First Steps Towards Taming Description Logics with Strings

Stéphane Demri

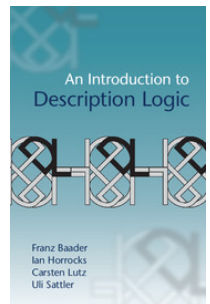
Joint work with Karin Quaas
(Leipzig University, Germany)

GT FM & AI, January 2024

Reasoning on Ontologies with Description Logics

- Ontology: formal specification of some domain with concepts, objects, relationships between concepts, objects, etc.
- Backbone of ontologies includes:
  - taxonomy (classification of objects),
  - axioms (to constrain the models of the defined terms).
- Description logics are well-known logical formalisms dedicated to ontologies.

  [Baader et al, Book 2017]



- BioPortal (`http://bioportal.bioontology.org/`): huge amount of ontologies such as Cell Ontology and Plant Ontology to facilitate scientific activities.
- Computational properties.
  - Acceptable trade-off between expressivity and complexity.
  - Decidability and often tractability.
  - Implementation in tools of the main reasoning tasks.

Description Logics with Concrete Domains

- Need to express concrete properties about data in ontologies. (e.g. age, duration, name, size, etc.)

- Concrete domain $\mathscr{D} = (\mathbb{D}, R_1, R_2, \dots)$: fixed non-empty domain with a family of relations.

- $(\mathbb{N}, <, +1)$, $(\mathbb{Q}, <, =)$, $(\mathbb{N}, <, =)$, $(\{0,1\}^*, <_{\mathsf{pre}}, <_{\mathsf{suf}})$.

- Concrete domain RCC8 with space regions in $\mathbb{R}^2$ contains topological relations between spatial regions.

  See e.g. [Wolter & Zakharyaschev, KR'00]

- General scheme for integrating concrete domains in DLs.

  [Baader & Hanschke, IJCAI'91]

  - declarative semantics close to the usual semantics for DLs,
  - generic extensions of DLs with various concrete domains,
  - tableaux-based algorithms combined with theory reasoning.

Methods for Handling Concrete Domains

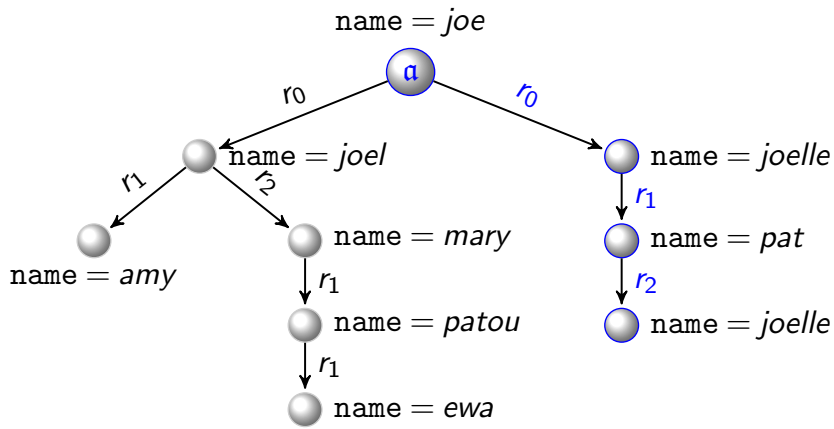
- Tableaux-based decision procedures for ω-admissible concrete domains. [Lutz & Miličić, JAR 2007]
  - $\mathscr{R} = (\mathbb{R}, <, =, (=_r)_{r \in \mathbb{R}})$ is ω-admissible.
  - $\mathscr{N} = (\mathbb{N}, <, =, (=_n)_{n \in \mathbb{N}})$ is not ω-admissible.

- Translation into a decidable extension of MSO with bounding quantifier B. [Carapelle & Turhan, ECAI'16]
  - EHD approach developped with Bool(MSO,WMSO+B) over infinite trees of finite branching degree. [Carapelle & Kartzow & Lorhey, JCSS 2016]
  - Decidability of concept satisfiability problem w.r.t. general TBoxes for $\mathscr{ALC}(\mathscr{N})$.

- Translation into Rabin tree automata over finite alphabets using approximations for satisfiable symbolic interpretations.
  - Concept satisfiability problem w.r.t. general TBoxes for $\mathscr{ALC}(\mathscr{N})$ in ExpTime. [Labai & Ortiz & Šimkus, KR'20]

Finite Strings with the Prefix Relation

- $\mathscr{D}_\Sigma = (\Sigma^*, <_{\text{pre}}, =, (=_w)_{w \in \Sigma^*})$.
- $(\mathbb{N}, <, =, (=_n)_{n \in \mathbb{N}})$ corresponds to $\mathscr{D}_\Sigma$ with singleton Σ.



$$\big(\exists r_0 r_1 r_2 \cdot (\texttt{name} <_{\text{pre}} SSS\ \texttt{name})\big) \wedge \forall r_0 \cdot (\texttt{name} <_{\text{pre}} S\ \texttt{name})$$

('S' similar to neXt in temporal logic)

- Concept satisfiability problem w.r.t. general TBoxes for $\mathscr{ALC}(\mathscr{W})$ with $\mathscr{W} = (\Sigma^*, \cdot, =, (=_w)_{w \in \Sigma^*})$ is undecidable.

[Lutz, PhD 2002]

- No known decidability results for description logics with $\mathscr{D}_\Sigma$.

$\mathscr{ALC}$ in a Nutshell

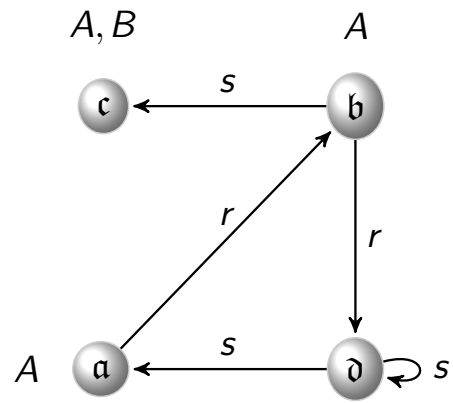
- Complex concepts.

$$C ::= \top \mid \bot \mid A \mid \neg C \mid C \sqcap C \mid C \sqcup C \mid \exists r.C \mid \forall r.C,$$

with concept names A and role names r.

- Interpretation $\mathscr{I} \stackrel{\text{def}}{=} (\Delta^{\mathscr{I}}, \cdot^{\mathscr{I}})$
  - $\Delta^{\mathscr{I}}$: non-empty set (the *domain*).
  - $\cdot^{\mathscr{I}}$: *interpretation function* such that

$$A^{\mathscr{I}} \subseteq \Delta^{\mathscr{I}} \qquad r^{\mathscr{I}} \subseteq \Delta^{\mathscr{I}} \times \Delta^{\mathscr{I}}$$



Concept name A/role name r

$\approx$

unary predicate/binary predicate

Set-Theoretical Semantics for Complex Concepts

$$
\begin{aligned}
\top^{\mathscr{I}} &\stackrel{\text{def}}{=} \Delta^{\mathscr{I}} \\
\bot^{\mathscr{I}} &\stackrel{\text{def}}{=} \emptyset \\
(\neg C)^{\mathscr{I}} &\stackrel{\text{def}}{=} \Delta^{\mathscr{I}} \setminus C^{\mathscr{I}} \\
(C_1 \sqcup C_2)^{\mathscr{I}} &\stackrel{\text{def}}{=} C_1^{\mathscr{I}} \cup C_2^{\mathscr{I}} \\
(C_1 \sqcap C_2)^{\mathscr{I}} &\stackrel{\text{def}}{=} C_1^{\mathscr{I}} \cap C_2^{\mathscr{I}} \\
(\exists r.C)^{\mathscr{I}} &\stackrel{\text{def}}{=} \{\mathfrak{a} \in \Delta^{\mathscr{I}} \mid r^{\mathscr{I}}(\mathfrak{a}) \cap C^{\mathscr{I}} \neq \emptyset\} \\
(\forall r.C)^{\mathscr{I}} &\stackrel{\text{def}}{=} \{\mathfrak{a} \in \Delta^{\mathscr{I}} \mid r^{\mathscr{I}}(\mathfrak{a}) \subseteq C^{\mathscr{I}}\}
\end{aligned}
$$

$$
\mathscr{R}(\mathfrak{a}) \stackrel{\text{def}}{=} \{\mathfrak{b} \mid (\mathfrak{a}, \mathfrak{b}) \in \mathscr{R}\}
$$

Inclusion and Decision Problem $\mathrm{TSAT}(\mathscr{ALC})$

- General concept inclusions $C \sqsubseteq D$ (GCIs).
  E.g., $\texttt{Employee} \sqsubseteq \exists \texttt{WorksFor}.\ \top$

$$\mathscr{I} \models C \sqsubseteq D \quad \overset{\text{def}}{\Leftrightarrow} \quad C^{\mathscr{I}} \subseteq D^{\mathscr{I}}$$

- Terminological Box (TBox) $\mathscr{T}$: finite collection of GCIs.
- Interpretation $\mathscr{I} = (\Delta^{\mathscr{I}}, \cdot^{\mathscr{I}})$, TBox $\mathscr{T}$.

$$\mathscr{I} \models \mathscr{T} \quad \overset{\text{def}}{\Leftrightarrow} \quad \text{for all } C \sqsubseteq D \in \mathscr{T},\ \mathscr{I} \models C \sqsubseteq D$$

- *Concept satisfiability problem w.r.t. general TBoxes* ($\mathrm{TSAT}(\mathscr{ALC})$):

  Input: A concept C_0 and a TBox $\mathscr{T}$.

  Question: Is there an interpretation $\mathscr{I}$ such that $\mathscr{I} \models \mathscr{T}$ and $C_0^{\mathscr{I}} \neq \emptyset$?

- $\mathrm{TSAT}(\mathscr{ALC})$ is ExpTime-complete.

Description Logic $\mathscr{ALCF}^{\mathscr{P}}(\mathscr{D}_\Sigma)$

$$\exists r_0 r_1 r_2 \cdot (\mathtt{name} <_{\mathsf{pre}} SSS\ \mathtt{name})$$

- New (atomic) concepts of the form $\exists P.[\![\Theta]\!]$ and $\forall P.[\![\Theta]\!]$:
  - non-empty sequence P of role names (*role path*),
  - Boolean constraint Θ built over terms of the form $S^j\mathtt{x}$ with $j \leq |P|$ and atomic constraints of the form

$$\mathtt{t} <_{\mathsf{pre}} \mathtt{t}' \qquad \mathtt{t} = \mathtt{t}' \qquad =_{\mathfrak{w}} (\mathtt{t}) \text{ (also written } \mathtt{t} = \mathfrak{w})$$

- Interpretation $\mathscr{I} = (\Delta^{\mathscr{I}}, \cdot^{\mathscr{I}}, \mathfrak{v})$ with $\mathfrak{v}\colon \Delta^{\mathscr{I}} \times \mathrm{VAR} \to \Sigma^*$.

- $(r_1 r_2 \ldots r_n)^{\mathscr{I}} \stackrel{\text{def}}{=}$ set of tuples $(\mathfrak{a}_0, \ldots, \mathfrak{a}_n)$ in $(\Delta^{\mathscr{I}})^{n+1}$ such that $(\mathfrak{a}_{i-1}, \mathfrak{a}_i) \in r_i^{\mathscr{I}}$ for all $i \in [1, n]$.

- Satisfaction relation $\mathscr{I}, \pi = (\mathfrak{a}_0, \mathfrak{a}_1, \ldots, \mathfrak{a}_n) \models \Theta$:

  - $\mathscr{I}, \pi \models S^i\mathtt{x} <_{\mathsf{pre}} S^j\mathtt{y} \stackrel{\text{def}}{\Leftrightarrow} \mathfrak{v}(\mathfrak{a}_i, \mathtt{x}) <_{\mathsf{pre}} \mathfrak{v}(\mathfrak{a}_j, \mathtt{y})$, (similar for $=$)
  - $\mathscr{I}, \pi \models S^i\mathtt{x} = \mathfrak{w} \stackrel{\text{def}}{\Leftrightarrow} \mathfrak{v}(\mathfrak{a}_i, \mathtt{x}) = \mathfrak{w}$,

Clauses for Interpreting the Concepts Involving $\mathscr{D}_\Sigma$

$$(\exists P.[\![\Theta]\!])^{\mathscr{I}} \stackrel{\text{def}}{=}$$

$$\{\mathfrak{a}_0 \in \Delta^{\mathscr{I}} \mid \exists\, \mathfrak{a}_1, \ldots, \mathfrak{a}_n \in \Delta^{\mathscr{I}} \text{ s.t. } \pi = (\mathfrak{a}_0, \mathfrak{a}_1, \ldots, \mathfrak{a}_n) \in P^{\mathscr{I}} \text{ and } \mathscr{I}, \pi \models \Theta\}$$

$$(\forall P.[\![\Theta]\!])^{\mathscr{I}} \stackrel{\text{def}}{=}$$

$$\{\mathfrak{a}_0 \in \Delta^{\mathscr{I}} \mid \forall\, \mathfrak{a}_1, \ldots, \mathfrak{a}_n \in \Delta^{\mathscr{I}}, \pi = (\mathfrak{a}_0, \mathfrak{a}_1, \ldots, \mathfrak{a}_n) \in P^{\mathscr{I}} \text{ implies } \mathscr{I}, \pi \models \Theta\}$$

- $\mathscr{ALCF}^{\mathscr{P}}(\mathscr{D}_\Sigma)$ has also functional role names, omitted today.
- Integrating concrete domains may come with some extra costs: we need to solve a (potentially infinite) constraint satisfaction problem.
- No finite interpretation property, e.g. with

$$\mathscr{T} = \{\top \sqsubseteq \exists r.[\![\mathrm{x} <_{\mathsf{pre}} S\mathrm{x}]\!]\}$$

Main Steps For Getting ExpTime Upper Bound

- Automata-based approach with tree constraint automata (TCA) accepting infinite data trees with domain Σ^*.

- Step 0: to transform the instance so that every concept is in *simple form*, proper form to perform Step 1.

- Step 1: $C_0, \mathscr{T} \longrightarrow \mathbb{A}$ s.t. $C_0, \mathscr{T}$ is a positive instance of TSAT iff $\mathrm{L}(\mathbb{A}) \neq \emptyset$.

- Step 2: $\mathbb{A}$ (for $\mathscr{D}_\Sigma$) $\longrightarrow \mathbb{A}'$ (for $\mathscr{N}$) such that

$$\mathrm{L}(\mathbb{A}) \neq \emptyset \;\text{ iff }\; \mathrm{L}(\mathbb{A}') \neq \emptyset$$

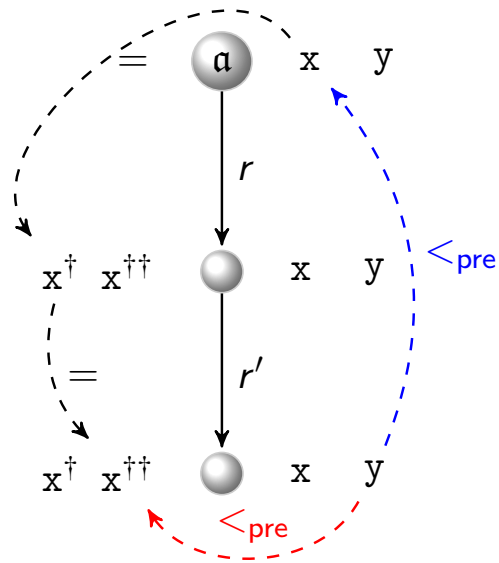
- Step 3: complexity analysis to get ExpTime-completeness of TSAT. (based on [Demri & Quaas, CONCUR'23]).

Concepts in Simple Form

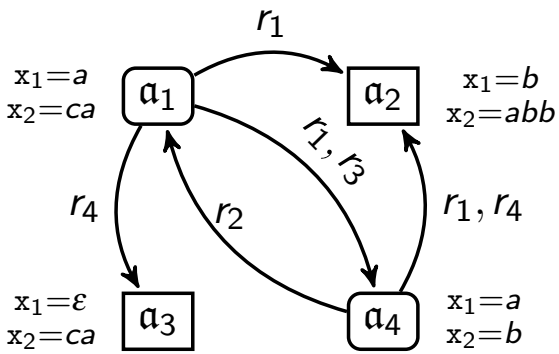
- A concept is in simple form if it is in NNF and all its role paths are of length at most one.
- $\exists rr'.[\![S^2\mathrm{y} <_{\mathsf{pre}} \mathrm{x}]\!]$ not in simple form and concepts below in simple form:

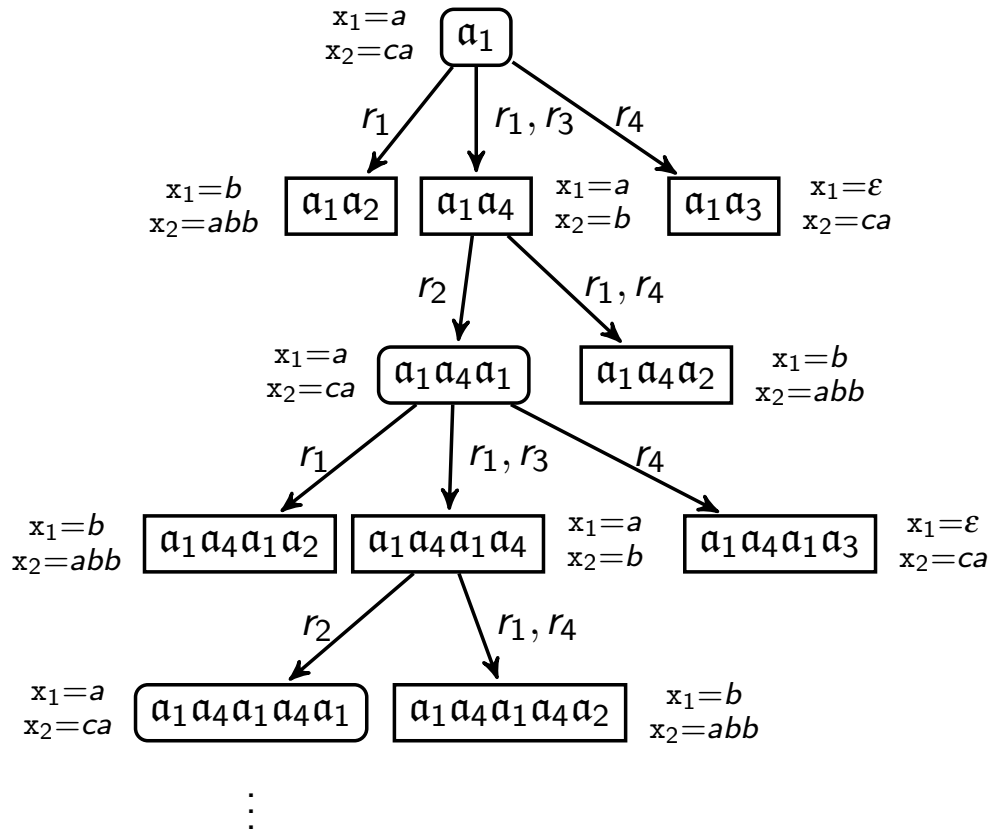
  $\exists r.\exists r'.\exists \varepsilon.[\![\mathrm{y} <_{\mathsf{pre}} \mathrm{x}^{\dagger\dagger}]\!] \quad \top \sqsubseteq \forall r.[\![\mathrm{x} = S\mathrm{x}^{\dagger}]\!] \quad \top \sqsubseteq \forall r'.[\![\mathrm{x}^{\dagger} = S\mathrm{x}^{\dagger\dagger}]\!]$

- Given $C_0, \mathscr{T}$, one can construct in polynomial-time $C_0', \mathscr{T}'$ in simple form s.t. $C_0, \mathscr{T}$ positive instance of TSAT iff $C_0', \mathscr{T}'$ positive instance of TSAT.



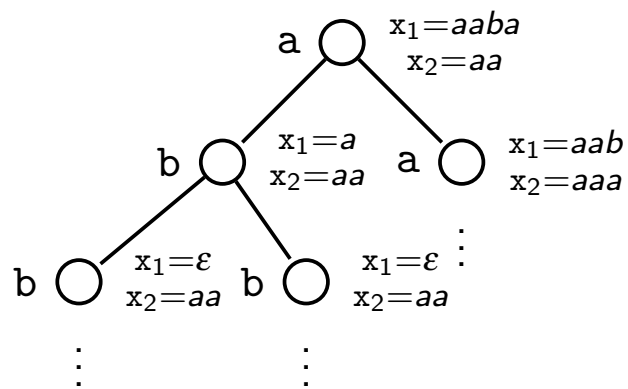
Tree Interpretation Property Needed!

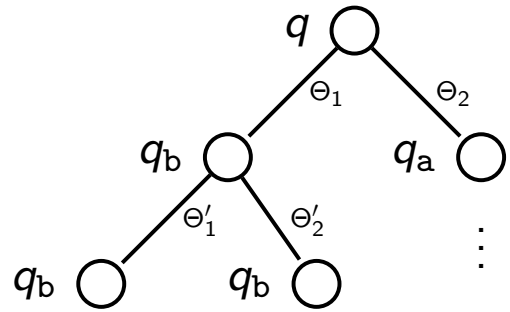




Tree Automata on Strings

(or how to recognize infinite data trees)





- $\mathbb{A}$ accepts infinite trees $\mathbf{t} : [0, d-1]^* \to (\mathbf{\Sigma} \times (\Sigma^*)^\beta)$.

(ye

- Transitions $(q, \mathrm{a}, (\Theta_1, q_1), \ldots, (\Theta_d, q_d))$ put constraints on values of current node and children nodes.
- Büchi and Rabin acceptance conditions.
- Can be adapted to many concrete domains and extends similar definitions for the linear case with $d = 1$.

  E.g. [Segoufin & Toruńczyk, STACS'11; Kartzow & Weider, arXiv 2015]

From TSAT to Nonemptiness

(or how to apply the standard automata-based approach)

- Technically involved construction following a standard pattern.
  - $C_0, \mathscr{T}$ in simple form positive instance of TSAT iff $\mathrm{L}(\mathbb{A}) \neq \emptyset$.
  - Locations are propositionally $\mathscr{T}$-consistent set of subconcepts.
  - Each role name has dedicated directions in $[1, d]$.
  - Constraints at the level of concepts translated at the level of transitions.

- Postponing the actual problem to the nonemptiness problem.

- Advantages of translating into TCA:
  - Reveals the size of automaton (depending on parameters like number of variables, maximal size for constants, etc.) – important for complexity!
  - Same construction for other concrete domains.

From String Constraints to Integer Constraints

- Intuition: encoding of prefix of strings $\mathfrak{w}$ and $\mathfrak{w}'$ by length of common prefix (which is a nonnegative integer).

- $\mathit{clen}(\mathfrak{w}, \mathfrak{w}')$: length of longest common prefix btw. $\mathfrak{w}$ and $\mathfrak{w}'$. E.g. $\mathit{clen}(\mathit{aba}, \mathit{abbbab}) = 2$.

- Properties (I)–(III) are "complete" to *recover string values in a greedy way* ($k = \mathrm{card}(\Sigma)$). [Demri & Deters, JLC 2015]

(I) For $\mathfrak{w}, \mathfrak{w}' \in \Sigma^*$, $|\mathfrak{w}| = \mathit{clen}(\mathfrak{w}, \mathfrak{w}) \geq \mathit{clen}(\mathfrak{w}, \mathfrak{w}')$.

(II) For all $\mathfrak{w}_0, \mathfrak{w}_1, \ldots, \mathfrak{w}_k \in \Sigma^*$ such that
  - $\mathit{clen}(\mathfrak{w}_0, \mathfrak{w}_1) = \cdots = \mathit{clen}(\mathfrak{w}_0, \mathfrak{w}_k)$ and,
  - for all $i \in [0, k]$, $\mathit{clen}(\mathfrak{w}_0, \mathfrak{w}_1) < |\mathfrak{w}_i|$,

  there are $i \neq j \in [1, k]$ such that $\mathit{clen}(\mathfrak{w}_0, \mathfrak{w}_1) < \mathit{clen}(\mathfrak{w}_i, \mathfrak{w}_j)$.

(III) For all $\mathfrak{w}_0, \mathfrak{w}_1, \mathfrak{w}_2 \in \Sigma^*$,
$\mathit{clen}(\mathfrak{w}_0, \mathfrak{w}_1) < \mathit{clen}(\mathfrak{w}_1, \mathfrak{w}_2)$ implies $\mathit{clen}(\mathfrak{w}_0, \mathfrak{w}_1) = \mathit{clen}(\mathfrak{w}_0, \mathfrak{w}_2)$.

Lifting at the Level of Automata

- TCA $\mathbb{A} = (Q, \mathbf{\Sigma}, d, \beta, Q_{\mathsf{in}}, \delta, F)$ on $\mathscr{D}_\Sigma$ translated into $\mathbb{A}' = (Q, \mathbf{\Sigma}, d, \beta', Q_{\mathsf{in}}, \delta', F)$ on $\mathscr{N}$.

- $\mathrm{L}(\mathbb{A}) \neq \emptyset$ iff $\mathrm{L}(\mathbb{A}') \neq \emptyset$.

- $\mathrm{L}(\mathbb{A}') \neq \emptyset$ checked in time

$$R_1\big(\mathsf{card}(Q) \times \mathsf{card}(\delta') \times \mathtt{MCS}(\mathbb{A}') \times \mathsf{card}(\mathbf{\Sigma}) \times R_2(\beta')\big)^{\mathscr{O}(R_2(\beta') \times R_3(d))}$$

[Demri & Quaas, CONCUR'23]

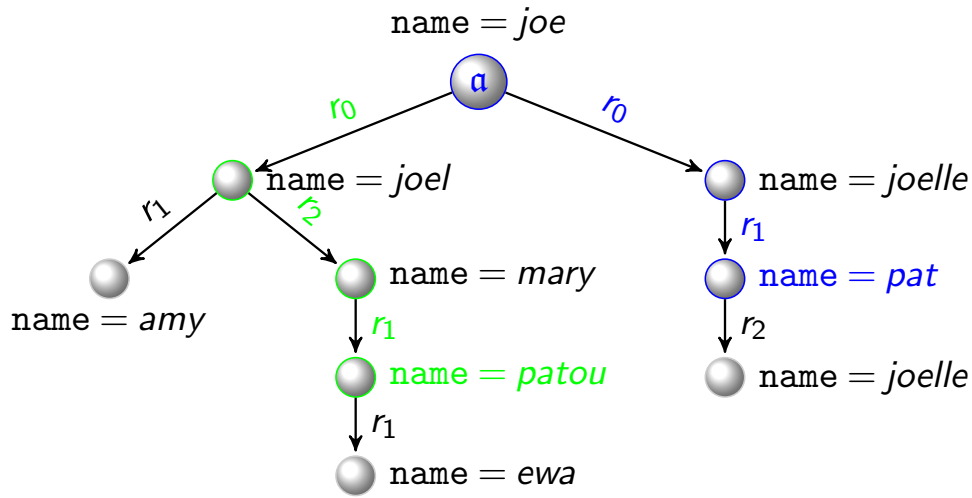
  - the R_i's are polynomials,
  - $\mathtt{MCS}(\mathbb{A}')$: maximal size of a constraint in $\mathbb{A}'$,
  - $\mathtt{MCS}(\mathbb{A}')$ in $(\beta + \mathtt{MCS}(\mathbb{A}) \times \mathit{card}(\delta) \times d)^{\mathscr{O}(\mathsf{card}(\Sigma)+3)}$,
  - β' polynomial in β and in the number of constant strings in $\mathbb{A}$.

Final Complexity Analysis

$C_0, \mathscr{T}$ is a positive instance iff $\mathrm{L}(\mathbb{A}) \neq \emptyset$ and $\mathbb{A}$ satisfies the following quantitative properties:

- Degree d bounded by $\mathtt{size}(C_0, \mathscr{T})$.
- Number of locations in $2^{\mathscr{O}(\mathtt{size}(C_0, \mathscr{T}))}$.
- Number of transitions in $2^{\mathscr{O}(R(\mathtt{size}(C_0, \mathscr{T})))}$ for some polynomial $R(\cdot)$.
- Number of variables β bounded by $\mathtt{size}(C_0, \mathscr{T})$.
- Cardinality of finite alphabet $\mathbf{\Sigma}$ bounded by $2^{\mathtt{size}(C_0, \mathscr{T})}$.
- $\mathtt{MCS}(\mathbb{A})$ quadratic in $\mathtt{size}(C_0, \mathscr{T})$.

Open Problem Related to XPath on Data Trees



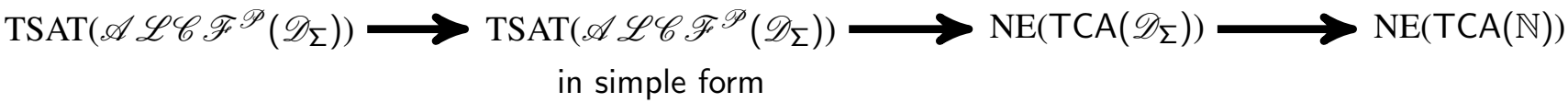
$$\left(\exists r_0 r_1 \cdot \texttt{name} <_{\text{pre}} \exists r_0 r_2 r_1 \cdot \texttt{name}\right)$$

- How to handle this extension and characterise its complexity? ExpTime-membership ?

- Can we adapt results about XPath on data trees?

  See e.g. [Figueira, ToCL 2012]

Concluding Remarks



- First steps towards taming description logics over strings.
- Automata-based approach with tree constraint automata.
- Reuse or adaptations of several results from literature with new insights to combine them.
- How to extend the results with suffix relation $<_{\mathsf{suf}}$ or regularity constraints?