Perfect Rewriting for Ontology Based Query Answering over Spatial Databases

⌘ Ralf Moeller
Hamburg University of Technology

Joint work with
Oezguer Oezcep, Rolf Gruetter and Aleksandar Gudov

Agenda

1. Ontology Based Query Answering (OBQA)
2. GDL-Lite-8 = DL-Lite with extensions to cope with certain aspects of spatial data (RCC8)
3. Query rewriting algorithm for GDL-Lite-8
4. Insights from an implementation and its evaluation

Query answering w.r.t. ontologies

TBox:

$\text{MALE} \sqsubseteq \text{PERSON}$
$\text{MALE} \sqsubseteq \neg\text{FEMALE}$
$\exists \text{hasFather}^- \sqsubseteq \text{MALE}$
$\exists \text{hasMother}^- \sqsubseteq \text{FEMALE}$
$\text{FEMALE} \sqsubseteq \text{PERSON}$
$\text{PERSON} \sqsubseteq \exists \text{hasFather}$
$\text{PERSON} \sqsubseteq \exists \text{hasMother}$

ABox:

MALE(Bob)
MALE(Paul)
FEMALE(Ann)
hasFather(Ann, Paul)
hasMother(Paul, Mary)

input query:

$q(x) \leftarrow \text{PERSON}(x)$

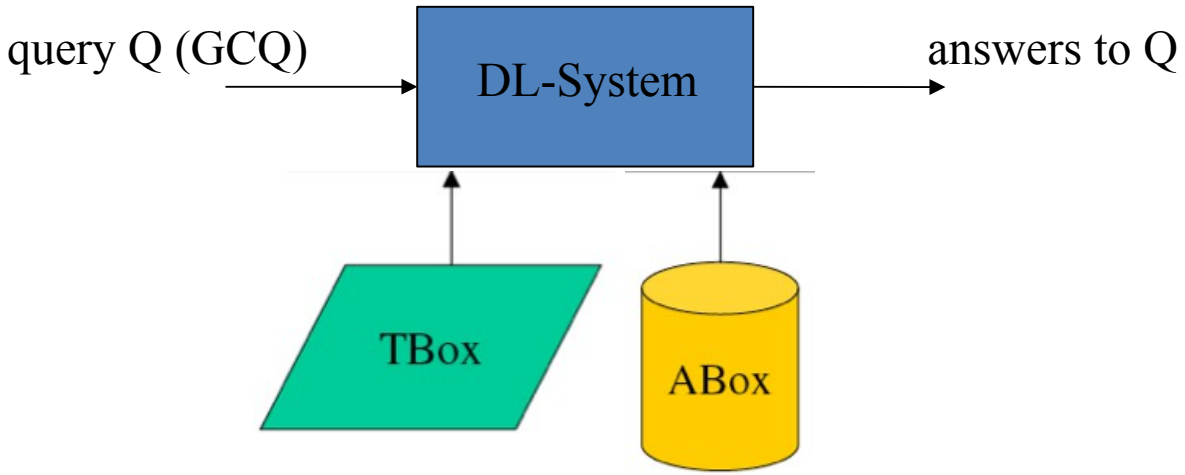
answers to query:

{ Bob, Paul, Ann, Mary }

Example adapted from a presentation by Riccardo Rosati

To Some Extent Solved by DL Systems…

- ALNHf: CLASSIC (90-96)
- SRIQ(D)$^-$: RacerPro (99-today)
- SROIQ/OWL 2: Pellet (04-today)



Problems and Solutions

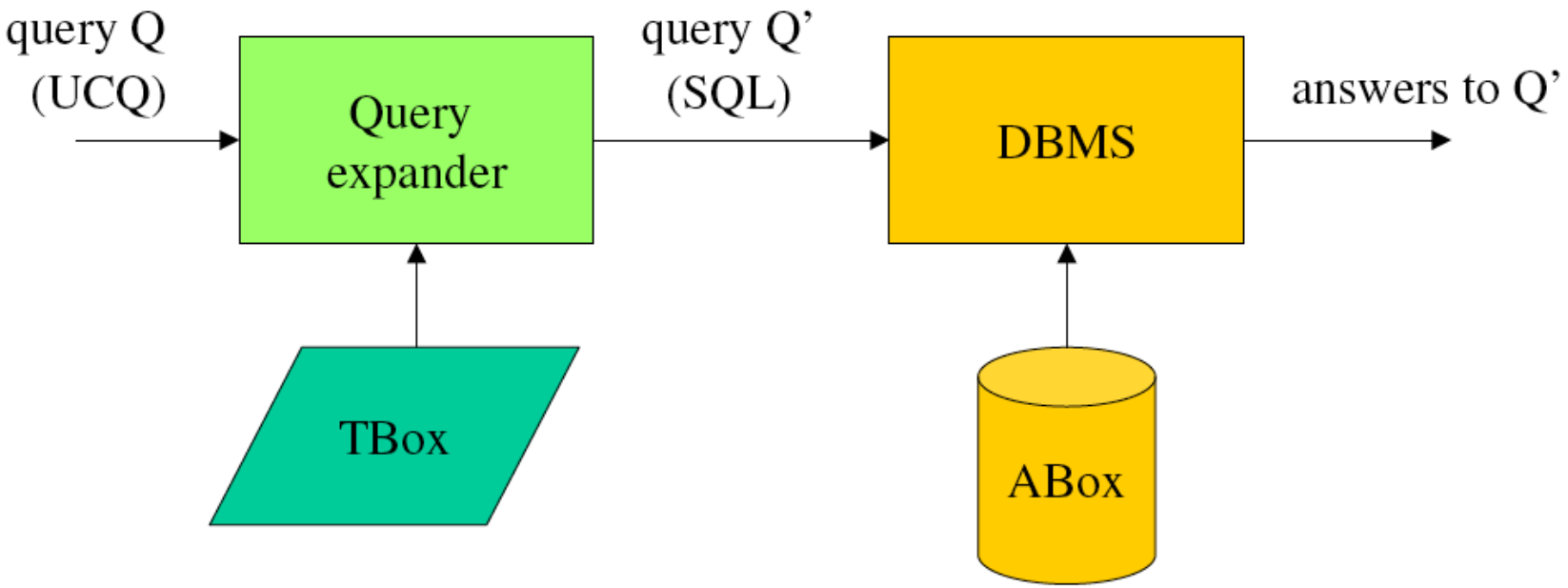
- GIS deal with spatial data

➢ Extensions to DLs such as DL-RCC8 (e.g., PelletSpatial, RacerPro, ...)

- GIS deal with big data
- (→ Partitioning approaches)

➢ Query rewriting (e.g., ontopQuest, Stardog)

Query Rewriting



Taken from a presentation by Riccardo Rosati

DL-Lite

Definition ($\text{DL-Lite}^{\sqcap}_{\mathcal{F},\mathcal{R}}$)

$P \in RN$ (role symb.); $A \in CN$ (concept symb.); $a, b \in Const$ (constants).

$$R \longrightarrow P \mid P^{-} \quad B \longrightarrow A \mid \exists R \quad C_l \longrightarrow B \mid C_l \sqcap B \quad C_r \longrightarrow B \mid \neg B$$

TBox*): $C_l \sqsubseteq C_r, (\text{funct } R), R_1 \sqsubseteq R_2$

ABox: $A(a), R(a, b)$

*) Restriction: If R occurs in a functionality axiom, then R and R^{-} do not occur on the right-hand side of a role inclusion axiom $R_1 \sqsubseteq R_2$.

Rewriting: Example

TBox:

$\text{MALE} \sqsubseteq \text{PERSON}$
$\text{MALE} \sqsubseteq \neg\text{FEMALE}$
$\exists\text{hasFather}^- \sqsubseteq \text{MALE}$
$\exists\text{hasMother}^- \sqsubseteq \text{FEMALE}$
$\text{FEMALE} \sqsubseteq \text{PERSON}$
$\text{PERSON} \sqsubseteq \exists\text{hasFather}$
$\text{PERSON} \sqsubseteq \exists\text{hasMother}$

input query:

$q(x) \leftarrow \text{PERSON}(x)$

rewritten query:

$q'(x) \leftarrow \text{PERSON}(x) \vee$
$\text{FEMALE}(x) \vee$
$\text{MALE}(x) \vee$
$\text{hasFather}(y,x) \vee$
$\text{hasMother}(y,x)$

Taken from a presentation by Riccardo Rosati

Example (cntd.)

rewritten query:

$q'(x) \leftarrow \text{PERSON}(x) \vee$
$\text{FEMALE}(x) \vee$
$\text{MALE}(x) \vee$
$\text{hasFather}(y,x) \vee$
$\text{hasMother}(y,x)$

ABox:

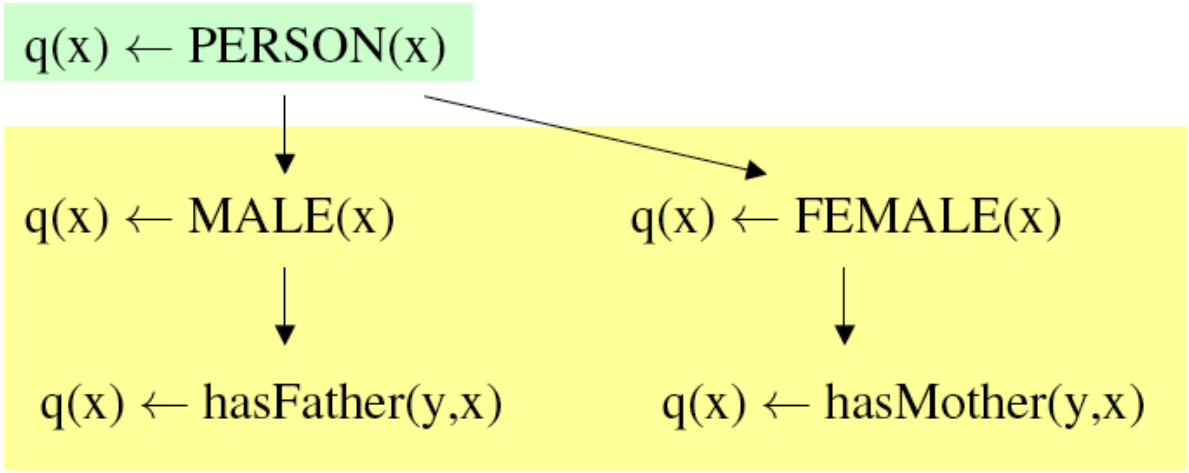
MALE(Bob)
MALE(Paul)
FEMALE(Ann)
hasFather(Ann, Paul)
hasMother(Paul, Mary)

answers to query:

{ Bob, Paul, Ann, Mary }

Taken from a presentation by Riccardo Rosati

Perfect Reformulation Algorithm for DL-Lite (Backward Chaining)



how to avoid the infinite chase of the ABox?

CHASE of the query:

- inclusions are applied "from right to left"
- this chase always terminates
- this chase is computed independently of the ABox

Taken from a presentation by Riccardo Rosati

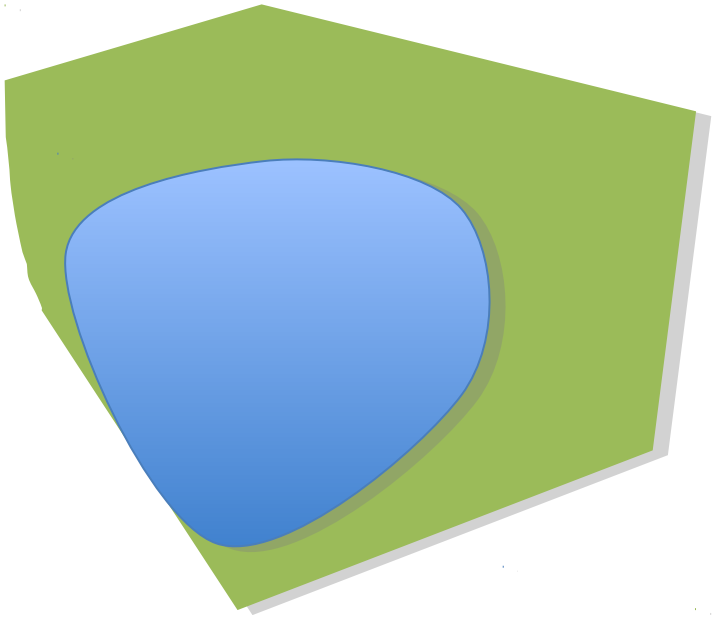
Further Problems and Solutions

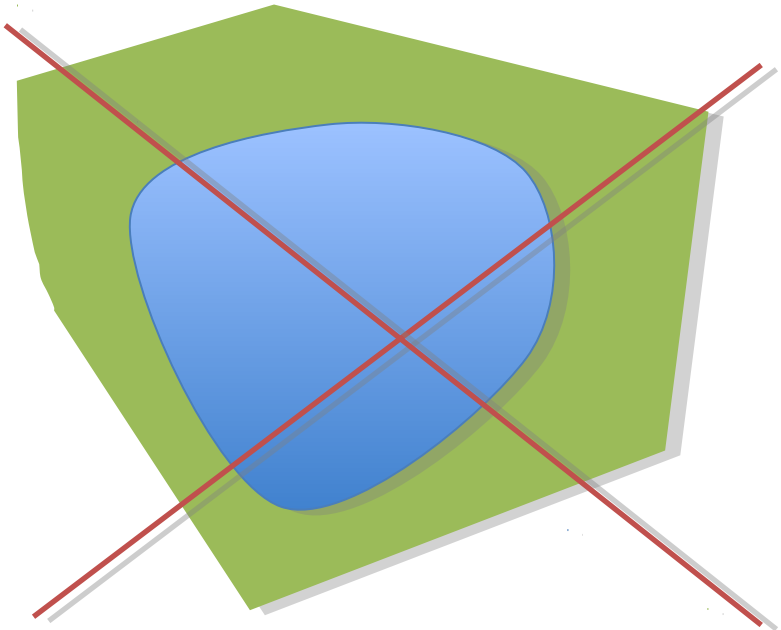
- Existing DBs store values in n-ary tables (→ Virtual Abox and mapping of query atoms to SQL in global-as-view style)
- Query rewriting might cause exponential blowup of the original query (disjunctive normal form), which might lead SQL optimizer into combinatorial explosion (→ Optimizations, e.g., in Optique, EU-FP7)
- Rewriting for spatial reasoning

➢ [Oezcep&M ISWC-2012]

Example (1)

- Park with lake (Park+Lake)
- Lake should be reachable from the outside (e.g., for easy access from roads)

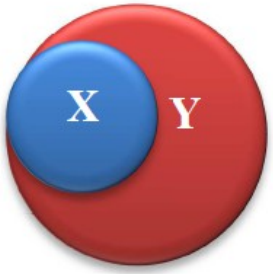




$\mathcal{B}_{RCC8}$ [KR-92]

DC(x,y)
disjointness

EC(x,y)
externally connected

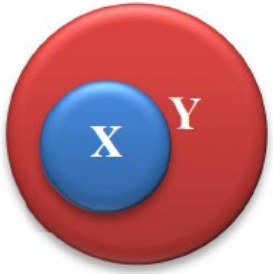


TPP(x,y)
tangential proper part

TPPi(x,y)
tangential proper part inverse

PO(x,y)
partial overlap

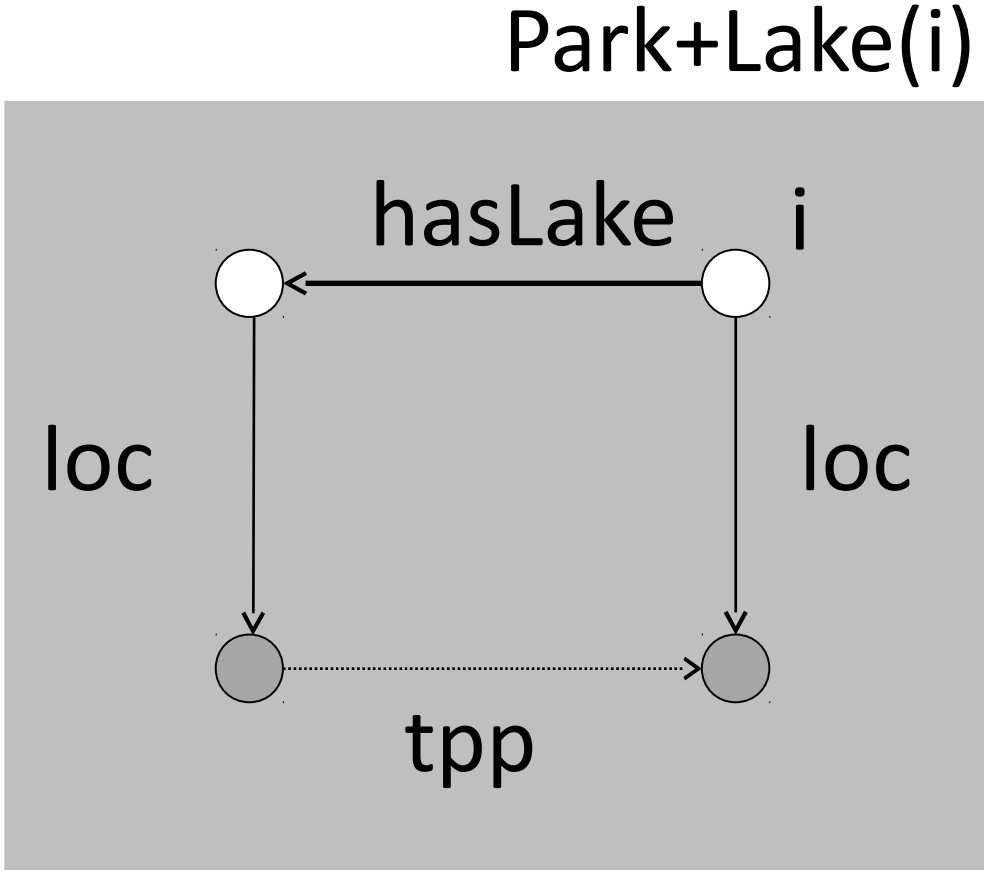
EQ(x,y)
equal



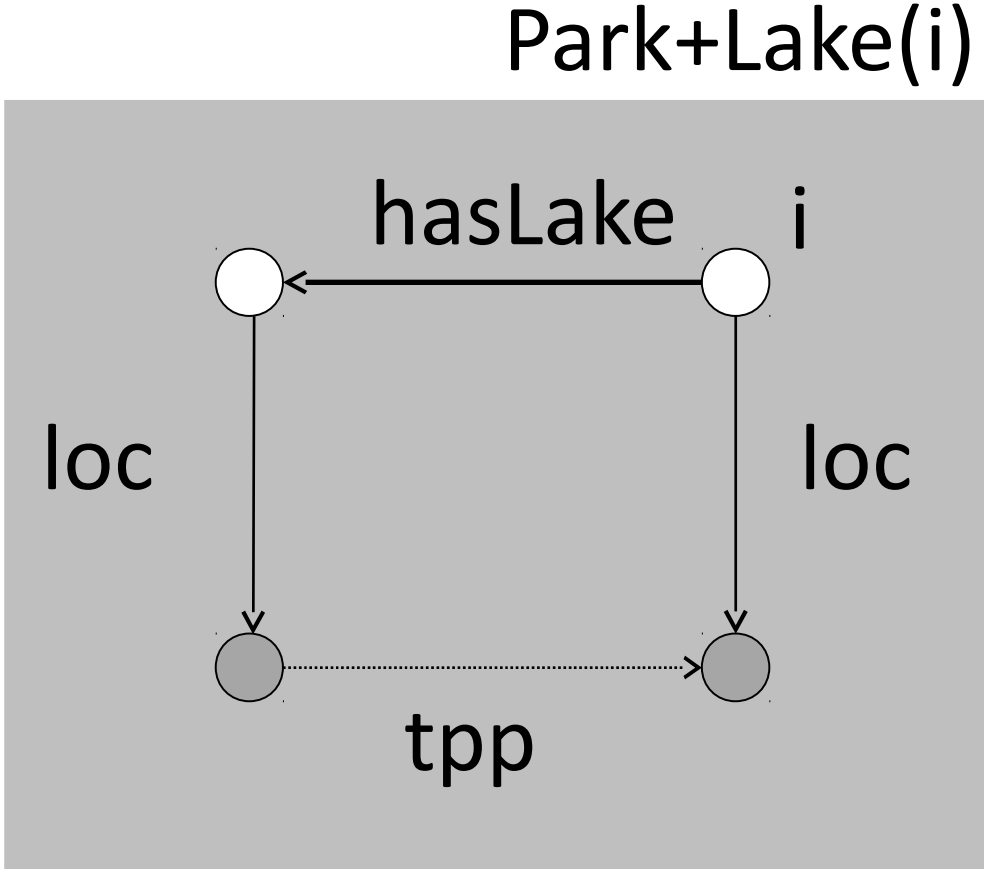
NTPP (x,y)
non-tangential proper part

NTPPi (x,y)
non-tangential proper part inverse

Relational Representation (1)



Relational Representation (1)

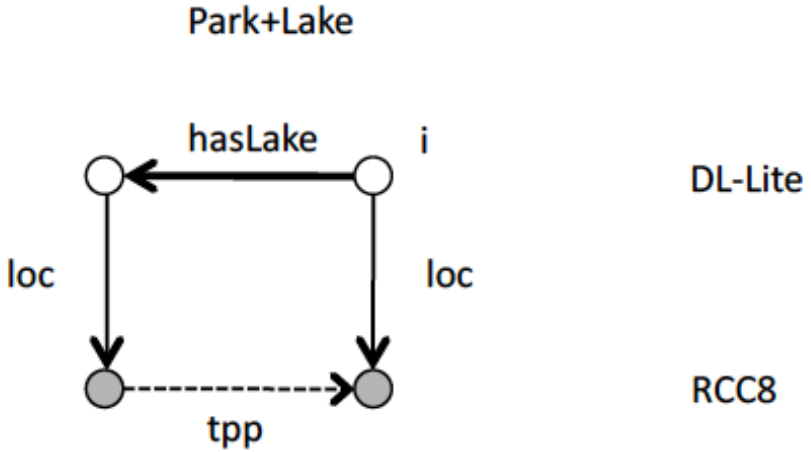


DL

RCC8

Representation Language: GDL-Lite-8

- Restricted combination with paths from thematic to spatial component (Lutz, Miličić 2007)
- $\exists hasLake \circ loc, loc.tpp(i)$



Definition (GDL-Lite-8)

$R \longrightarrow P \mid P^-$ $\quad U \longrightarrow loc \mid R \circ loc$ $\quad B \longrightarrow A \mid \exists R \mid \exists loc$

$C_l \longrightarrow B \mid C_l \sqcap B$ $\quad C_r \longrightarrow B \mid \neg B \mid \exists U_1, U_2.r$

TBox*): $C_l \sqsubseteq C_r, (\text{funct } loc), (\text{funct } R), R_1 \sqsubseteq R_2$

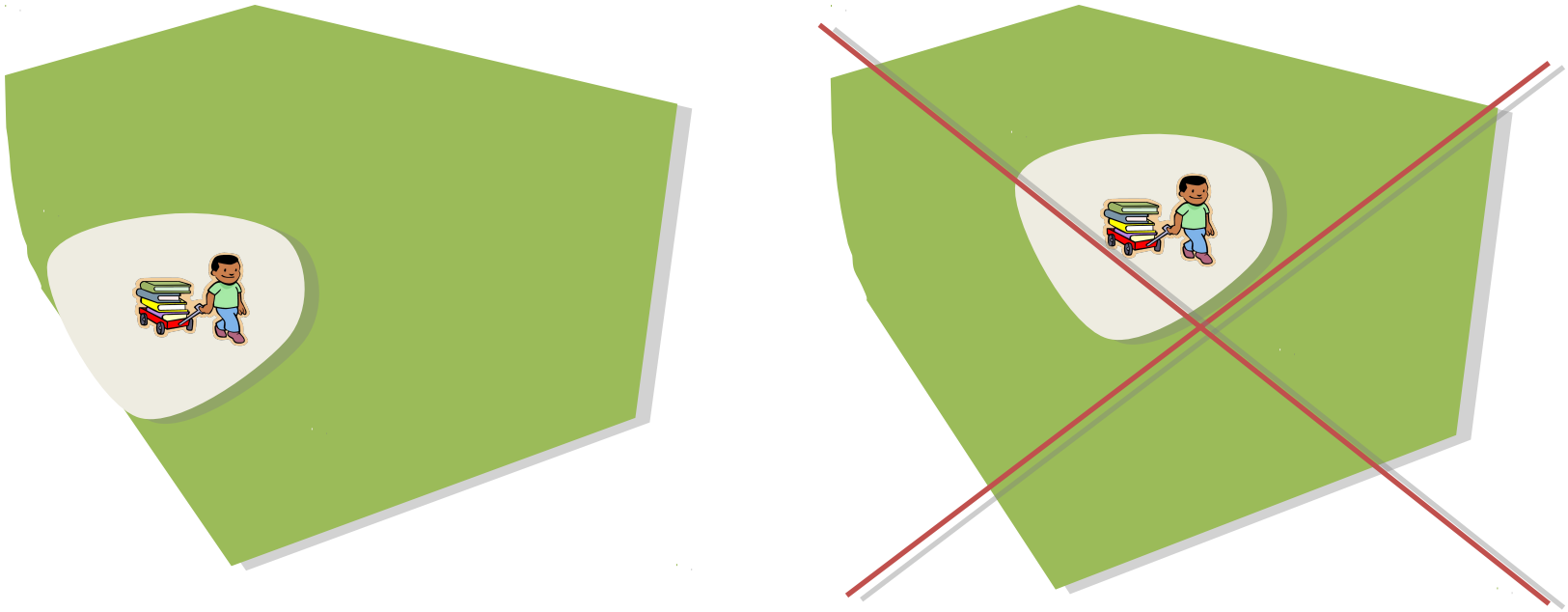
ABox: $A(a), R(a,b), loc(a,a^*), r(a^*,b^*)$ for $r \in Rel_{RCC8}$

*) If $(\text{funct } R) \in \mathcal{T}$, then R and R^- do not occur on the right-hand side of a role inclusion axiom or in a concept of the form $\exists U_1, U_2.r$.

[Oezcep and Moeller ISWC-2012]

Example (2)

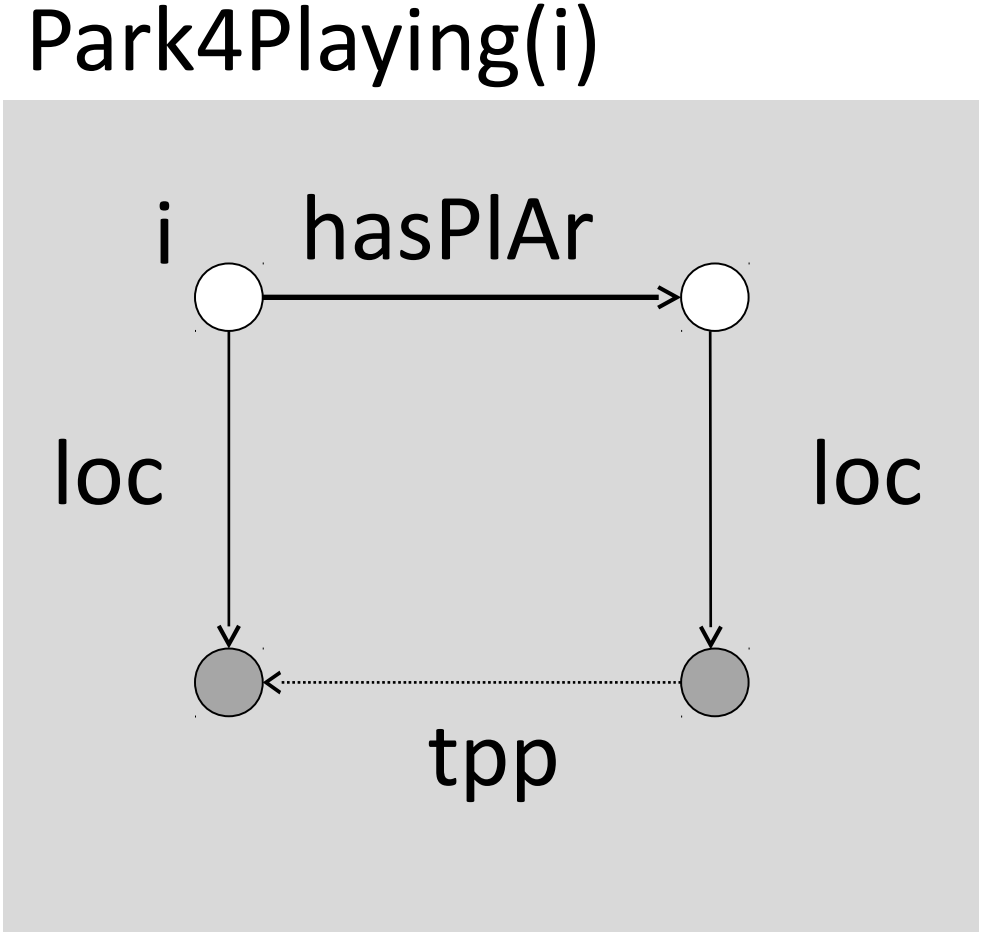
- Park with playing area (Park4Playing)



Relational Representation (2)

DL

RCC8



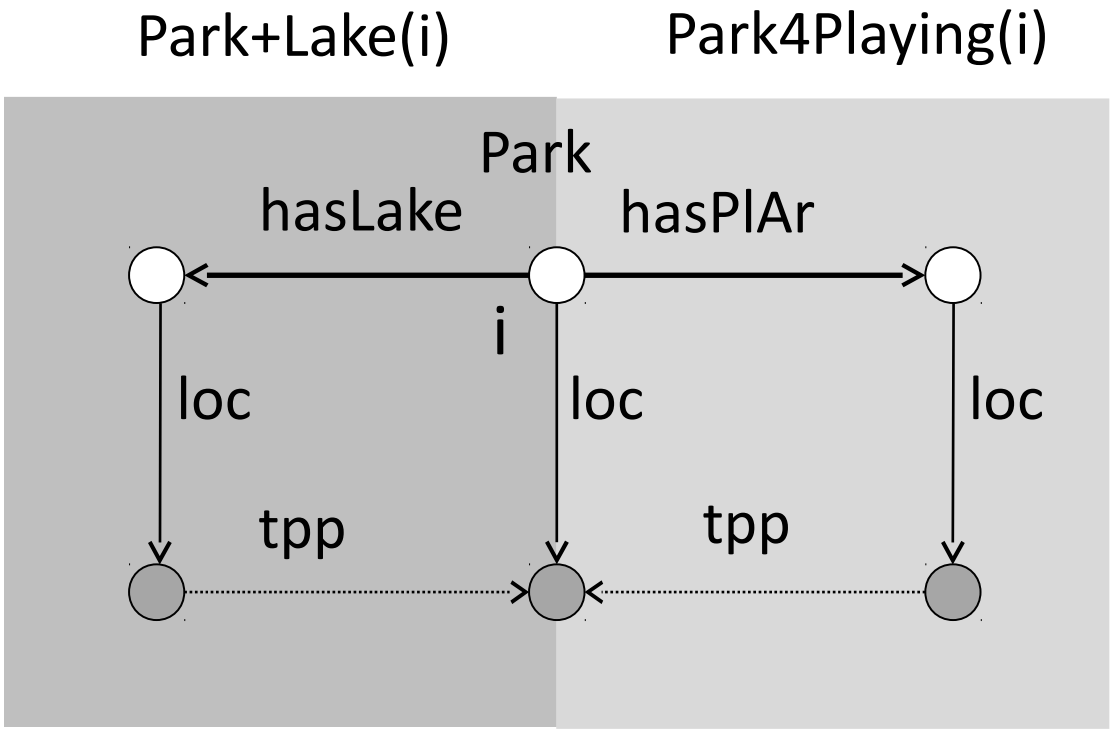
Combined Representation (KB)

- TBox:

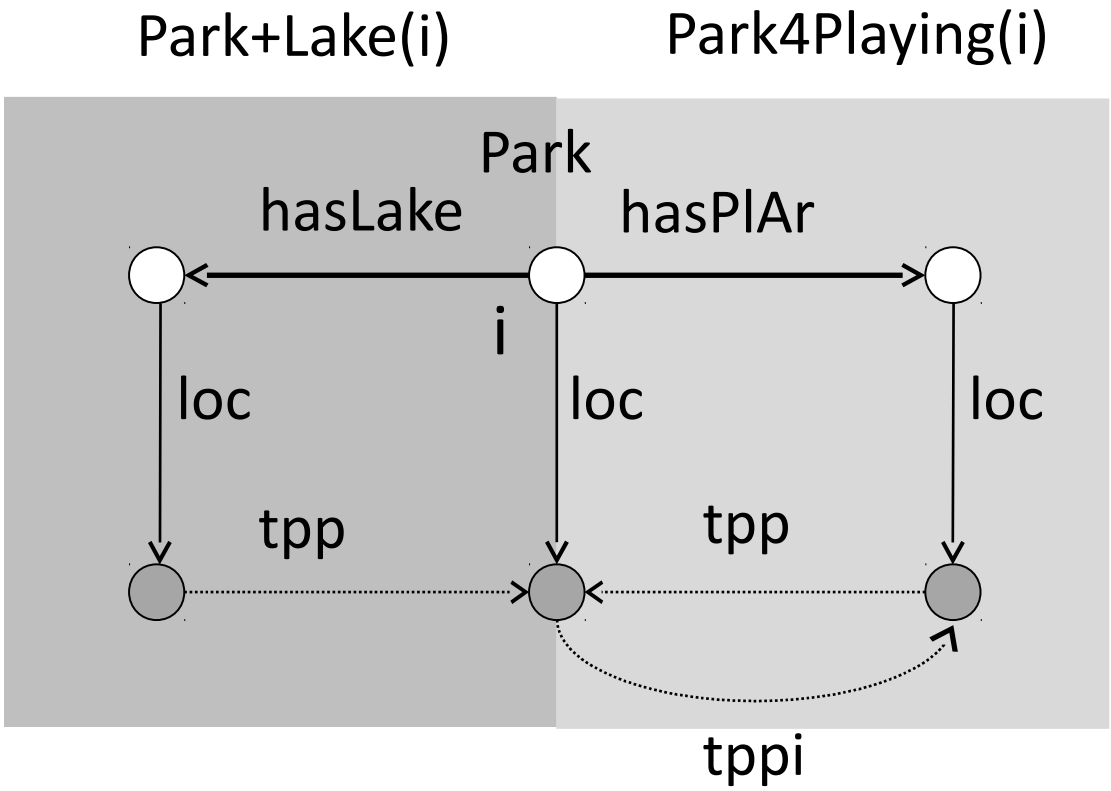
$$\begin{aligned} \mathit{Park{+}Lake} &\sqsubseteq \mathit{Park} \\ \mathit{Park4Playing} &\sqsubseteq \mathit{Park} \\ \mathit{Park{+}Lake} &\sqsubseteq \exists \mathit{hasLake} \circ \mathit{loc}, \mathit{loc}.\mathrm{tpp} \\ \mathit{Park4Playing} &\sqsubseteq \exists \mathit{hasPlAr} \circ \mathit{loc}, \mathit{loc}.\mathrm{tpp} \end{aligned}$$

- ABox contains $\mathit{Park{+}Lake}(i)$, $\mathit{Park4Playing}(i)$

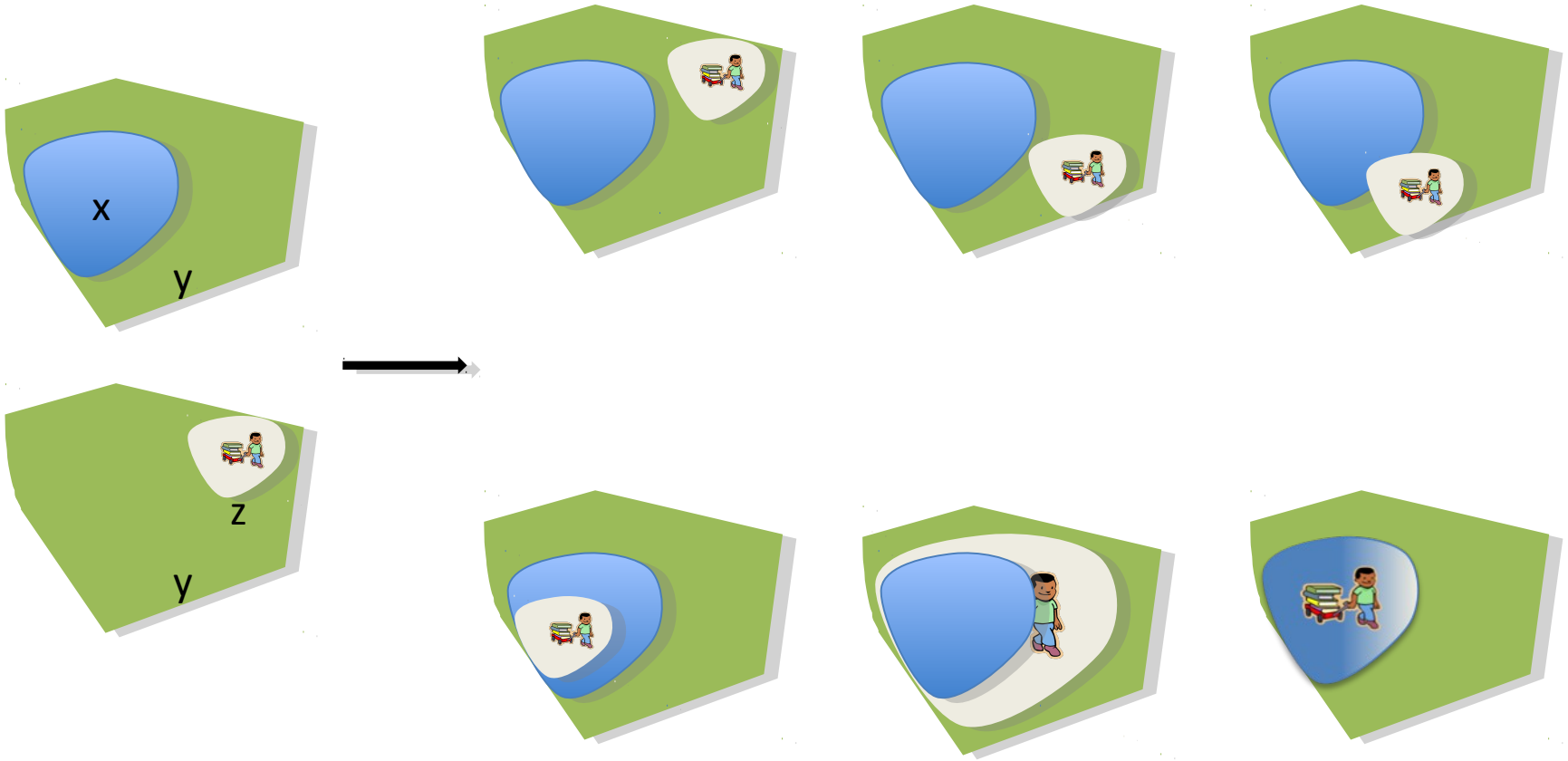
Combined Representation (Models)



Combined Representation (Models)



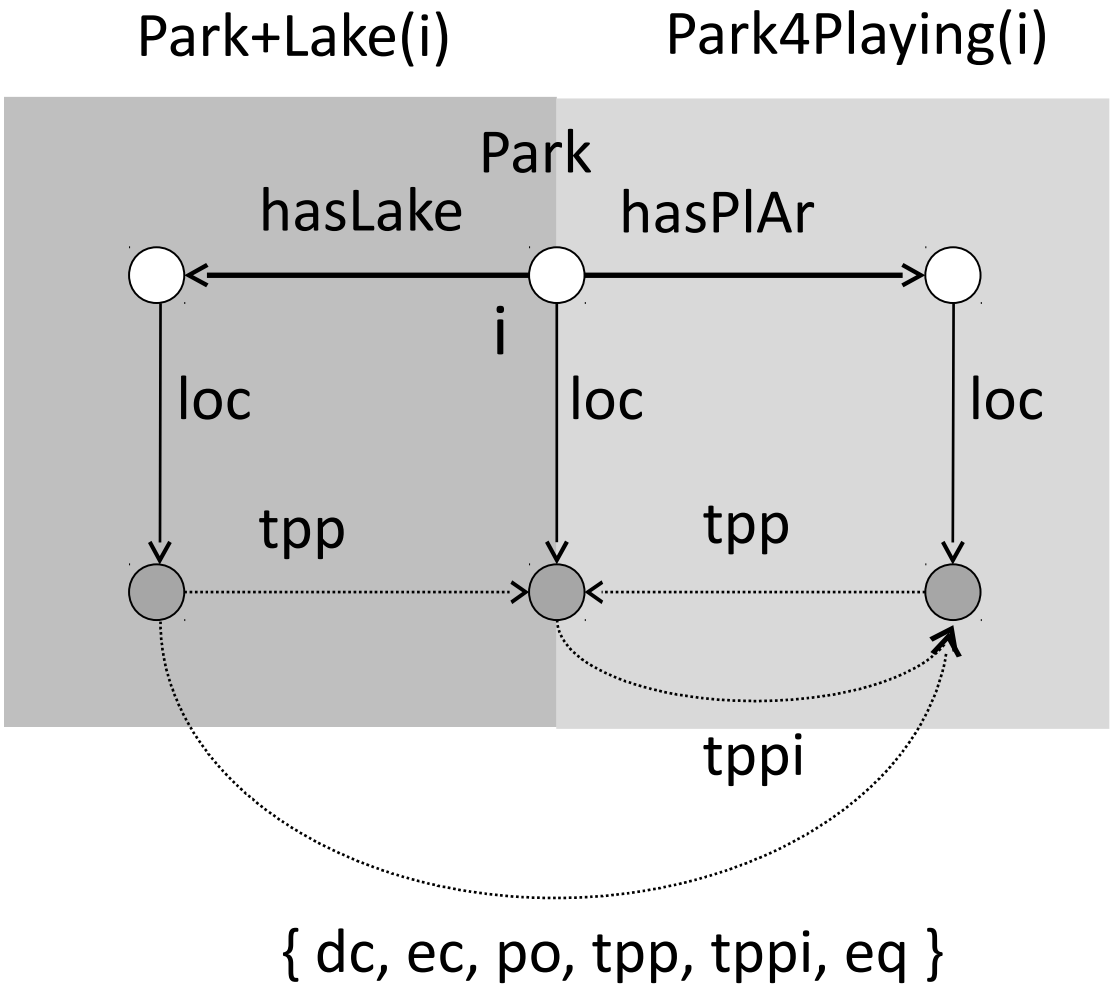
tpp(x,y) and tppi(y,z) Resulting Models



Composition Table

∘	DC	EC	PO	TPP	NTPP	TPPi	NTPPi	EQ
DC	*	DC EC PO TPP NTPP	DC EC PO TPP NTPP	DC EC PO TPP NTPP	DC EC PO TP NTPP	DC	DC	DC
EC	DC EC PO TPPi NTPPi	DC EC PO TPP TPPi EQ	DC EC PO TPP NTPP	EC PO TPP NTPP	PO TPP NTPP	DC EC	DC	EC
PO	DC EC PO TPPi NTPPi	DC EC PO TPPi NTPPi	*	PO TPP NTPP	PO TPP NTPP	DC EC PO TPPi NTPPi	DC EC PO TPPi NTTPi	PO
TPP	DC	DC EC	DC EC PO TPP NTPP	TPP NTPP	NTPP	DC EC PO TPP TPPi EQ	DC EC PO TPPi NTPPi	TPP
NTPP	DC	DC	DC EC PO TPP NTPP	NTPP	NTPP	DC EC PO TPP NTPP	*	NTPP
TPPi	DC EC PO TPPi NTPPi	EC PO TPPi NTPPi	PO TPPi NTPPi	PO EQ TPP TPPi	PO TPP NTPP	TPPi NTPPi	NTPPi	TPPi
NTPPi	DC EC PO TPPi NTPPi	PO TPPi NTPPI	PO TPPi NTPPi	PO TPPi NTPPi	PO TPPi TPP NTPP NTPPi EQ	NTPPi	NTPPi	NTPPi
EQ	DC	EC	PO	TPP	NTPP	TPPi	NTPPi	EQ

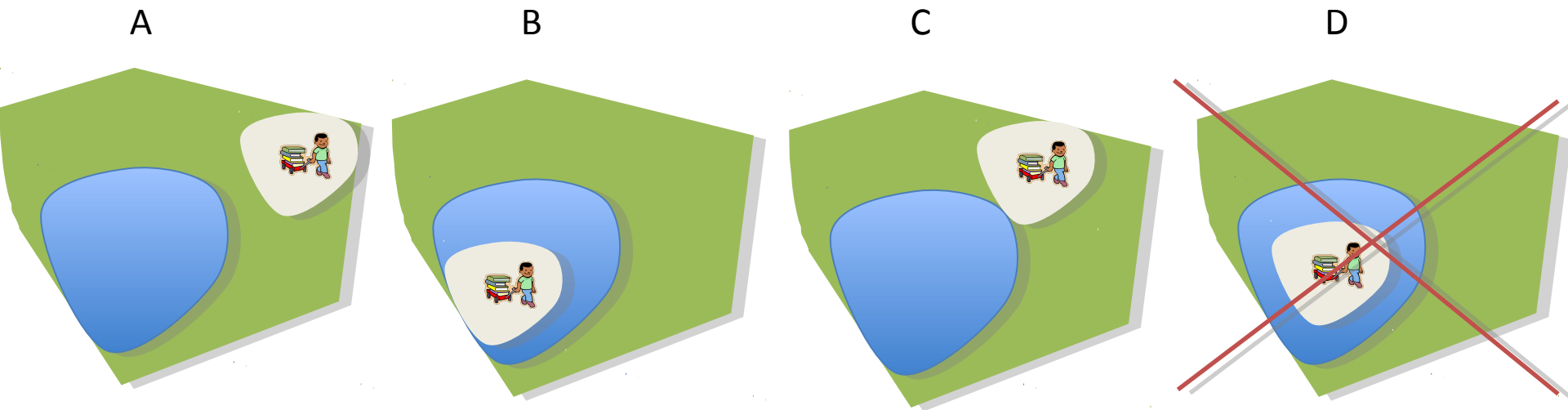
Combined Representation (Models)



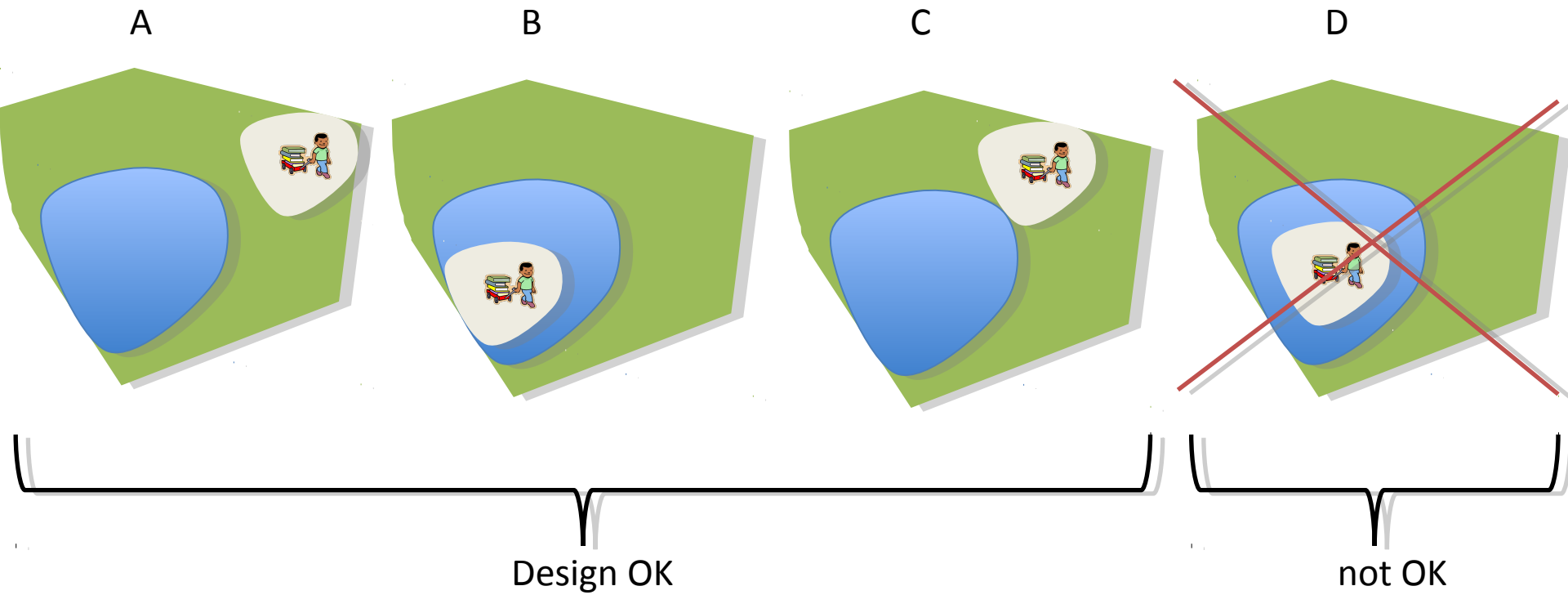
Application Scenario

- Large repository of assertions modeling "spatial designs" in an engineering company
- Safety conditions for design support
  - Formalization as queries
  - Non-results are counterexamples

Querying for Safe Designs



Querying for Safe Designs



Query Language GCQ^+

Example query

Q(x) = Find all parks x with lakes such that x contains a playing area that is not contained as island in the lake.

- Query formalizable in a special query language GCQ^+
  - allows atoms of form $C(x)$ where C is a GDL-Lite-8 concept without $\neg$
  - active domain semantics for atoms of the form $R(x, y)$, $r(x^*, y^*)$ etc.

Query formally

$$Q(x) = Park(x) \wedge \exists hasLake \circ loc, hasPlAr \circ loc.\mathcal{B}_{RCC8} \setminus \{\text{ntpp}\}(x)$$

The Complete Example

- TBox:

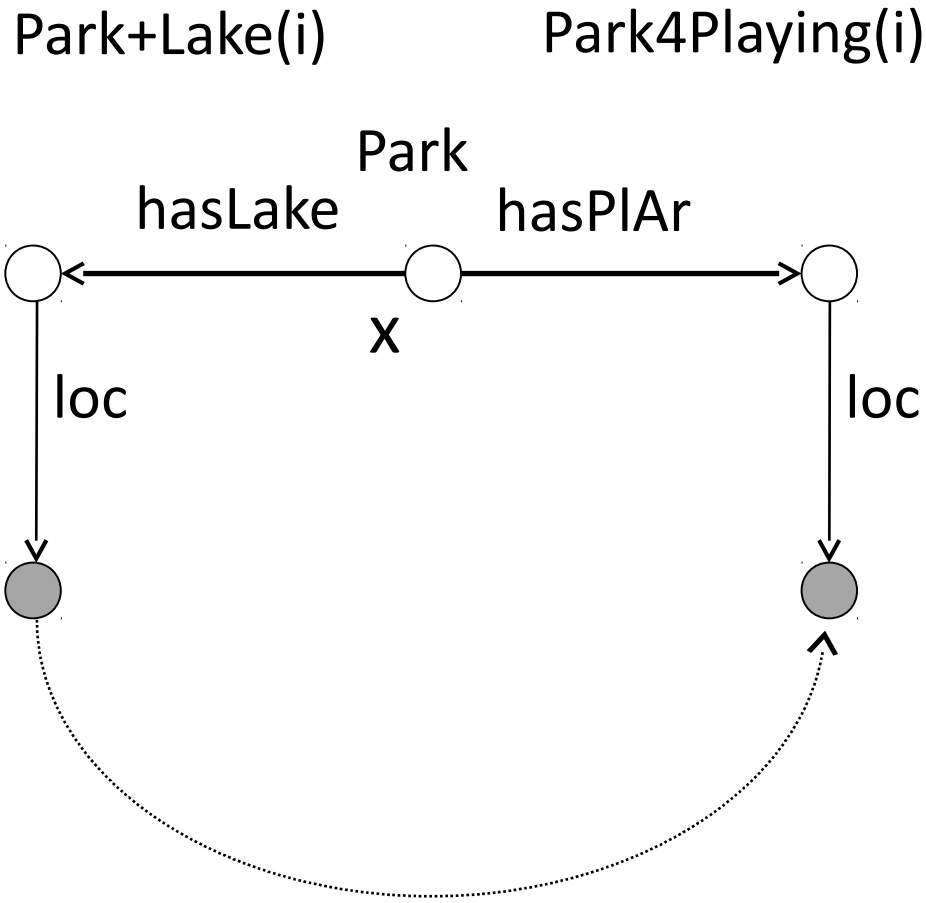
$$\begin{aligned} \mathit{Park{+}Lake} &\sqsubseteq \mathit{Park} \\ \mathit{Park4Playing} &\sqsubseteq \mathit{Park} \\ \mathit{Park{+}Lake} &\sqsubseteq \exists \mathit{hasLake} \circ \mathit{loc}, \mathit{loc}.\mathsf{tpp} \\ \mathit{Park4Playing} &\sqsubseteq \exists \mathit{hasPlAr} \circ \mathit{loc}, \mathit{loc}.\mathsf{tpp} \end{aligned}$$

- ABox contains $\mathit{Park{+}Lake}(i)$, $\mathit{Park4Playing}(i)$

Query formally

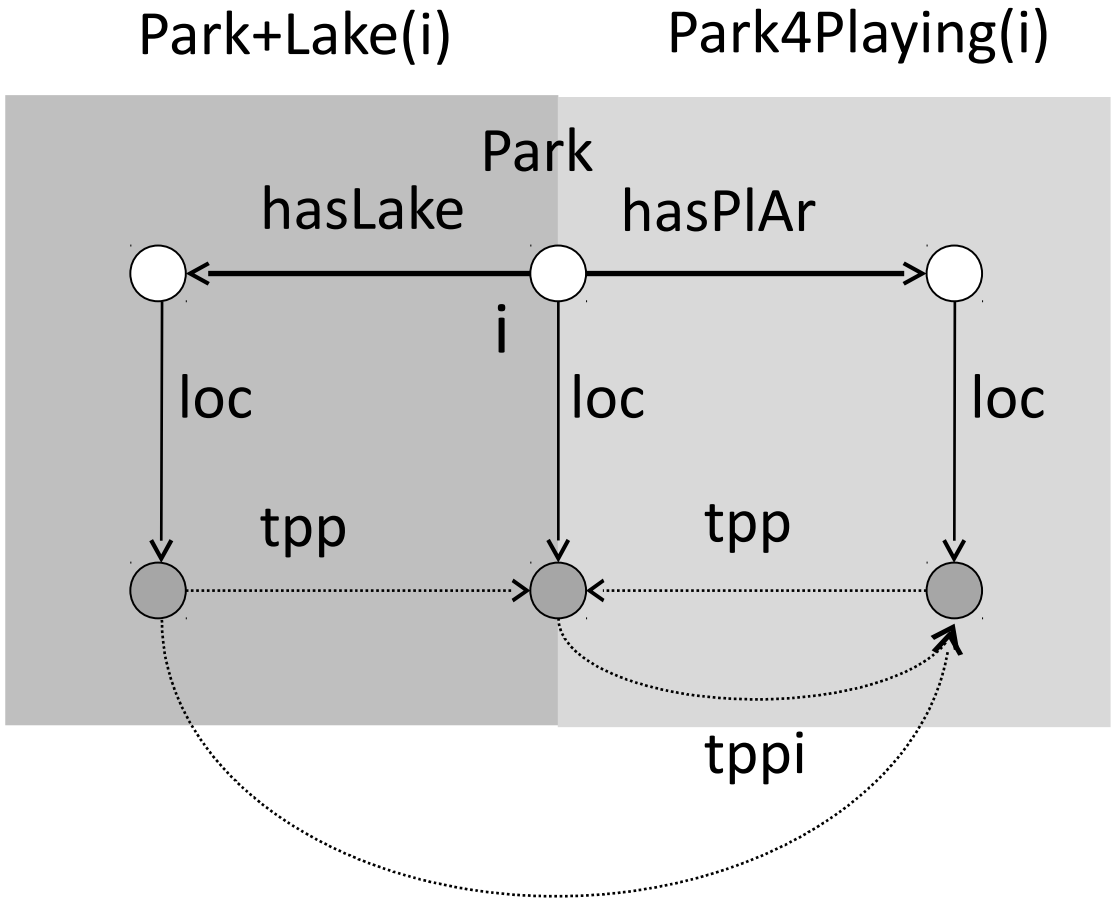
$$Q(x) = \mathit{Park}(x) \wedge \exists \mathit{hasLake} \circ \mathit{loc}, \mathit{hasPlAr} \circ \mathit{loc}.\mathcal{B}_{RCC8} \setminus \{\mathsf{ntpp}\}(x)$$

Is Design “i” safe?



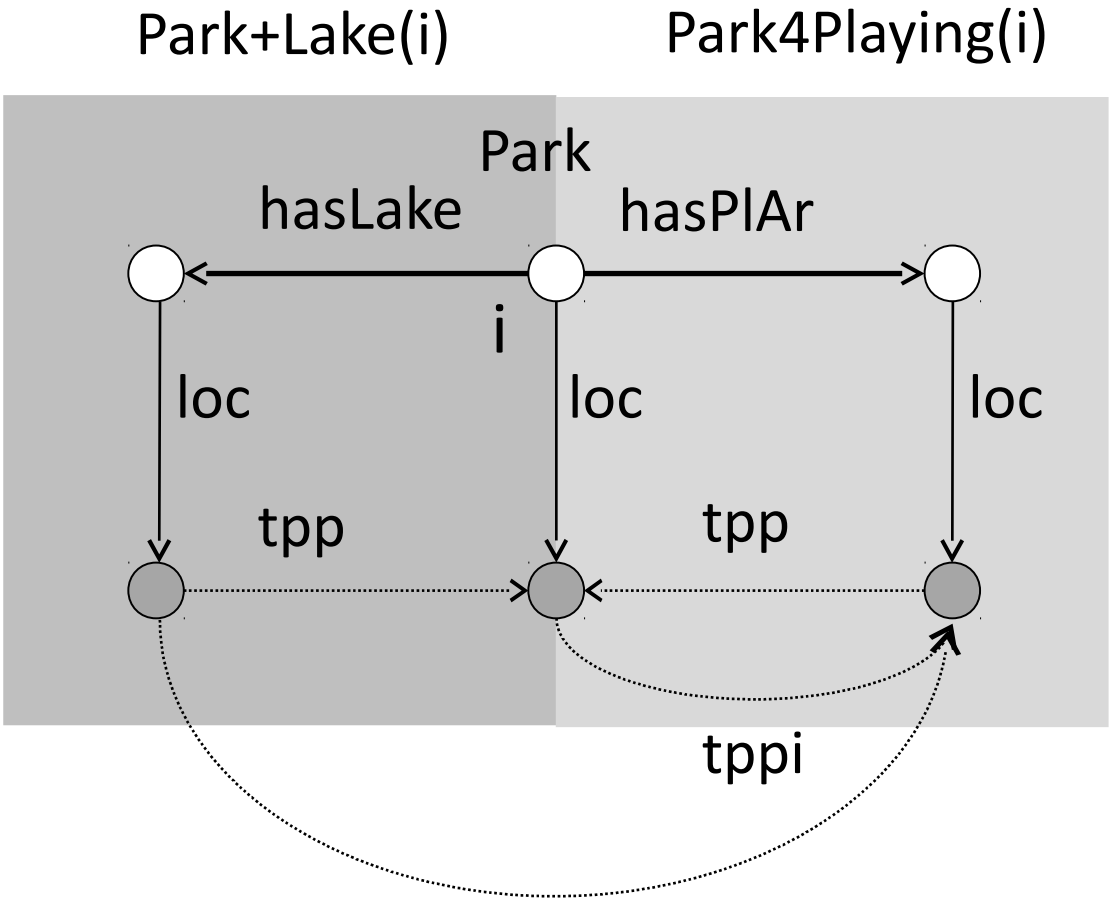
$\mathcal{B}_{RCC8} \setminus \{\text{ntpp}\}$

Is Design “i” safe?



{ dc, ec, po, tpp, tppi, eq }

Design “i” is safe



{ dc, ec, po, tpp, tppi, eq } $\subseteq \mathcal{B}_{RCC8} \setminus \{\text{ntpp}\}$

QA with a DL System?

- GDL-Lite-8 is decidable [Lutz-Milicic-2007]

- Why not extending the query rewriting idea to GDL-Lite-8?

The Whole Example

- TBox:

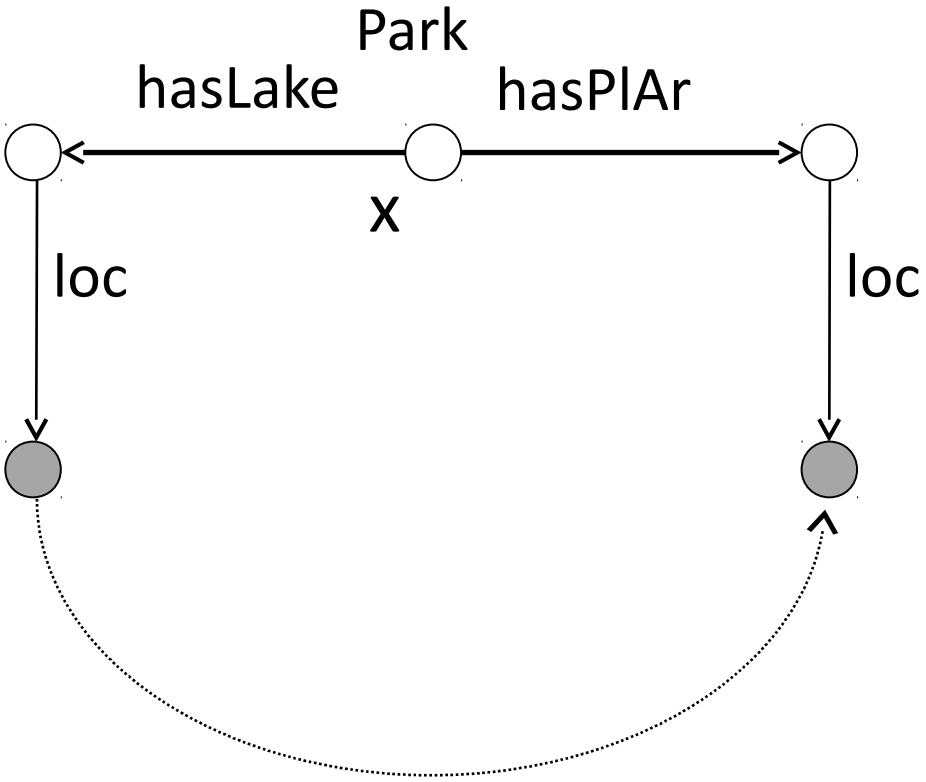
$$\begin{aligned} \mathit{Park{+}Lake} &\sqsubseteq \mathit{Park} \\ \mathit{Park4Playing} &\sqsubseteq \mathit{Park} \\ \mathit{Park{+}Lake} &\sqsubseteq \exists \mathit{hasLake} \circ \mathit{loc}, \mathit{loc}.\mathsf{tpp} \\ \mathit{Park4Playing} &\sqsubseteq \exists \mathit{hasPlAr} \circ \mathit{loc}, \mathit{loc}.\mathsf{tpp} \end{aligned}$$

- ABox contains $\mathit{Park{+}Lake}(i)$, $\mathit{Park4Playing}(i)$

Query formally

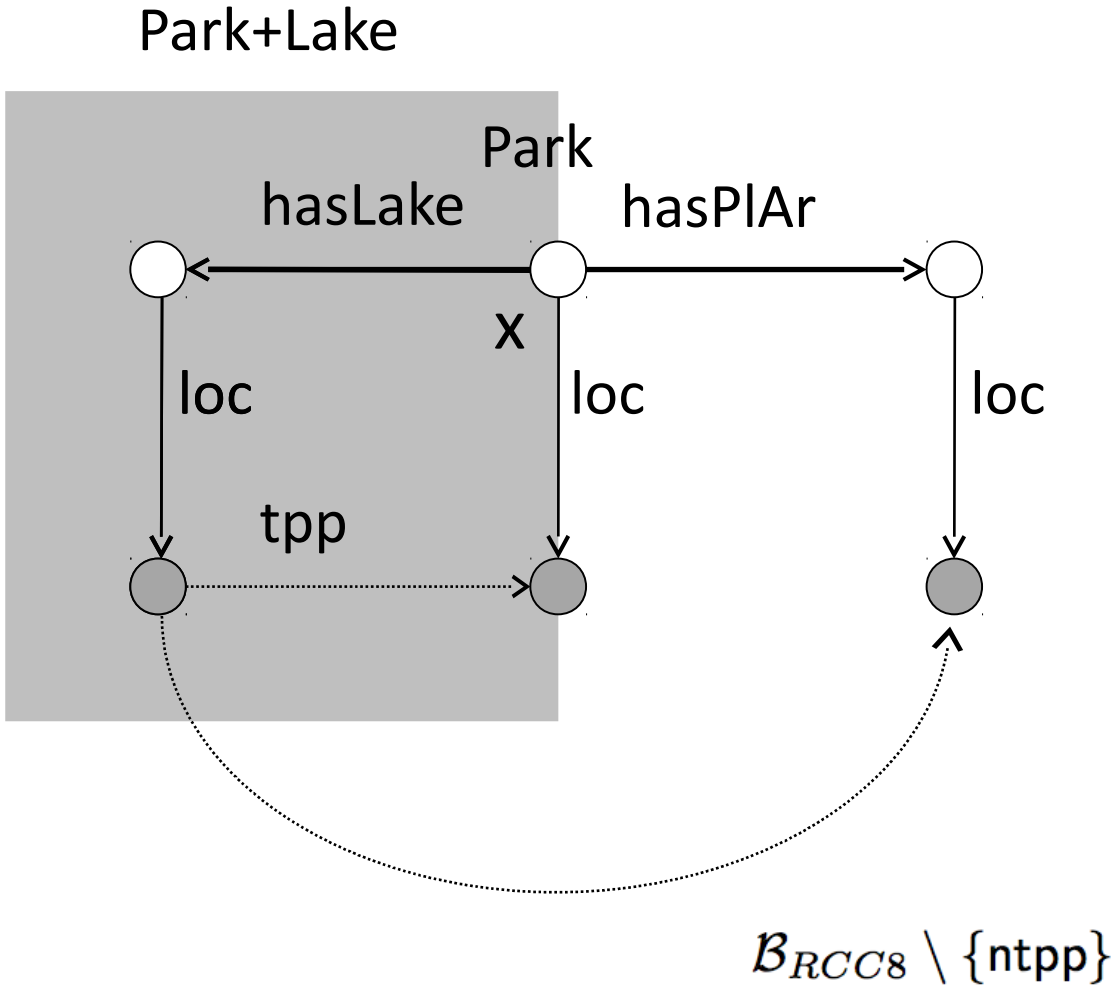
$$Q(x) = \mathit{Park}(x) \wedge \exists \mathit{hasLake} \circ \mathit{loc}, \mathit{hasPlAr} \circ \mathit{loc}.\mathcal{B}_{RCC8} \setminus \{\mathsf{ntpp}\}(x)$$

Rewriting the Example Query

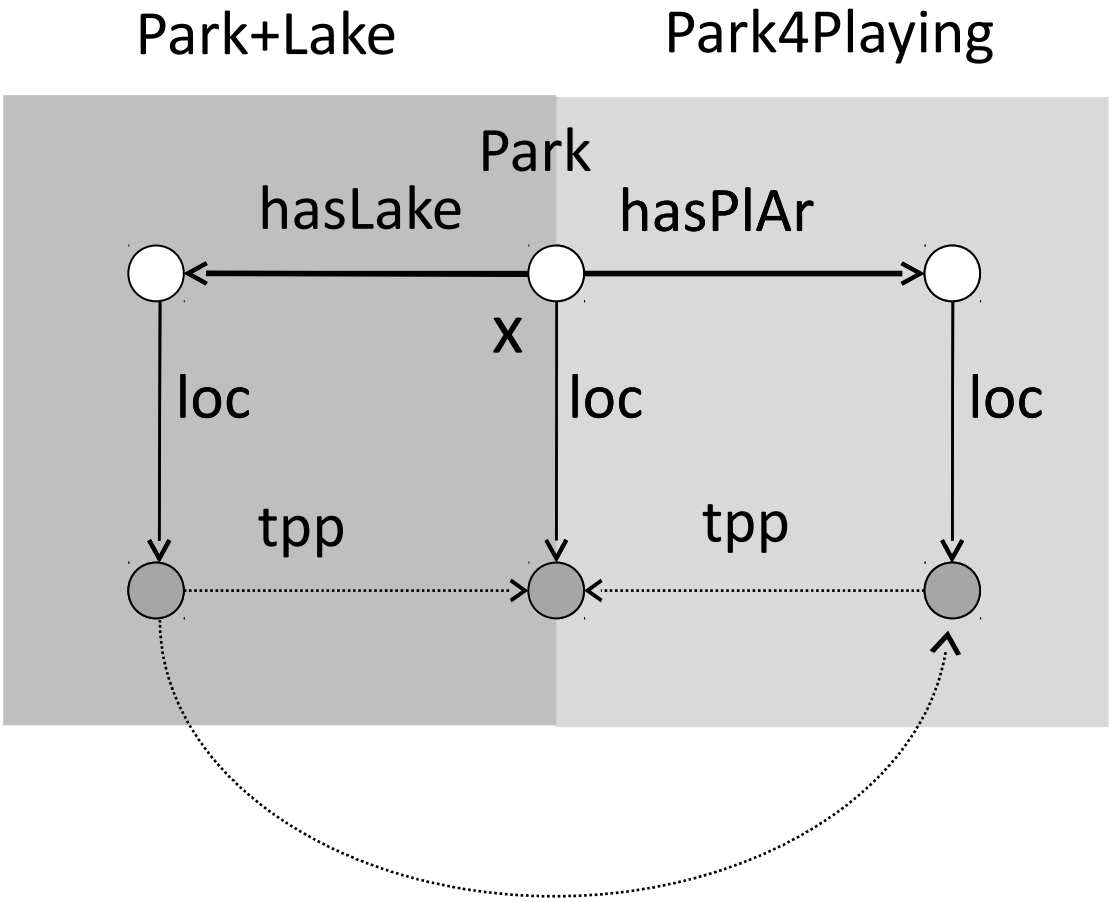


$\mathcal{B}_{RCC8} \setminus \{\text{ntpp}\}$

Rewriting the Example Query

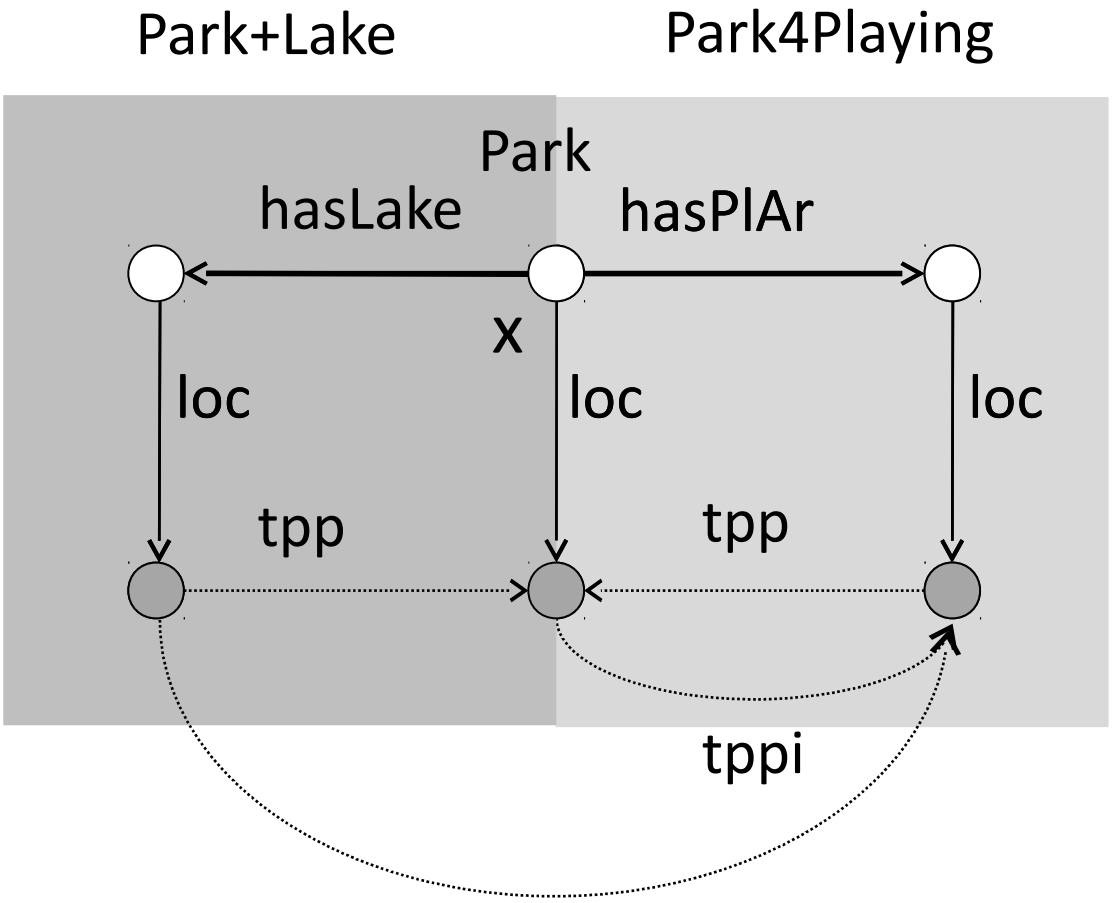


Rewriting the Example Query



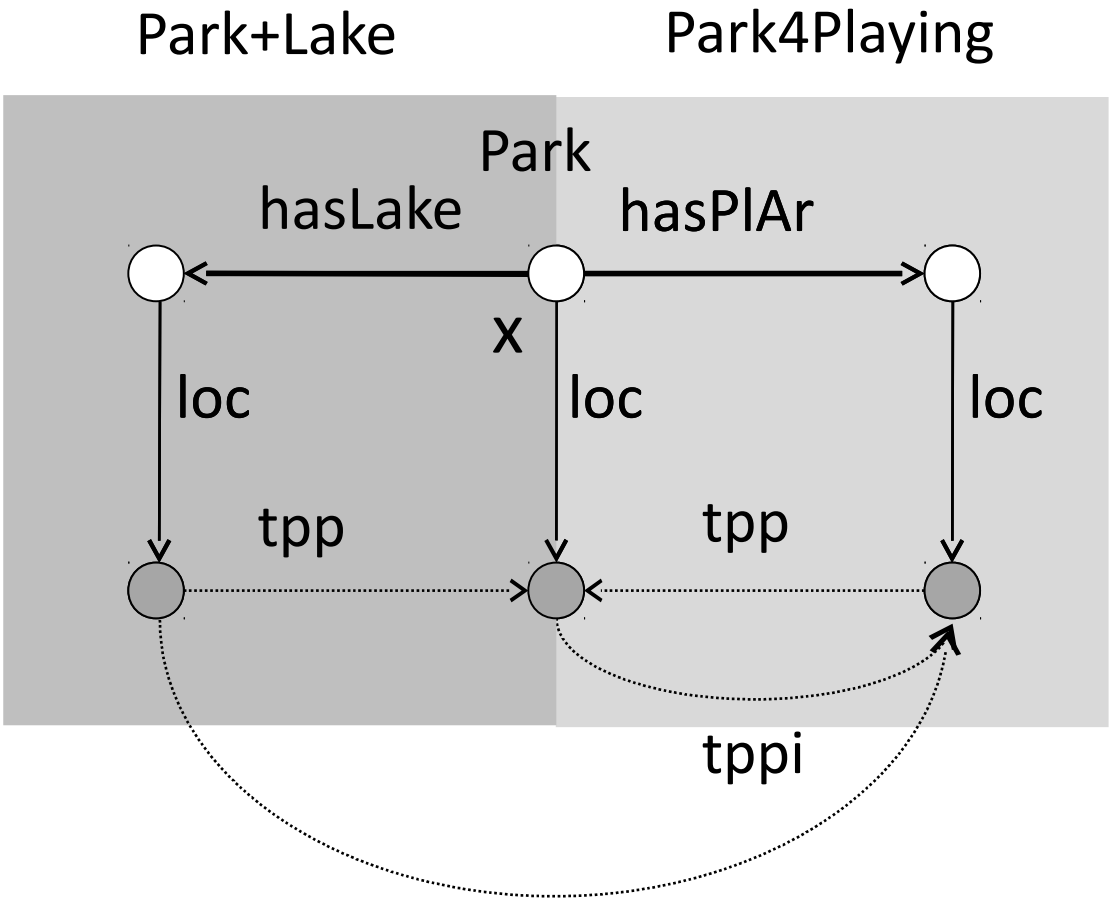
$\mathcal{B}_{RCC8} \setminus \{\text{ntpp}\}$

Rewriting the Example Query



$\mathcal{B}_{RCC8} \setminus \{\text{ntpp}\}$

Rewriting the Example Query

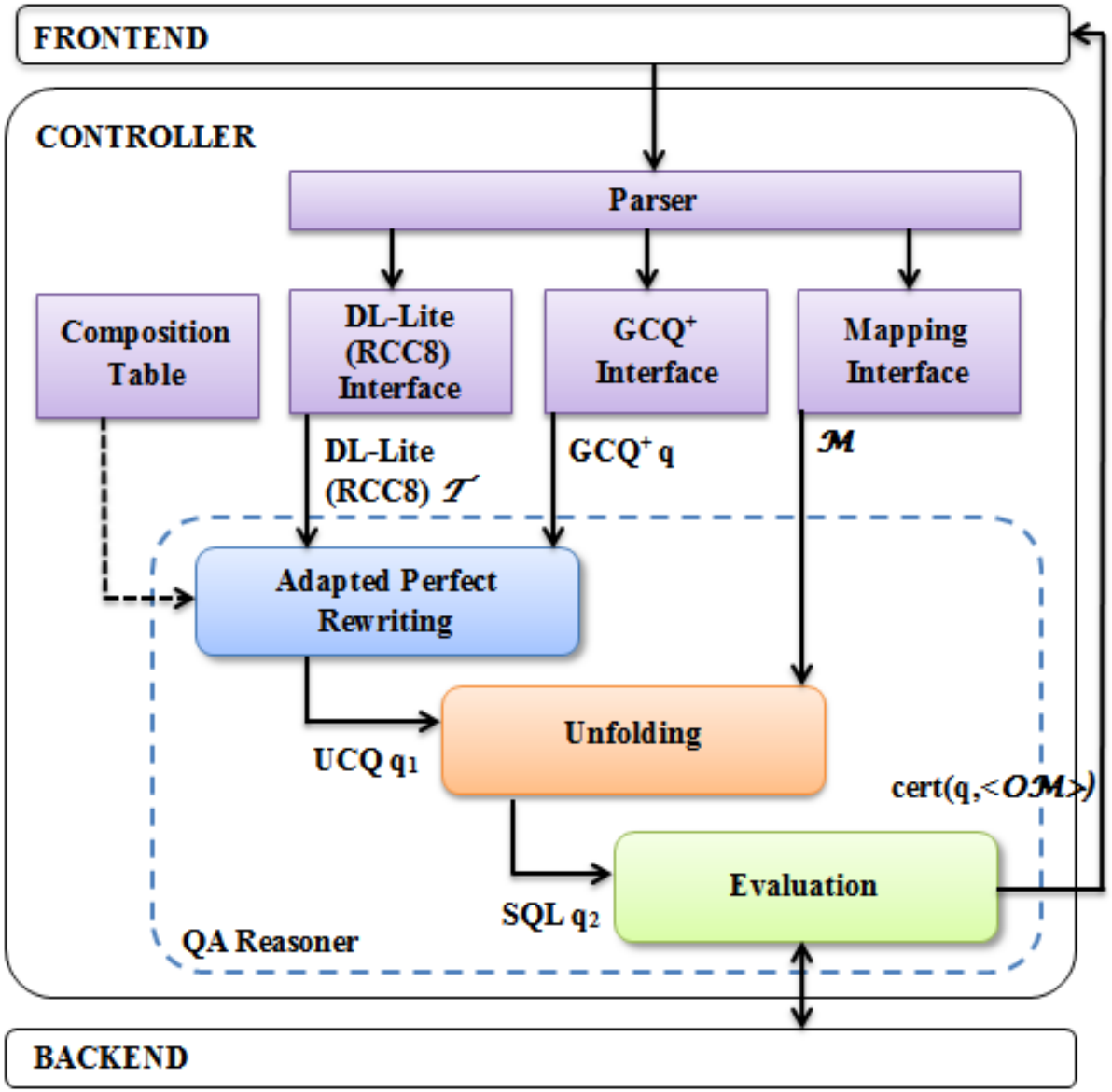


{ dc, ec, po, tpp, tppi, eq } $\subseteq \mathcal{B}_{RCC8} \setminus \{\text{ntpp}\}$

What‘s in the Repository?

- Only simple tables:
  - Park+Lake(i)
  - Park4Playing(i)
- Spatial query answering without spatial data!
- Standard SQL can do with UCQ→SQL mappings (Unfolding)

System Architecture

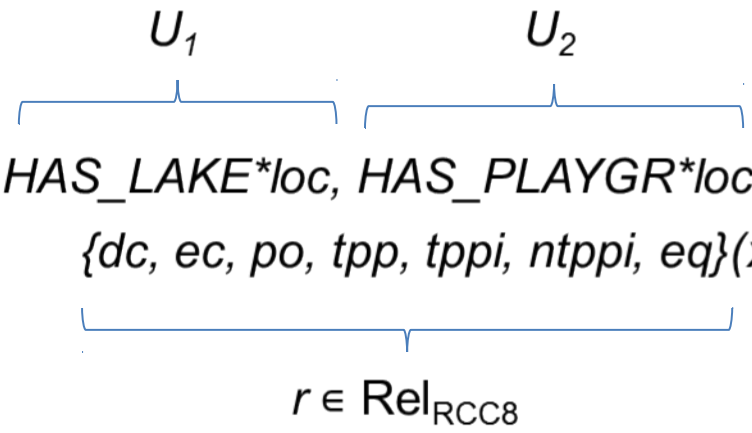


Adapted Query Rewriting

Adapted Perfect Rewriting Algorithm : an extension of the Perfect Rewriting Algorithm, handling GCQ$^+$ atoms of the form $\exists U_1, U_2 \,.r$ for $r \in \text{Rel}_{\text{RCC8}}$, by introducing 4 rewriting rules.

E.g.

GCQ$^+$ Query:



Performance Optimization

- **1st rule:** If a GCQ+ atom of the form $\exists R_1 \circ loc, R_2 \circ loc.r_3(x)$ occurs during the rewriting process, then it can be substituted by the conjunct of two new atoms of the form $\exists R_1 \circ loc, loc.r_1(x)$ and $\exists loc, R_2 \circ loc.r_2(x)$ in a new CQ for all r_1, r_2, contained in Rel_{RCC8} such that $r_1;r_2 \subseteq r_3$, namely where all possible compositions of the sets $(r_1 \circ r_2)$ from a full composition table are refinements of r_3.

- **Disadvantage**: The full composition table has 65025 possible combinations of pairs (r_1, r_2) => exponential blow-up, generating up to 130050 new query atoms for every input query atom in the form $\exists R_1 \circ loc, R_2 \circ loc.r_3(x)$

- **Optimization**: do not search for all $r_1;r_2 \subseteq r_3$, but search for all **maximal pairs** $\boldsymbol{r_1, r_2}$ such that $r_1;r_2 \subseteq r_3$ and do the reformulation process only w.r.t. these pairs.

  E.g. if $r_1;r_2 \subseteq r_3$, $r_4;r_5 \subseteq r_3$ and $r_4 \subseteq r_1$, $r_5 \subseteq r_2$, => then r_4, r_5 is redundant, since r_1, r_2 is the maximal pair

Rewriting the Example

$$Q = Park(x) \wedge \exists hasLake \circ loc, hasPlAr \circ loc.(\mathcal{B}_{RCC8} \setminus \{\mathsf{ntpp}\})(x)$$

$$Q' = (\exists hasLake \circ loc, loc.\mathsf{tpp})(x) \wedge (\exists loc, hasPlAr \circ loc.\mathsf{tppi})(x)$$

$$Q'' = Park{+}Lake(x) \wedge Park4Playing(x)$$

Preliminary Evaluation

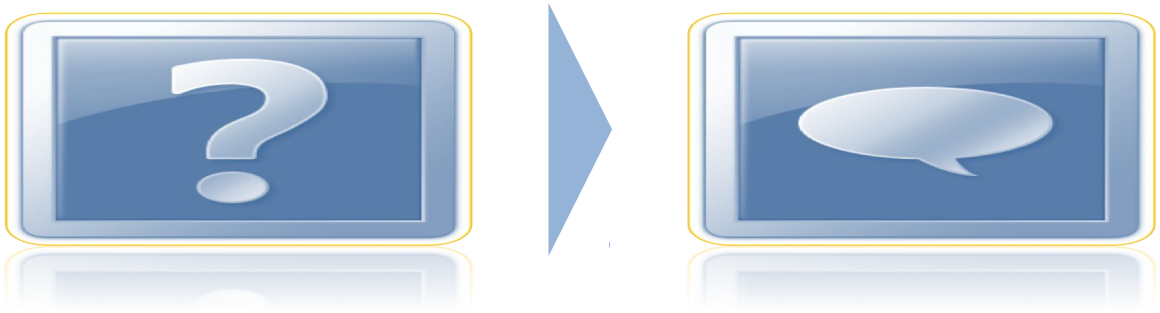
Test parameter	Exp. 1/ Result	Exp. 2/ Result	Exp. 3/ Result	Exp. 4/ Result	Exp. 5/ Result
Tbox					
concepts	5	5	3	5	5
roles	4	4	2	4	4
axioms	10	10	4	8	8
PIs	9	9	4	8	8
FeaturePaths	0	0	2	4	4
Input Query					
atoms	2	2	2	2	3
concept atoms	0	0	1	1	2
role atoms	2	2	0	0	0
FeaturePath atoms	0	0	1	1	1
Perfect Rewr. Algorithm					
version	adapted	original	adapted	adapted	adapted
iterations	3	3	4	4	4
added queries case 1	0	x	9	9	9
added queries case 2/3	0	x	11	11	11
added queries case 4	0	x	0	0	0
queries before drop( )	5	x	21	21	21
queries after drop( )	5	x	1	1	1
execution time in ms	5	4	47817	47985	48711

Summary and Outlook

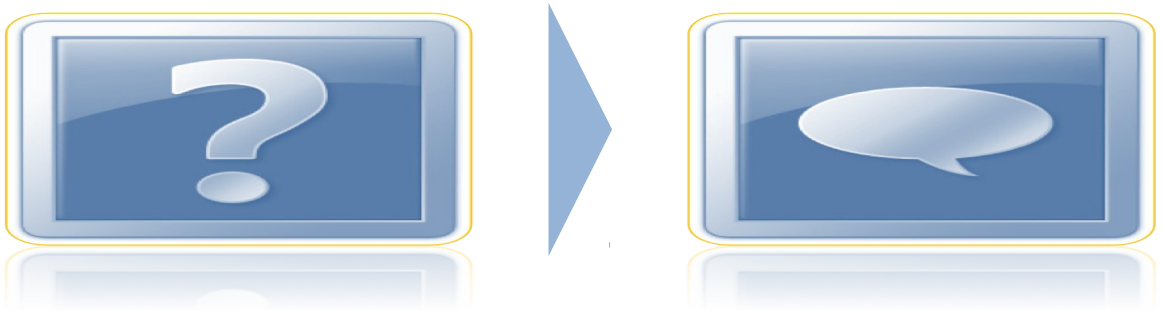
- GDL-Lite-8 modeling language
- GCQ+ as a querying language
- First steps towards implementation

- What if there is spatial information in the DB → Wednesday
- GDL-Lite-2, -3 → Wednesday
- TDL-Lite-13 (Allen) seems equally possible

Questions & Answers...



Questions & Answers...



This work was supported by
Deutsche Forschungsgemeinschaft (DFG)
as part of a grant for the project **GeoDL**

The work is now supported by the **European Commision** as part of the **FP7** project **Optique** (http://optique.project.ifi.uio.no)

Optique