Conservative Extensions in Expressive Ontology Languages

Thomas Schneider

AG Theorie der künstlichen Intelligenz (TdKI), Universität Bremen

Joint work with
Jean Christoph Jung, Carsten Lutz, Mauricio Martel, Frank Wolter
What is a conservative extension?

**Fundamental notion in mathematical logic** to relate theories

Useful in **computer science**:
- for formalising modularity in **software specification**
- for composing subgoals in **higher-order theorem proving**
- for formalising various notions in **ontology engineering**

Remarkably **positive results** for **description logics** and **modal logics**:

- ✓ turn out to be **decidable in many relevant cases**
- ✓ often have natural and insightful **model-theoretic characterisations**

**Natural question**: How far do these extend? To FO$^2$? **Guarded fragment**? **Existential rules** (aka Datalog$^\pm$)?
Conservative extensions for ontology design

W3C Web Ontology Language **OWL**

- ... is based on an expressive description logic (**SROIQ**)
- ... admits the design of **ontologies**, e.g., SNOMED CT, NCI Thesaurus, FMA (**100,000s of logical axioms**)
- Standard reasoning problems (e.g., SAT) well-understood
  reasoners: Racer, FaCT++, Pellet, HermiT, Konclude, ...

**Challenges** for designing/using large ontologies:

- Navigation, comprehension
- Efficient (incremental) reasoning
- Efficient reuse
- Versioning and more ...
A reuse scenario

Assume that ...

- you want to buy a subset of a medical ontology $\mathcal{O}$ from me that covers the subdomain of, say, diseases
- I offer two subsets $\mathcal{M}_1$ and $\mathcal{M}_2$

Q: which one do you choose?
A: the one that “knows more” about diseases!

Q: which is the best subset I can offer?
A: a subset $\mathcal{M} \subseteq \mathcal{O}$ that is

- ... indistinguishable from $\mathcal{O}$ w.r.t. all terms relevant for diseases
- ... as small as possible
Let $\varphi_1, \varphi_2$ be sentences and $\Sigma$ a signature (set of symbols).

- $\varphi_1 \Sigma$-entails $\varphi_2$, written $\varphi_1 \models_{\Sigma} \varphi_2$, if $\varphi_2 \models \psi$ and $\text{sig}(\psi) \subseteq \Sigma$ implies $\varphi_1 \models \psi$.

- $\varphi_1$ and $\varphi_2$ are $\Sigma$-inseparable, written $\varphi_1 \equiv_{\Sigma} \varphi_2$, if $\varphi_1 \models_{\Sigma} \varphi_2$ and $\varphi_2 \models_{\Sigma} \varphi_1$.

- $\varphi_1 \land \varphi_2$ is a conservative extension of $\varphi_1$ if $\varphi_1 \equiv_{\Sigma} \varphi_1 \land \varphi_2$ for $\Sigma = \text{sig}(\varphi_1)$.

In the reuse scenario, you should want to ...

- buy some $\mathcal{M} \subseteq \mathcal{O}$ with $\mathcal{M} \equiv_{\Sigma} \mathcal{O}$ ($\mathcal{M} \models_{\Sigma} \mathcal{O}$ suffices)
A closer look at $\Sigma$-entailment

$\Sigma$-entailment is the most general notion of the previous 3.

Two variants:

- **Deductive:** $\varphi_1$ $\Sigma$-entails $\varphi_2$
  if $\varphi_2 \models \psi$ and $\text{sig}(\psi) \subseteq \Sigma$ implies $\varphi_1 \models \psi$.

- **Model-theoretic:** $\varphi_1$ $\Sigma$-entails $\varphi_2$
  if every model of $\varphi_1$ can be extended to a model of $\varphi_2$ without changing the interpretation of the $\Sigma$-symbols.

Model-theoretic $\Sigma$-entailment is **highly undecidable** already for a very small FO fragment, that is, the description logic $\mathcal{EL}$:

$$\forall x \varphi(x) \quad \text{with} \quad \varphi(x) \text{ built from } \text{true}, \land, \exists y (Rxy \land \varphi(y))$$

and when $\Sigma = \text{sig}(\varphi_1)$.  

[Konev et al., AIJ 2013]
Examples

- $\varphi_1 \Sigma$-entails $\varphi_2$
  
  if $\varphi_2 \models \psi$ and $\text{sig}(\psi) \subseteq \Sigma$ implies $\varphi_1 \models \psi$.

Examples in the guarded fragment GF of FO (with equality):

- $\varphi ::= x = y \mid R\overline{x} \mid \neg \varphi \mid \varphi \land \varphi \mid \exists y (Rxy \land \varphi(\overline{xy}))$

- $\varphi_1 = \forall x \exists y Rxy \not\models \{R\} \quad \varphi_2 = \forall x ((\exists y Rxy \land Ay) \land (\exists y Rxy \land \neg Ay))$
  
  witnessed by $\psi = \exists x \exists y (Rxy \land x \neq y)$

- $\varphi'_1 = \forall x \exists y (Rxy \land x \neq y) \models \{R\} \quad \varphi_2$

Choice of separating logic essential: $\varphi'_1 \not\models \{R\} \varphi_2$ in FO
# Overview

1. An overview of $\Sigma$-entailment
2. $\Sigma$-entailment in FO fragments
3. Query entailment in expressive Horn description logics
4. Outlook
An overview of \( \Sigma \)-entailment

\( \Sigma \)-entailment in FO fragments

Query entailment in expressive Horn description logics

Outlook
\[ \Sigma \text{-entailment in modal and description logics} \]

**Basic description logic** \( \mathcal{ALC} \) \( \approx \) multi-modal logic K:

\[ \forall x \varphi(x) \text{ with } \varphi(x) \text{ built from } \text{true}, \neg, \land, \exists y (Rxy \land \varphi(y)) \]

**Theorem** (Lutz & Wolter 2011, “sloppy version”)

\( \varphi_1 \models_{\Sigma} \varphi_2 \iff \) every model of \( \varphi_1 \) can be extended to a model of \( \varphi_2 \), up to \( \Sigma \)-bisimulation.

\( \hat{=} \) model-theoretic \( \Sigma \)-entailment up to “what the logic can express”

Enables **decision procedure** using (amorphous) alternating tree automata

Problem is **2ExpTime-complete** (SAT in \( \mathcal{ALC} \): ExpTime); smallest witnesses are triple exponential in the worst case!

Similar characterisations for many other description and modal logics
\( \Sigma \)-entailment beyond \( ALC \)

✓ **Decidable and 2ExpTime-complete, too:** extensions of \( ALC \) with

- inverse roles (aka past modalities) \( \exists y (Ryx \land \varphi(y)) \)
- counting (aka graded modalities)

✗ **Undecidable:**

combination of the above two with nominals \( x = c \)

[Lutz et al., IJCAI 2007]
### Description logic $\mathcal{EL}$ ("half $\mathcal{ALC}$"):

$$\forall x \varphi(x) \text{ with } \varphi(x) \text{ built from } \text{true}, \land, \exists y (Rxy \land \varphi(y))$$

- analogous model-theoretic characterisation via simulations ("half-bisimulations")
- **ExpTime-complete** (SAT: in PTime)
- Restriction to acyclic terminologies: in **PTime**  
  (deductive and model-theoretic variant)

[Lutz et al., KR 2012; Konev et al., JAIR 2012 & AIJ 2013]
Further variants of $\Sigma$-entailment

$\Sigma$-query entailment / $\Sigma$-query inseparability:
separating formulas $\psi$ are queries

Relevant for ontology-based data access (OBDA) aka ontology-mediated querying (OMQ)

Various notions of $\Sigma$-query entailment obtained by . . .

- varying the query language: CQs, UCQs, PEQs, C2RPQs, . . .
- allowing whether $\varphi_1$, $\varphi_2$ contain data or not
  (if not, then relative to all possible instances)

Actively studied for basic and lightweight DLs (we’ll get back later)
Relationship with modularity of ontologies

Back to the reuse scenario:
Given ontology $\mathcal{O}$ and signature $\Sigma$, you want to buy a subset $\mathcal{M} \subseteq \mathcal{O}$ such that

(a) $\mathcal{M} \equiv_\Sigma \mathcal{O}$ and
(b) $\mathcal{M}$ small (possibly minimal with (a)): a module of $\mathcal{O}$ for $\Sigma$

Previous results: it is hard to decide whether a given $\mathcal{M} \subseteq \mathcal{O}$ is a module

There are tractable approximations guaranteeing (a) but not (b), e.g., locality-based modules [Cuenca Grau et al., JAIR 2008]
Relationship with uniform $\Sigma$-interpolants

Let $\varphi$ be a formula and $\Sigma \subseteq \text{sig}(\varphi)$.

Formula $\psi$ is a uniform $\Sigma$-interpolant of $\varphi$ if

1. $\text{sig}(\psi) \subseteq \Sigma$,
2. $\psi \equiv_\Sigma \varphi$

Relevant for forgetting:
eliminate non-$\Sigma$ predicates while preserving $\Sigma$-consequences

Applications:

- Ontology reuse
- Predicate hiding
- Ontology summary
1. An overview of $\Sigma$-entailment
2. $\Sigma$-entailment in FO fragments
3. Query entailment in expressive Horn description logics
4. Outlook
**Guarded Fragment and FO$^2$**

**GF$^k$, FO$^k$: k-variable fragment of GF or FO**

**Theorem**

- $\Sigma$-entailment, $\Sigma$-inseparability, and conservative extensions are **undecidable** in every logic that contains GF$^3$ or FO$^2$ (such as the guarded negation fragment GNF);
- **2ExpTime-complete** in GF$^2$.

(2) is based on a model-theoretic characterisation, but it is **much more complex** than, e.g., for ALC.
Undecidability

**GF^3**
- Reduction from the halting problem of 2-register machines
- Crucial: \( \varphi_2 \) uses ternary guard that is not in \( \Sigma \), thus breaks guardedness
- Little expressive power needed for separation: \( ALC \) suffices

**FO^2**
- Reduction from tiling problem
- Little expressive power needed in \( \varphi_1 \) and \( \varphi_2 \): \( ALC \) suffices
- Crucial: ability to use full \( FO^2 \) expressive power in witnessing formula
Characterisation

Recall that in $\mathcal{ALC}$:

**Theorem (Lutz & Wolter 2011, “sloppy version”)**

\[ \varphi_1 \models_\Sigma \varphi_2 \iff \text{every model of } \varphi_1 \text{ can be extended to a model of } \varphi_2, \text{ up to } \Sigma\text{-bisimulation.} \]

**Q:** Can’t we simply replace bisimulations with $\text{GF}^2$-bisimulations?

**A:** No!

\[ \varphi_1 = \exists x A x \land \forall x (A x \rightarrow \exists y (R x y \land A y)) \quad \Sigma = \{ R \} \]

“There exists a path $\bullet \xrightarrow{R} \bullet \xrightarrow{R} \bullet \rightarrow$”

\[ \varphi_2 = \varphi_1 \land \exists x (A x \land B x) \land \forall x (B x \rightarrow \exists y (R y x \land B y)) \]

“There exists a path $\rightarrow \bullet \xrightarrow{R} \bullet \xrightarrow{R} \bullet \rightarrow$”

Then $\varphi_1 \models_\Sigma \varphi_2$, but $\Sigma$-$\text{GF}^2$-bisimulations fail.
**Bounded bisimulations**

\(k\)-bounded bisimulations:

\[ a \sim^k_{\Sigma} b \iff a \text{ and } b \text{ are } \Sigma\text{-GF}^2\text{-bisimilar up to depth } k \]

---

**Theorem**

\( \varphi_1 \models_{\Sigma} \varphi_2 \iff \) for every model \( \mathcal{A} \) of \( \varphi_1 \) of finite outdegree and every \( k \geq 0 \), there is a model \( \mathcal{B} \) of \( \varphi_2 \) such that:

1. for every \( a \in A \), there is \( b \in B \) with \( a \sim_{\Sigma} b \) and
2. for every \( b \in B \), there is \( a \in A \) with \( a \sim_{\Sigma}^k b \).

---

But it is **not so easy** to deal with **bounded bisimulations** when using tree automata or related techniques!
“Marker-delimited” bisimulations

Substitute for $k$-bounded bisimulation between $A$ and $B$:

- Decorate $A$ with unary predicate $X$ such that
  - on every infinite path there are infinitely many $X$
  - the distance between two $X$ is $\geq k$
- Break off bisimulations at second $X$ seen (both back and forth)

Does not travel exactly $k$ steps, but we need it for every $k$ anyway

Problem: even in forest models, decoration does not exist when $k > 2$

Solution: we need bounded bisimulations only when travelling upwards, but not when travelling downwards.

Then the distance between markers only matters on upwards paths
Automata-friendly characterisation

Theorem

\( \varphi_1 \models_{\Sigma} \varphi_2 \) iff for every forest model \( \mathcal{A} \) of \( \varphi_1 \) of finite outdegree and every marking \( X \subseteq \mathcal{A} \), there is a model \( \mathcal{B} \) of \( \varphi_2 \) such that:

1. for every \( a \in \mathcal{A} \), there is \( b \in \mathcal{B} \) with \( a \sim_{\Sigma} b \) and
2. for every \( b \in \mathcal{B} \), there is \( a \in \mathcal{A} \) with \( a \sim_{\Sigma} X b \).

- downwards: unbounded
- upwards: stop after seeing second \( X \)
Decision procedure based on automata

2ATAs: 2-way alternating tree automata $\mathcal{A}$

- **Input:** non-empty node-labelled tree
  (unlimited depth, unbounded finite outdegree)

- **Transitions:**
  $\land, \lor$-formulas with atoms “send copy of $\mathcal{A}$ in state $q$ to ...”
  - current node
  - the predecessor node (if exists)
  - some or all successor node(s)

**Theorem**

Emptiness for 2ATAs can be solved in time exponential in $|Q|$.

**Proof** via reduction to 2ATAs on (exactly) $k$-ary trees;
their emptiness problem: ExpTime-complete [Vardi, ICALP 1998]
Decision procedure based on automata

Construct \( 2\text{ATA } \mathcal{A} \) such that \[ L(\mathcal{A}) = \emptyset \text{ iff } \varphi_1 \models_{\Sigma} \varphi_2 \]

with \(|Q|\) polynomial in \(|\varphi_1|\) and exponential in \(|\varphi_2|\)

**Input:** labelled tree representing a forest structure; labels contain information on unary/binary predicates and markers \( X \)

\( \mathcal{A} \) consists of three \( 2\text{ATAs} \) that check whether . . .

- the input tree represents a **model** \( \mathcal{A} \) of \( \varphi_1 \)
- the **marking** is correct
- there is \( \mathcal{B} \models \varphi_2 \) satisfying the **conditions from the theorem:**
  1. for every \( a \in A \), there is \( b \in B \) with \( a \sim_{\Sigma} b \) and
  2. for every \( b \in B \), there is \( a \in A \) with \( a \sim_{X \Sigma} b \).

(\( \mathcal{B} \) is constructed “locally”, memorising only guarded 1-/2-types)
Theorem

In $\text{GF}^2$, $\varphi_1 \models \Sigma \varphi_2$ can be decided in time single exponential in $|\varphi_1|$ and double exponential in $|\varphi_2|$. The problem is 2ExpTime-complete.

(2ExpTime lower bound via ATM reduction)

Corollary:

The same holds for $\Sigma$-inseparability, conservative extensions, and recognising uniform $\Sigma$-interpolants.
An overview of $\Sigma$-entailment

$\Sigma$-entailment in FO fragments

Query entailment in expressive Horn description logics

Outlook
Ontology-mediated querying

Idea:

- Database stores data – **ABox** $\mathcal{A}$, set of ground facts
- Ontology stores domain knowledge – **TBox** $\mathcal{T}$, think $\forall x \varphi(x)$
- Queries $q(\overline{x})$ are answered over **knowledge base (KB)** $(\mathcal{T}, \mathcal{A})$

Standard reasoning task  query answering:

given $(\mathcal{T}, \mathcal{A})$, $q(\overline{x})$, $\overline{a}$, does $(\mathcal{T}, \mathcal{A}) \models q(\overline{a})$ hold?

Query answering is well-understood ...

- for lightweight and “full Boolean” description logics
- and query languages CQs, UCQs, and sometimes PEQs, C2RPQs
As before, relevant for module extraction, versioning, etc.:

Let $T_1, T_2$ be TBoxes, $\Gamma, \Sigma$ signatures, and $Q$ a query language.

$T_1 (\Gamma, \Sigma)$-query entails $T_2$, written $T_1 \models^{Q, \Sigma} T_2$, if for all $\Gamma$-ABoxes $A$, $\Sigma$-queries $q(\overline{x}) \in Q$ and tuples $\overline{a}$:

$$(T_2, A) \models q(\overline{a}) \text{ implies } (T_1, A) \models q(\overline{a})$$

Inseparability and conservative extensions are again special cases.

**Variant $\Sigma$-query entailment between KBs:** $(T_1, A_1) \models^{Q, \Sigma} (T_2, A_2)$
An overview of $\Sigma$-query entailment

Analogous model-theoretic characterisations exist for the $\mathbf{KB}$ variant

- in $\mathbf{ALC}$: via homomorphisms between tree-shaped models
- in $\mathbf{EL}$ and Horn-$\mathbf{ALC}$: homomorphisms between canonical models

Useful for the $\mathbf{TBox}$ variant only if witness ABoxes can be restricted

Theorem (Botoeva et al., IJCAI’16)

$\Sigma$-CQ entailment is **undecidable** for $\mathbf{ALC}$ TBoxes and $\mathbf{2ExpTime}$-complete for $\mathbf{ALC}$ KBs.
An overview of $\Sigma$-query entailment

In $\mathcal{EL}$ and Horn-$\mathcal{ALC}$, the characterisations can be used to show

Theorem (Lutz & Wolter, JSC 2010; Botoeva et al., AIJ 2016)

$\Sigma$-CQ entailment is . . .

- ExpTime-complete for $\mathcal{EL}$-TBoxes and in PTime for $\mathcal{EL}$-KBs
- 2ExpTime-complete for Horn-$\mathcal{ALC}$ TBoxes and ExpTime-complete for Horn-$\mathcal{ALC}$ KBs

Further results for . . .

- DL-Lite dialects
- rooted (U)CQs
- data complexity of the KB variant
- $\mathcal{T}_1, \mathcal{T}_2$ of different expressivity (e.g., $\mathcal{ALC}$ and $\mathcal{EL}$)

Recent survey [Botoeva et al., RW 2016]
Beyond $\mathcal{EL}$ and Horn-$\mathcal{ALC}$

**Goal:** study $\Sigma$-query entailment for Horn-DLs with

1. **(I) inverse roles:** $\exists y (Ryx \land Ay)$
   - and more features:
     2. **(F) functionality:** $\forall x_1 x_2 y (Rx_1 y \land Rx_2 y \rightarrow x_1 = x_2)$
     3. **(H) role hierarchies:** $\forall xy (Rxy \rightarrow Sxy)$

and establish

- model-theoretic characterisations
- decidability/complexity
Model-theoretic characterisation

In $\mathcal{EL}$:

**Theorem (Lutz & Wolter 2010, “sloppy version”)**

$\mathcal{T}_1 \models_{\Gamma,\Sigma} \mathcal{T}_2$ iff for all $\Gamma$-ABoxes $A$ there is a $\Sigma$-homomorphism from the the canonical model of $(\mathcal{T}_2, A)$ to that of $(\mathcal{T}_1, A)$.

**Fails** in the presence of inverse roles for **very similar reasons**: infinite backward path not embeddable into infinite forward path!
**Bounded homomorphisms**

\(k\)-bounded homomorphisms:
\[ \mathcal{I}_1 \rightarrow^k_\Sigma \mathcal{I}_2 \text{ iff } \mathcal{I}_1 \text{ embeds homomorphically into } \mathcal{I}_2 \text{ up to depth } k \]

Shorthand for the canonical model of \((\mathcal{T}, \mathcal{A})\): \(\mathcal{I}_{\mathcal{T}, \mathcal{A}}\)

**Characterisation** for Horn-DLs *with inverse roles*:

**Theorem**
\[ \mathcal{T}_1 \models_{\Gamma, \Sigma} \mathcal{T}_2 \text{ iff for all } \Gamma\text{-ABoxes } \mathcal{A} \text{ and all } k \geq 0: \]
\[ \mathcal{I}_{\mathcal{T}_2, \mathcal{A}} \rightarrow^k_\Sigma \mathcal{I}_{\mathcal{T}_1, \mathcal{A}} \]

Again, bounded homomorphisms are **difficult** for tree automata!
Automata-friendly characterisation

**Lemma (sloppy):** it suffices to consider **tree-shaped** ABoxes and CQs.

**Theorem**

\[ T_1 \models_{CQ}^{\Gamma,\Sigma} T_2 \text{ iff for all tree-shaped } \Gamma\text{-ABoxes } \mathcal{A}: \]

(1) “the \( \Sigma \)-connected part of \( \mathcal{I}_{T_2,\mathcal{A}} \)” \( \rightarrow_{\Sigma} \mathcal{I}_{T_1,\mathcal{A}} \) and

(2) For every \( \Sigma \)-subtree \( \mathcal{I} \) in \( \mathcal{I}_{T_2,\mathcal{A}} \), one of the following holds:

(a) \( \mathcal{I} \rightarrow_{\Sigma} \mathcal{I}_{T_1,\mathcal{A}} \)

(b) there is a \( \Sigma \)-subtree of \( \mathcal{I}_2 \) rooted in the ABox part such that \( \forall k \geq 0 \), we have \( \mathcal{I} \rightarrow_{\Sigma}^k \mathcal{I}' \)

- (1) and (2a) use **unbounded** homomorphisms
  \( \sim \) decide via **2ATAs with counting**

- (2b) uses **bounded** homomorphisms
  \( \sim \) decide via **mosaic procedure**; “hard-code” into automaton
Theorem

In $\mathcal{ELI} \ldots$ Horn-$\mathcal{ALCHIF}$, $T_1 \models^{CQ}_{\Gamma, \Sigma} T_2$ can be decided in time single exponential in $|T_1|$ and double exponential in $|T_2|$. The problem is 2ExpTime-complete.

Corollary:
The same holds for $\Sigma$-inseparability and conservative extensions.
Next ...

1. An overview of $\Sigma$-entailment
2. $\Sigma$-entailment in FO fragments
3. Query entailment in expressive Horn description logics
4. Outlook
Deductive $\Sigma$-entailment in FO fragments:

What happens if we add

- guarded counting quantifiers,
- transitive relations or equivalence relations,
- fixed points?

Finite-model version of conservative extensions?

$\Sigma$-query entailment:

Extension to Datalog$^\pm$ languages (aka existential rules), in particular to frontier-guarded TGDs?
Thank you.
An example

$T_1$ (in FO notation): $\forall x \left( \text{PhdStud}(x) \rightarrow \exists y \left( \text{adv}(y, x) \land \text{Prof}(y) \right) \right)$

“every PhD student is advised by some prof”

$T_2 = T_1 \land \forall x_1 x_2 y \left( \text{adv}(x_1, y) \land \text{adv}(x_2, y) \rightarrow x_1 = x_2 \right)$

“everyone has $\leq 1$ advisor”

$\Gamma = \{ \text{PhdStud}, \text{adv} \}, \quad \Sigma = \{ \text{Prof} \}$

$T_1 \not\models_{\Gamma, \Sigma} T_2$:

- $\Gamma$-ABox $\mathcal{A} = \{ \text{PhdStud}(\text{john}), \text{adv}(\text{mary}, \text{john}) \}$
- $\Sigma$-CQ $q(a) = \text{Prof}(\text{mary})$
- $(T_2, \mathcal{A}) \models \text{Prof}(\text{mary})$ but $(T_1, \mathcal{A}) \not\models \text{Prof}(\text{mary})$
Uniform interpolation

Remember: $\psi$ is a uniform $\Sigma$-interpolant of $\varphi$ if

1. $\text{sig}(\psi) \subseteq \Sigma$,
2. $\psi \equiv_{\Sigma} \varphi$

Uniform interpolant recognition problem (UIRP):
Given $\varphi, \psi, \Sigma$, is $\psi$ a uniform $\Sigma$-interpolant of $\varphi$?

Easy reductions: to deciding $\Sigma$-entailment, “backwards” from deciding conservative extensions

Corollary

UIRP is $2\text{ExpTime}$-complete in $\text{GF}^2$ and undecidable in all extensions of $\text{FO}^2$ or $\text{GF}^3$ with Craig interpolation, e.g., GNF.
There is no decidable extension of $\text{FO}^2$ and of $\text{GF}^3$ that has effective uniform interpolation.