quad \mathcal{I}_{\mathcal{T},\mathcal{A}} \models \mathbf{Human}(\mathbf{rob})$$

Another Example

We consider the knowledge base $\mathcal{S} = (\mathcal{O}, \mathcal{B})$ given by the ABox $\mathcal{B}$ consisting of

$$\textbf{Person}(\textbf{john}), \quad \textbf{Person}(\textbf{nick}), \quad \textbf{Person}(\textbf{toni})$$

$$\textbf{hasFather}(\textbf{john}, \textbf{nick}), \quad \textbf{hasFather}(\textbf{nick}, \textbf{toni})$$

and the TBox $\mathcal{O}$ given by

$$\mathcal{O} = \{\textbf{Person} \sqsubseteq \exists \textbf{has\_Father}.\textbf{Person}\}.$$

We construct $\mathcal{I}_{\mathcal{S}}$.

Constructing $\mathcal{I}_{\mathcal{S}}$

The initial assignment is given by

$$\begin{aligned} S(d_{\textbf{Person}}) &= \{\textbf{Person}\} \\ S(\textbf{john}) &= \{\textbf{Person}\} \\ S(\textbf{nick}) &= \{\textbf{Person}\} \\ S(\textbf{toni}) &= \{\textbf{Person}\} \\ R(\textbf{hasFather}) &= \{(\textbf{john}, \textbf{nick}), (\textbf{nick}, \textbf{toni})\} \end{aligned}$$

Four applications of the rule (rightR) add

$$\{(\textbf{john}, d_{\textbf{Person}}), (\textbf{nick}, d_{\textbf{Person}}), (\textbf{toni}, d_{\textbf{Person}}), (d_{\textbf{Person}}, d_{\textbf{Person}})\}$$

to the original $R(\textbf{hasFather})$. After that, no rule is applicable.

The interpretation $\mathcal{I}_\mathcal{S}$

We obtain the interpretation $\mathcal{I}_\mathcal{S}$ defined as

$$\begin{array}{rcl}
\Delta^{\mathcal{I}_\mathcal{S}} & = & \{d_{\textsf{Person}}, \textsf{john}, \textsf{nick}, \textsf{toni}\} \\
\textsf{Person}^{\mathcal{I}_\mathcal{S}} & = & \{d_{\textsf{Person}}, \textsf{john}, \textsf{nick}, \textsf{toni}\} \\
\textsf{hasFather}^{\mathcal{I}_\mathcal{S}} & = & \{(\textsf{john}, \textsf{nick}), (\textsf{nick}, \textsf{toni}), (\textsf{john}, d_{\textsf{Person}}), \\
 & & (\textsf{nick}, d_{\textsf{Person}}), (\textsf{toni}, d_{\textsf{Person}}), (d_{\textsf{Person}}, d_{\textsf{Person}})\}
\end{array}$$

We have

$$\mathcal{S} \models C(a) \quad \Leftrightarrow \quad \mathcal{I}_\mathcal{S} \models C(a)$$

for all $\mathcal{EL}$ concepts C and a from $\mathcal{B}$. For example

$$\mathcal{I}_\mathcal{S} \models \exists \textsf{hasFather}.\exists \textsf{hasFather}.\textsf{Person}(\textsf{toni})$$