Counting in Trees for Free

Anca Muscholl

Joint work with:

Helmut Seidl, Thomas Schwentick, Peter Habermehl

TU Munich, Marburg, LIAFA & Univ. Paris 7

ICALP 2004

An e-donkey user:

```
<doc>
 <user>
  <name> Charlie Brown </name>
  <email> cb@comics.org </email>
  <music> ... </music>
  <video> ... </video>
  <images> ... </images>
 </user>
 ...
</doc>
```

His downloaded music:

```
<jazz>
 <album> Always let me go </album>
 <artist> Keith Jarrett </artist>
 <year> 2002 </year>
 <review> ... </review>
 <review> ... </review>
 <price> 24 </price>
</jazz>
<pop>
 <title> Just my imagination</title>
 <artist> The Cranberries </artist>
 <year> 2002 </year>
 <album> Stars </album>
 <price> 20 </price>
 <review> ... </review>
</pop>
```

```
<french>
 <title> Aux enfants de la chance </title>
 <artist> Serge Gainsbourg </artist>
 <album> Serge Gainsbourg, vol. 3 </album>
 <price> 30 </price>
 <review> ... </review>
 <review> ... </review>
</french>

<classic>
 <title> La Cenerentola </title>
 <comp> Rossini </comp>
 <performer> Bartoli </performer>
 <recorded> 2003 </recorded>
 <price> 40 </price>
</classic>
...
```

```
<jazz>
 <album> Kind of Blue </album>
 <artist> Miles Davis </artist>
 <year> 1997 </year>
 <price> 22 </price>
 <review> ... </review>
 <review> ... </review>
 <review> ... </review>
</jazz>
```

... in essence:

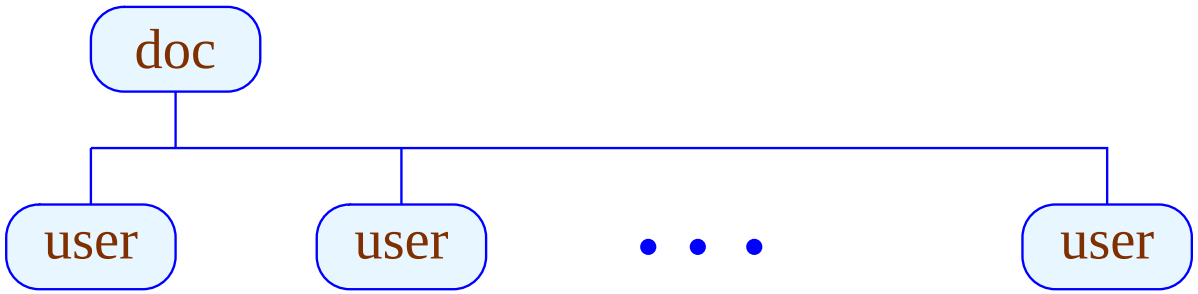
→ hierarchically structured through tags;

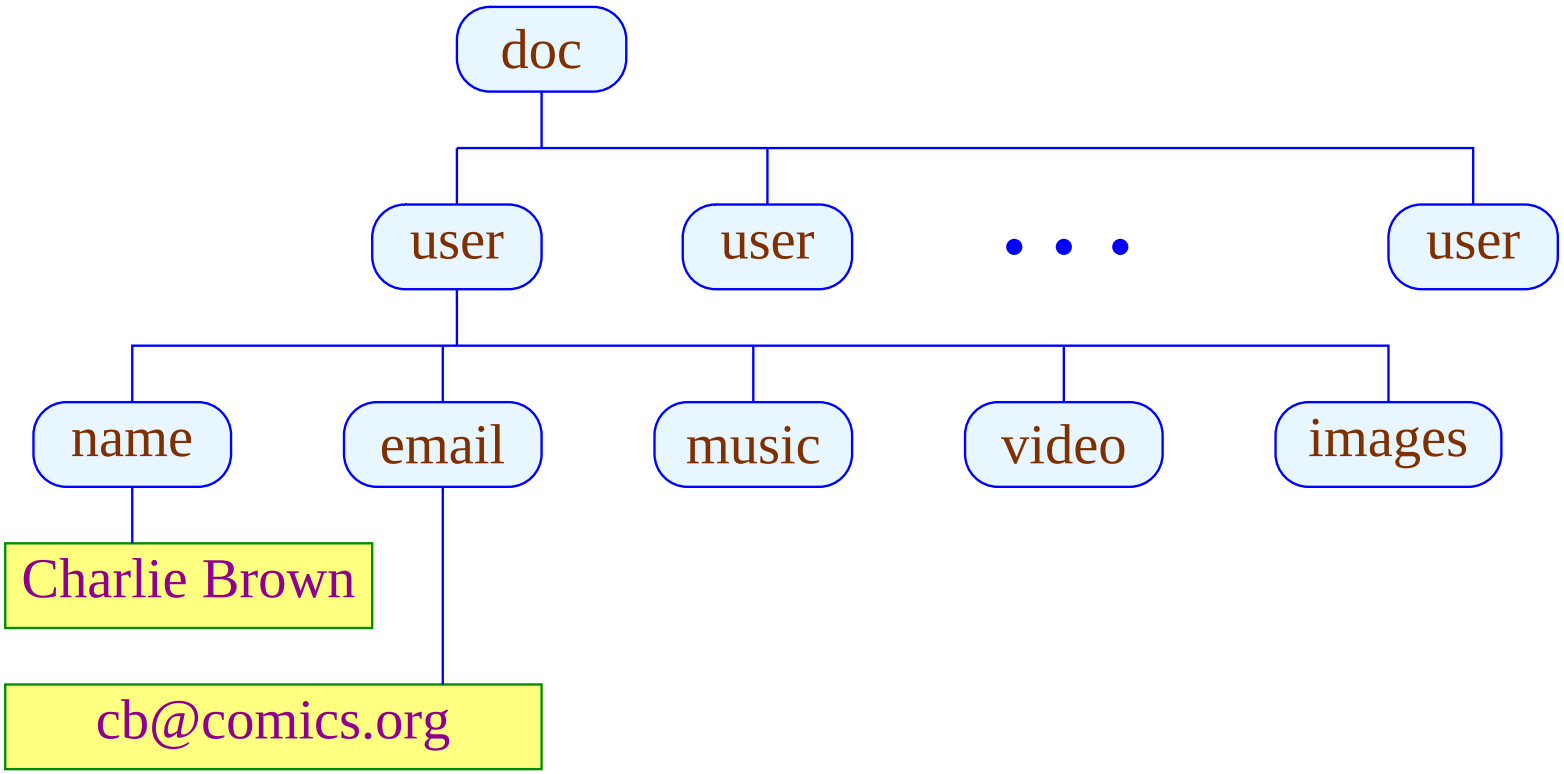
```
<jazz>
 <album> Kind of Blue </album>
 <artist> Miles Davis </artist>
 <year> 1997 </year>
 <price> 22 </price>
 <review> ... </review>
 <review> ... </review>
 <review> ... </review>
</jazz>
```

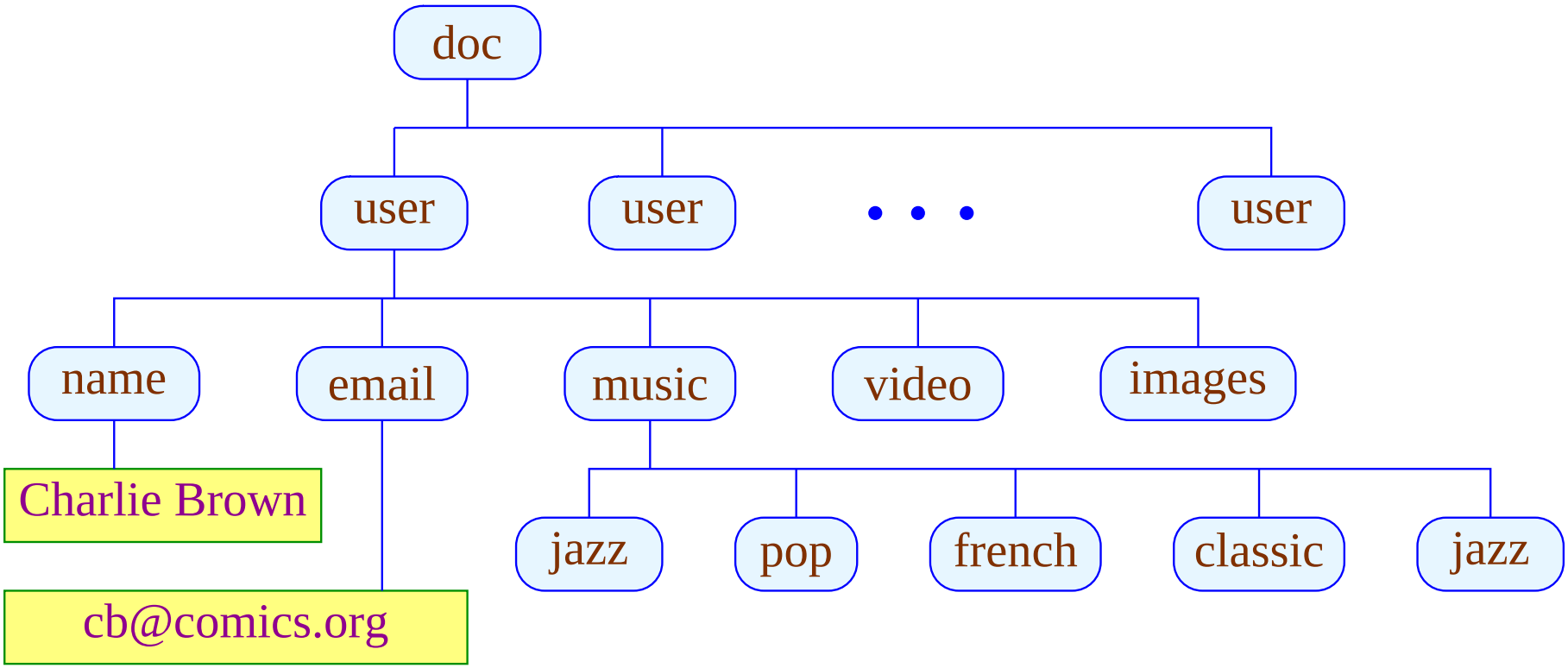
... in essence:

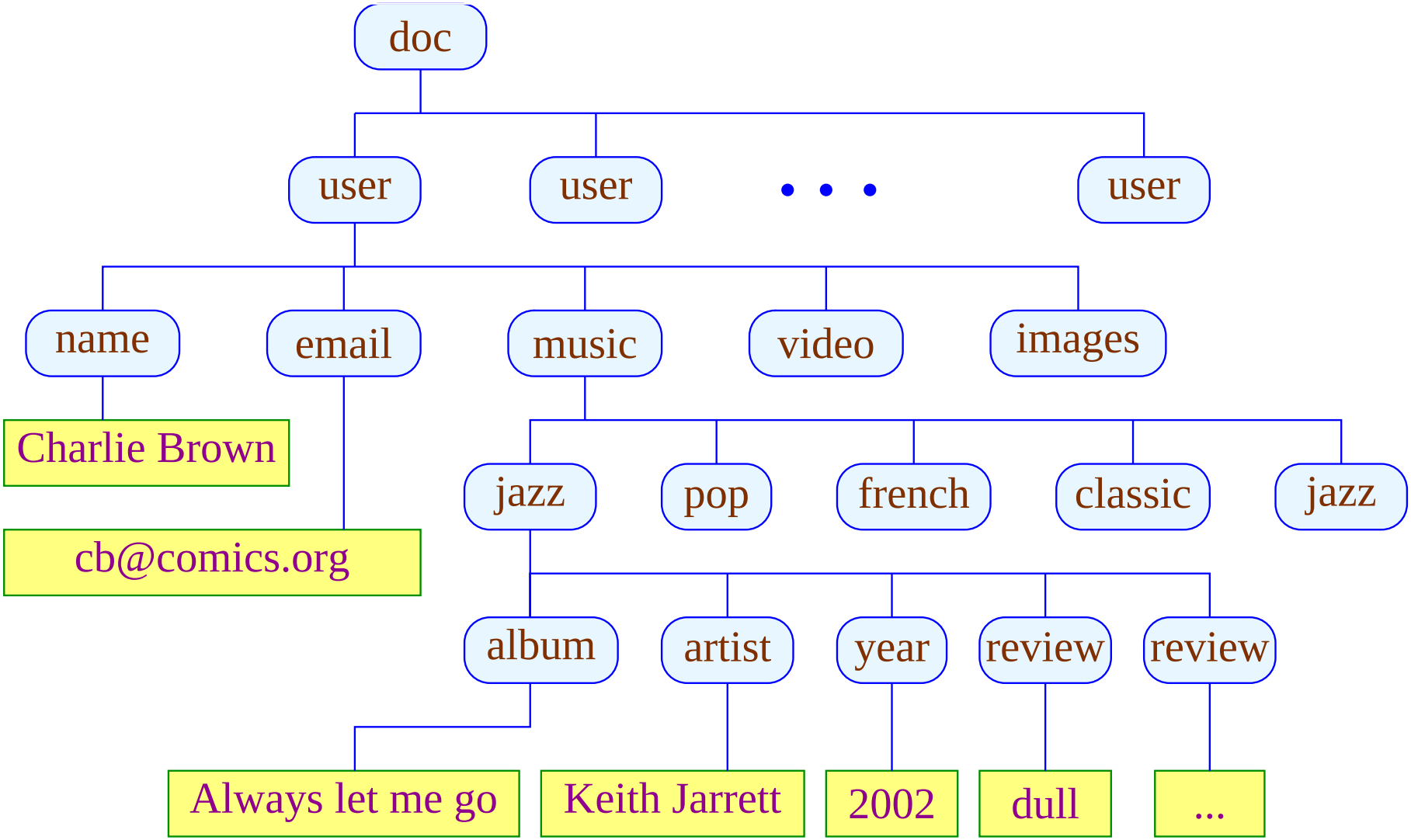
→ hierarchically structured through tags;

→ linear representation of (unranked) tree


doc
user
user
user


doc
user
user
user
name
email
music
video
images
Charlie Brown
cb@comics.org


doc
user
user
user
name
email
music
video
images
Charlie Brown
cb@comics.org
jazz
pop
french
classic
jazz


doc
user
user
user
name
email
music
video
images
Charlie Brown
cb@comics.org
jazz
pop
french
classic
jazz
album
artist
year
review
review
Always let me go
Keith Jarrett
2002
dull
...

Queries

Jazz with more than 2 reviews

Queries

Jazz with more than 2 reviews

$$\text{jazz}\langle \#\text{review} > 2 \rangle$$

Queries

Jazz with more than 2 reviews

$$\text{jazz}\langle \#\text{review} > 2 \rangle$$

Users who like jazz more than pop

Queries

Jazz with more than 2 reviews

$$\text{jazz}\langle \#\text{review} > 2 \rangle$$

Users who like jazz more than pop

$$\text{user} \wedge \mu x.\, (*\langle \_ x \_ \rangle \vee F)$$

Queries

Jazz with more than 2 reviews

$$\text{jazz}\langle \#\text{review} > 2 \rangle$$

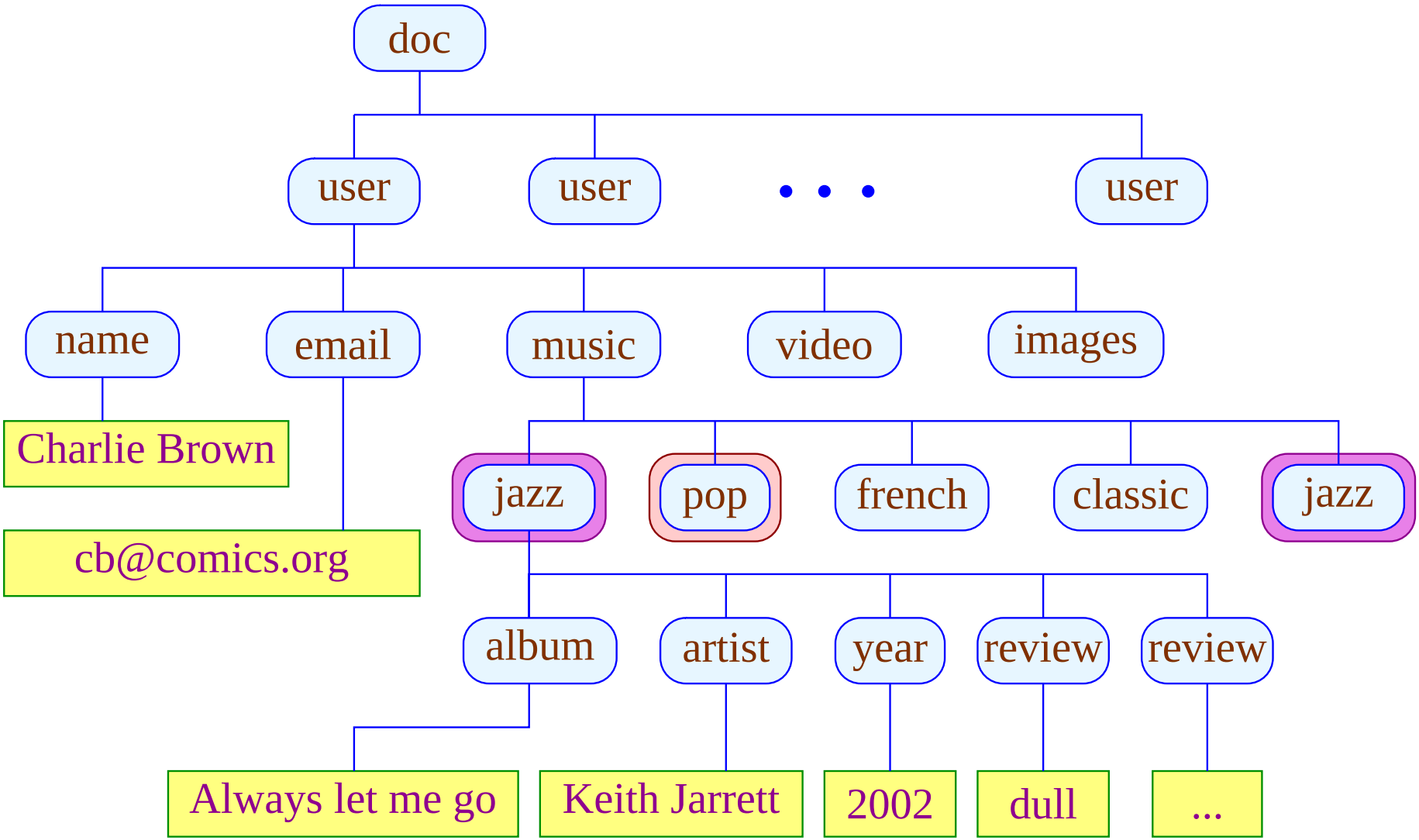
Users who like jazz more than pop

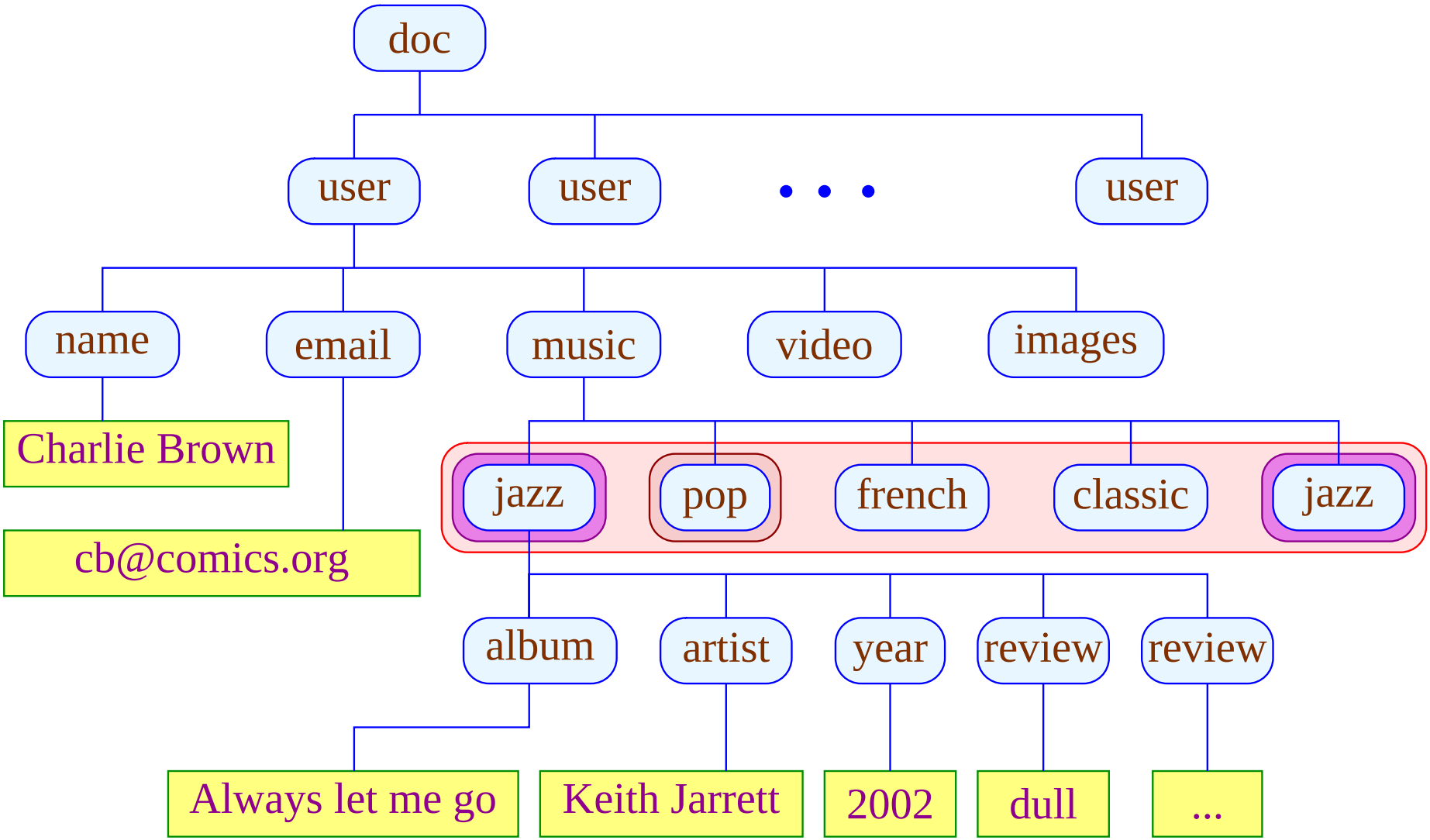
$$\text{user} \wedge \mu x.\, (*\langle \_ x \_ \rangle \vee F)$$

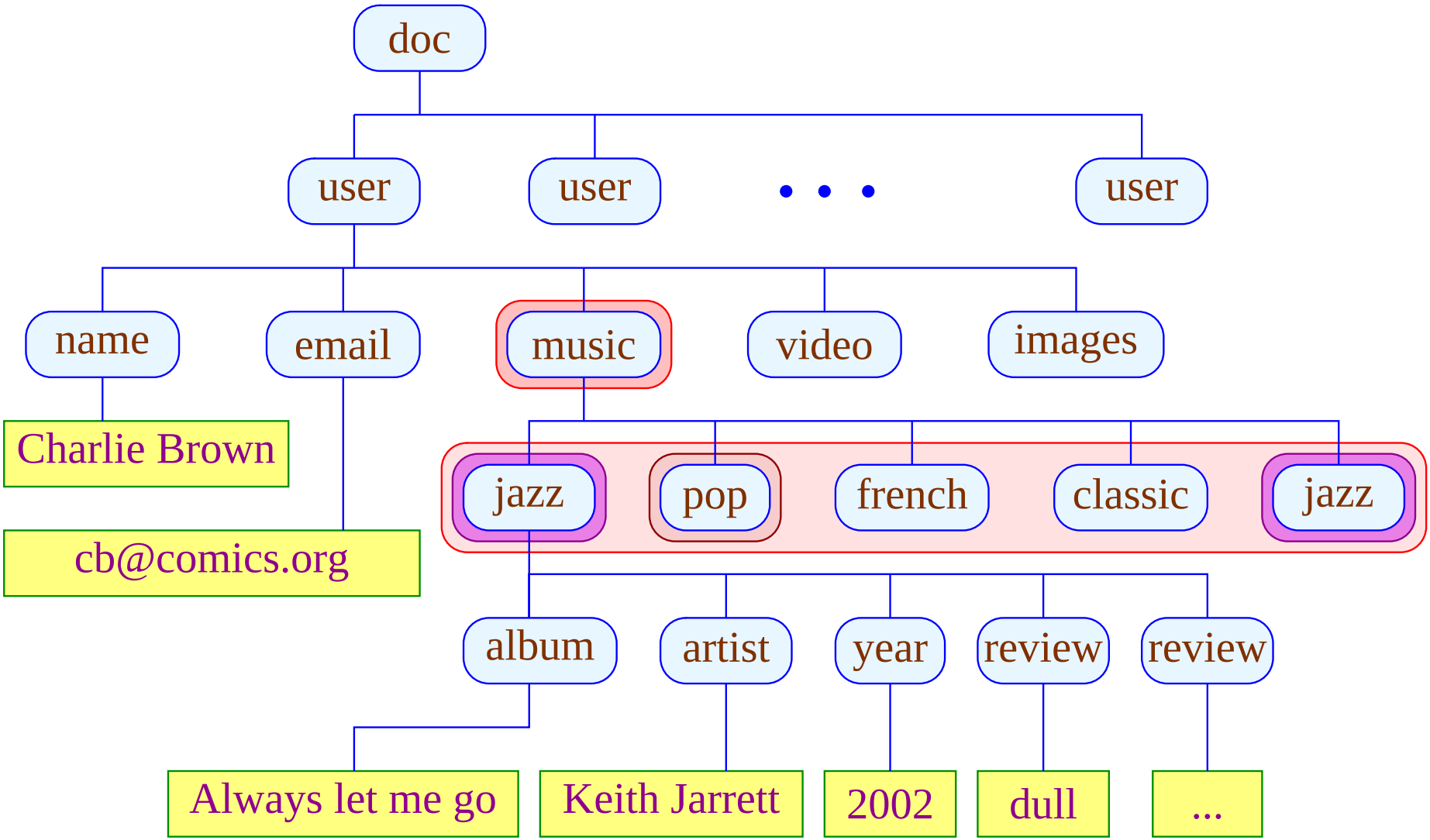
with

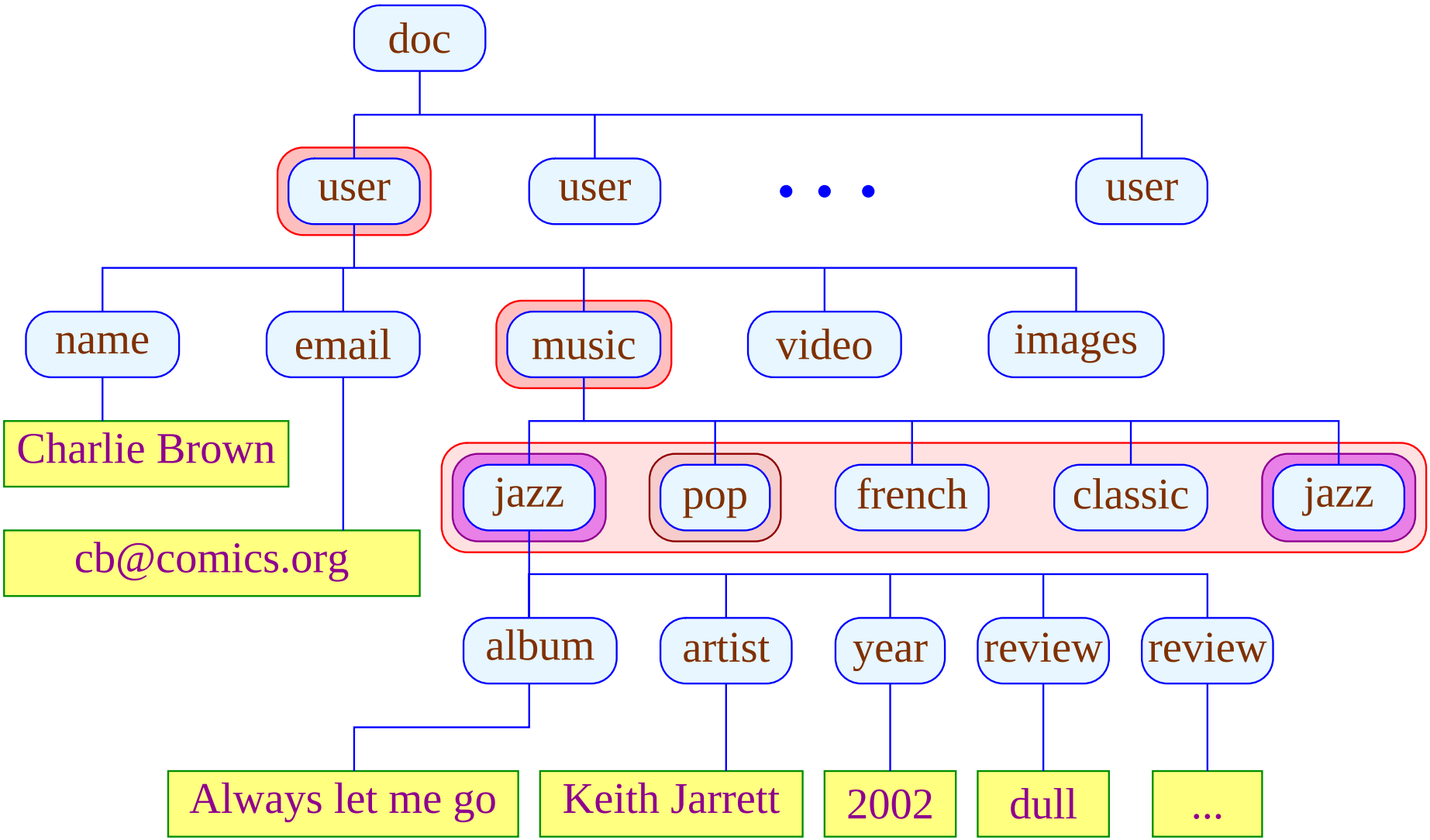
$$F = *\langle \#\text{jazz} > \#\text{pop} \rangle$$

(nodes with more jazz-children than pop-children)


doc
user
user
user
name
email
music
video
images
Charlie Brown
cb@comics.org
jazz
pop
french
classic
jazz
album
artist
year
review
review
Always let me go
Keith Jarrett
2002
dull
...


doc
user
user
user
name
email
music
video
images
Charlie Brown
cb@comics.org
jazz
pop
french
classic
jazz
album
artist
year
review
review
Always let me go
Keith Jarrett
2002
dull
...


doc
user
user
user
name
email
music
video
images
Charlie Brown
cb@comics.org
jazz
pop
french
classic
jazz
album
artist
year
review
review
Always let me go
Keith Jarrett
2002
dull
...


doc
user
user
user
name
email
music
video
images
Charlie Brown
cb@comics.org
jazz
pop
french
classic
jazz
album
artist
year
review
review
Always let me go
Keith Jarrett
2002
dull
...

Goal:

- → Extend query logics by numerical constraints on the children of nodes
- → Get implementation by extended tree automata

Fixpoint formulas:

$$\phi \quad ::= \quad \mathsf{tt} \mid \mathsf{ff} \mid x \mid \mu\, x.\, \phi \mid \phi \vee \phi \mid \phi \wedge \phi \mid$$

Fixpoint formulas:

$$\begin{array}{rcl} \phi & ::= & \mathsf{tt} \mid \mathsf{ff} \mid x \mid \mu\, x.\, \phi \mid \phi \vee \phi \mid \phi \wedge \phi \mid \\ & & a\langle F \rangle \mid\ * \langle F \rangle \end{array}$$

Fixpoint formulas:

$$\begin{array}{rcl} \phi & ::= & \mathsf{tt} \mid \mathsf{ff} \mid x \mid \mu\, x.\, \phi \mid \phi \vee \phi \mid \phi \wedge \phi \mid \\ & & a\langle F \rangle \mid *\langle F \rangle \end{array}$$

Precondition F: Boolean combination of regular expressions and Presburger constraints over formulas satisfied by children of node labeled a

Theorem

- $\mathcal{L} = \{t \mid t \models \phi\}$ for a Presburger fixpoint formula

  iff

  $\mathcal{L} = \mathcal{L}(A)$ for a deterministic Presburger automaton.
- Presburger fixpoint satisfiability is decidable in ExpTime.

Related Work:

Querying Ordered Documents:

XPath 2.0	W3C 2002
MSO	Schwentick et al. 2002

Related Work:

Querying Ordered Documents:

XPath 2.0 — W3C 2002

MSO — Schwentick et al. 2002

Querying Unordered Documents:

Ambient Logic — Cardelli, Ghelli et al. 2000ff

Related Work:

Querying Ordered Documents:

XPath 2.0	W3C 2002
MSO	Schwentick et al. 2002

Querying Unordered Documents:

Ambient Logic	Cardelli, Ghelli et al. 2000ff

Numerical Reasoning:

graded μ-calculus	Kupferman, Sattler, Vardi 2002
Sheaves Automata	Lugiez, Dal Zilio 2002
MSO + Presburger	Seidl, Schwentick, M. 2003

Presburger Arithmetic

$$\exists\, y.\ (5y \leq 3x \wedge 3x - 5y \leq 2)$$

Presburger Arithmetic

$$\exists\, y\ (5y \leq 3x \wedge 3x - 5y \leq 2)$$

Formally,

$$\begin{aligned} \phi \quad ::= \quad & x_1 + x_2 = x_3 \mid x = n \\ & \mid \phi_1 \vee \phi_2 \mid \neg\phi \\ & \mid \exists\, x.\ \phi \end{aligned}$$

Satisfying Variable Assignments:

$$\{x \mapsto 2\} \quad \models \quad \exists\, y\ (5y \leq 3x \wedge 3x - 5y \leq 2)$$

Satisfying Variable Assignments:

$$\{x \mapsto 2\} \quad \models \quad \exists\, y\ (5y \leq 3x \wedge 3x - 5y \leq 2)$$

$\rightarrow$ ... form semi-linear sets; (= union of $\bar{c} + \sum_i x_i \bar{p}_i$)

$\rightarrow$ ... which can be decided for emptiness / membership

$\rightarrow$ ... by constructing a DFA A_ϕ

Satisfying Variable Assignments:

$$\{x \mapsto 2\} \quad \models \quad \exists\, y\ (5y \leq 3x \wedge 3x - 5y \leq 2)$$

$\rightarrow$ ... form semi-linear sets (= union of $\bar{c} + \sum_i x_i \bar{p}_i$)

$\rightarrow$ ... which can be decided for emptiness / membership

$\rightarrow$ ... by constructing a DFA A_ϕ
(triple exponential, Klaedtke LICS 2004)

Presburger Tree Automata

Idea [S. et al. 2003 + Lugiez, Dal Zilio 2002]:

Allow as preconditions Boolean combinations of regular expressions and Presburger formulas.

Example: $q\ (q+p)^* \quad \wedge \quad \#p \geq 2(\#q) \quad \longrightarrow_a \quad s$

"Go to state s if node label is a, all children are in state p or q and there are at least twice as many in p, and the first child is in state q"

[Seidl et al. 2003]

For non-deterministic Presburger Tree Automata,

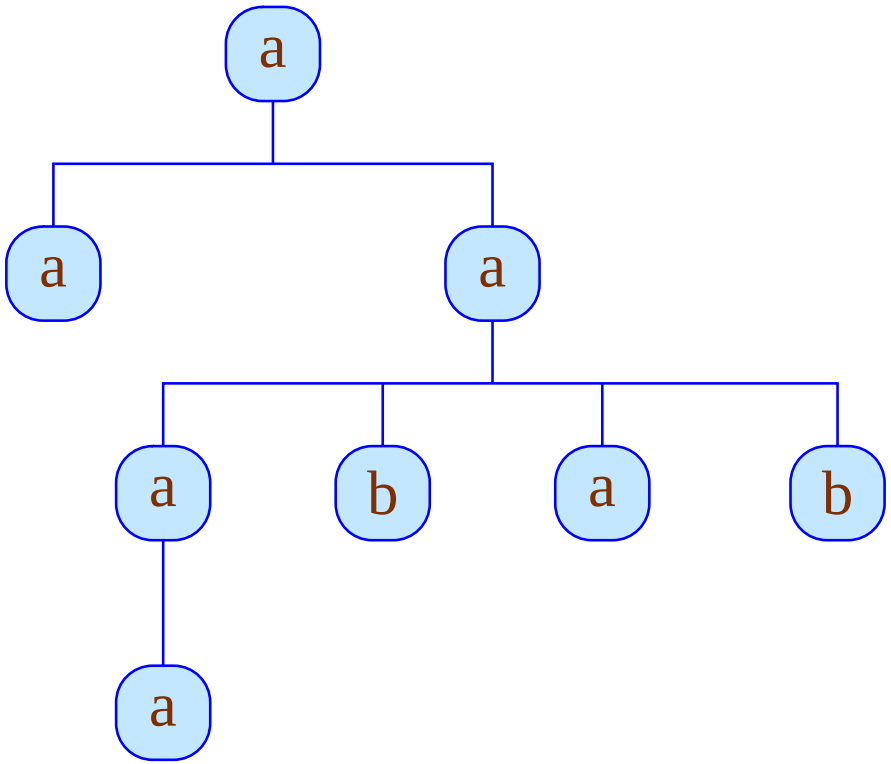
- Emptiness / membership are decidable;
- Universality is undecidable;
- Equivalence with Presburger EMSO (father/sibling + Presburger constraints on children)

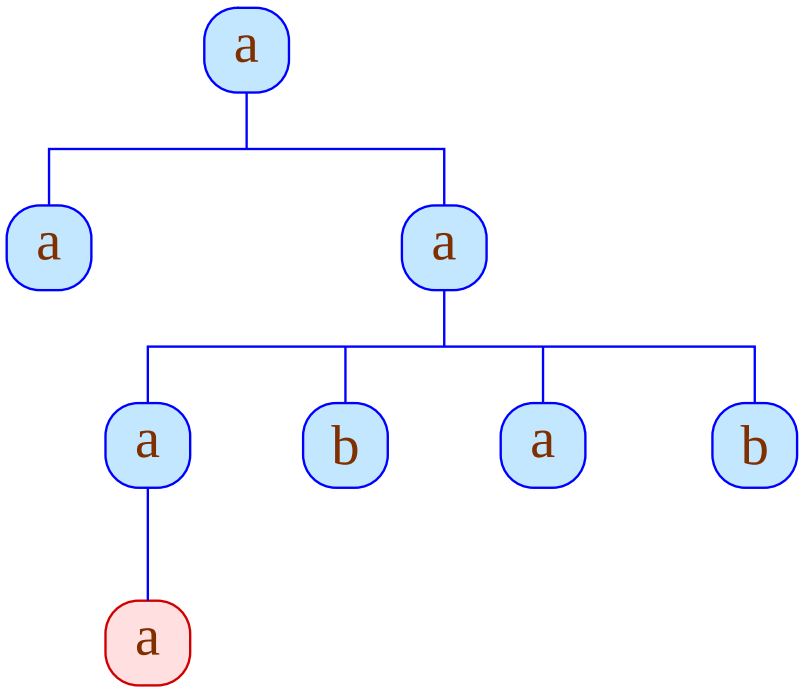
Deterministic Presburger Tree Automata

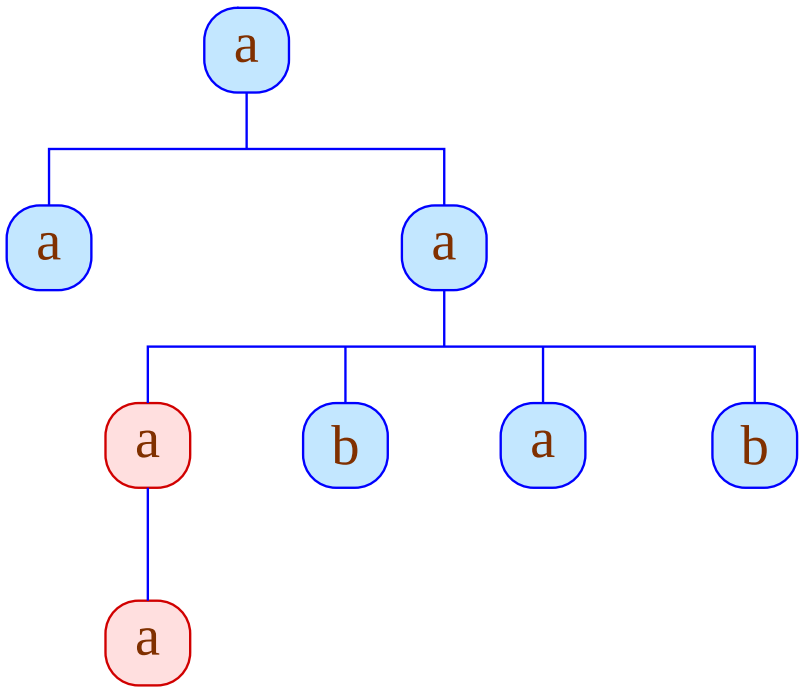
Example:

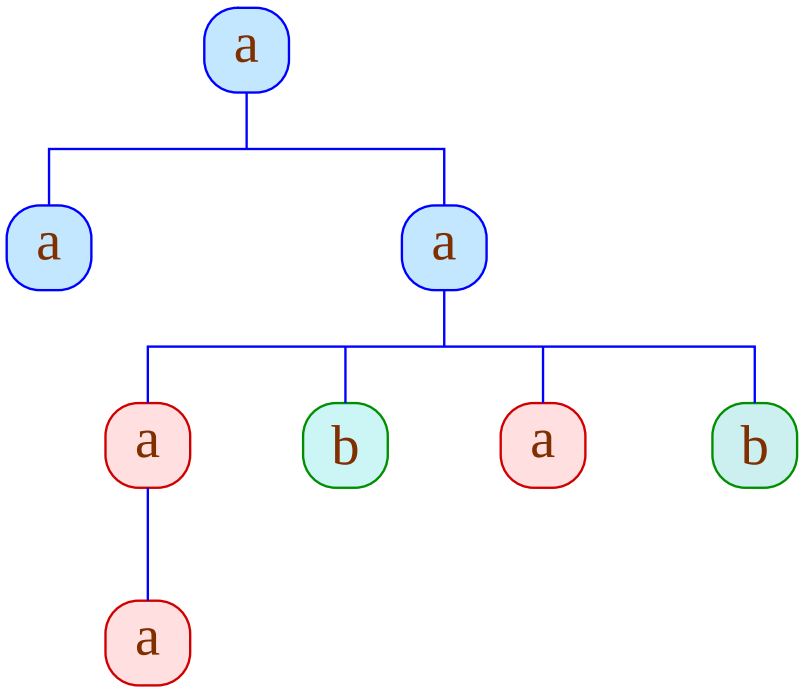
Set of all trees over $\{a, b\}$ such that:

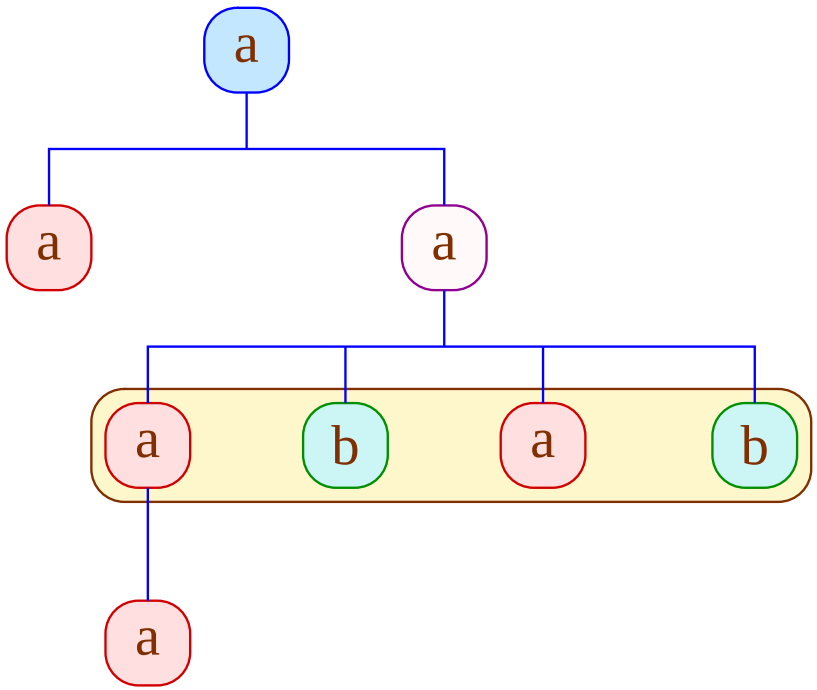
- Every internal node is labeled with a.
- The leftmost leaf is always labeled with a.
- The number of children with b-leaves is $\leq$ than the number of those with.

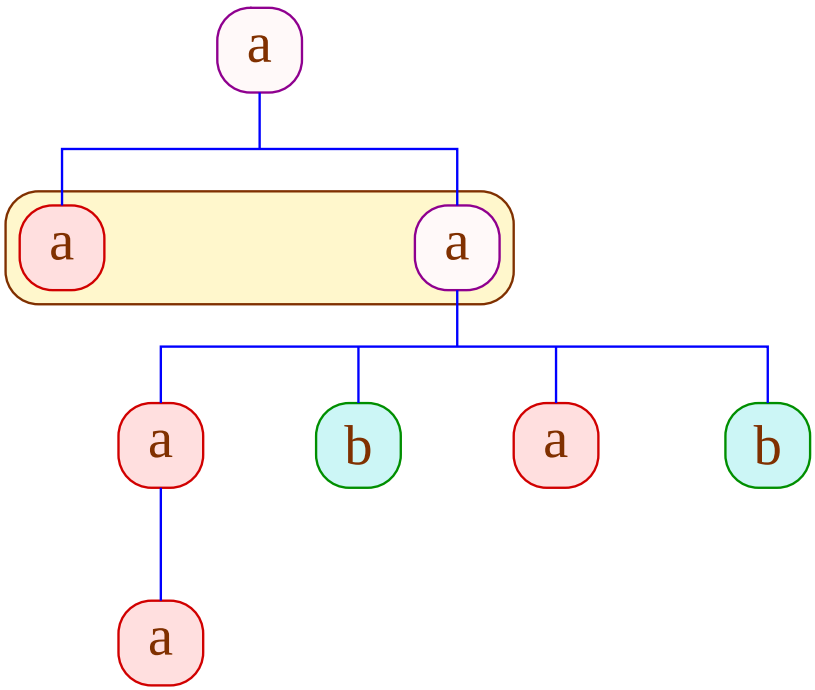

a
a
a
a
b
a
b
a


a
a
a
a
b
a
b
a


a
a
a
a
b
a
b
a


a
a
a
a
b
a
b
a


a
a
a
a
b
a
b
a


a
a
a
a
b
a
b
a

Complexity...

Use quantifier-free Presburger formulas...

Theorem

Emptiness for PTA is PSPACE-complete.

Proof: Determine reachable states bottom-up.

Complexity...

Transition precondition $\bigwedge_i R_i \wedge \bigwedge_j \overline{R'_j} \wedge \phi$: test emptiness

Complexity...

Transition precondition $\bigwedge_i R_i \wedge \bigwedge_j \overline{R'_j} \wedge \phi$: test emptiness

- quantifier-free Presburger formulas... boolean expressions over equations with integer coefficients

Complexity...

Transition precondition $\bigwedge_i R_i \wedge \bigwedge_j \overline{R'_j} \wedge \phi$: test emptiness

- quantifier-free Presburger formulas... boolean expressions over equations with integer coefficients
- Parikh image of an NFA of size n over alphabet of size k is a semi-linear set over vectors from $\{1, \ldots, n\}^k$.

Complexity...

Parikh image of $\bigwedge_i R_i \wedge \bigwedge_j \overline{R'_j}$:

Complexity...

Parikh image of $\bigwedge_i R_i \wedge \bigwedge_j \overline{R'_j}$:

exponentially many vectors (only :-), ...

Complexity...

Parikh image of $\bigwedge_i R_i \wedge \bigwedge_j \overline{R'_j}$:

exponentially many vectors (only :-), ...

but vector dimension is polynomial

Complexity...

Parikh image of $\bigwedge_i R_i \wedge \bigwedge_j \overline{R'_j}$:

exponentially many vectors (only :-), ...

but vector dimension is polynomial

[Papadimitriou'81] Size of minimal solution of diophantine system is only exponential in the number of equations

Complexity...

PTA Emptiness: Special case with better complexity

Presburger constraints $\bigvee_i (R_i \wedge \Phi_i)$, with Φ_i conjunction of equations

Complexity...

PTA Emptiness: Special case with better complexity

Presburger constraints $\bigvee_i (R_i \wedge \Phi_i)$, with Φ_i conjunction of equations

Theorem

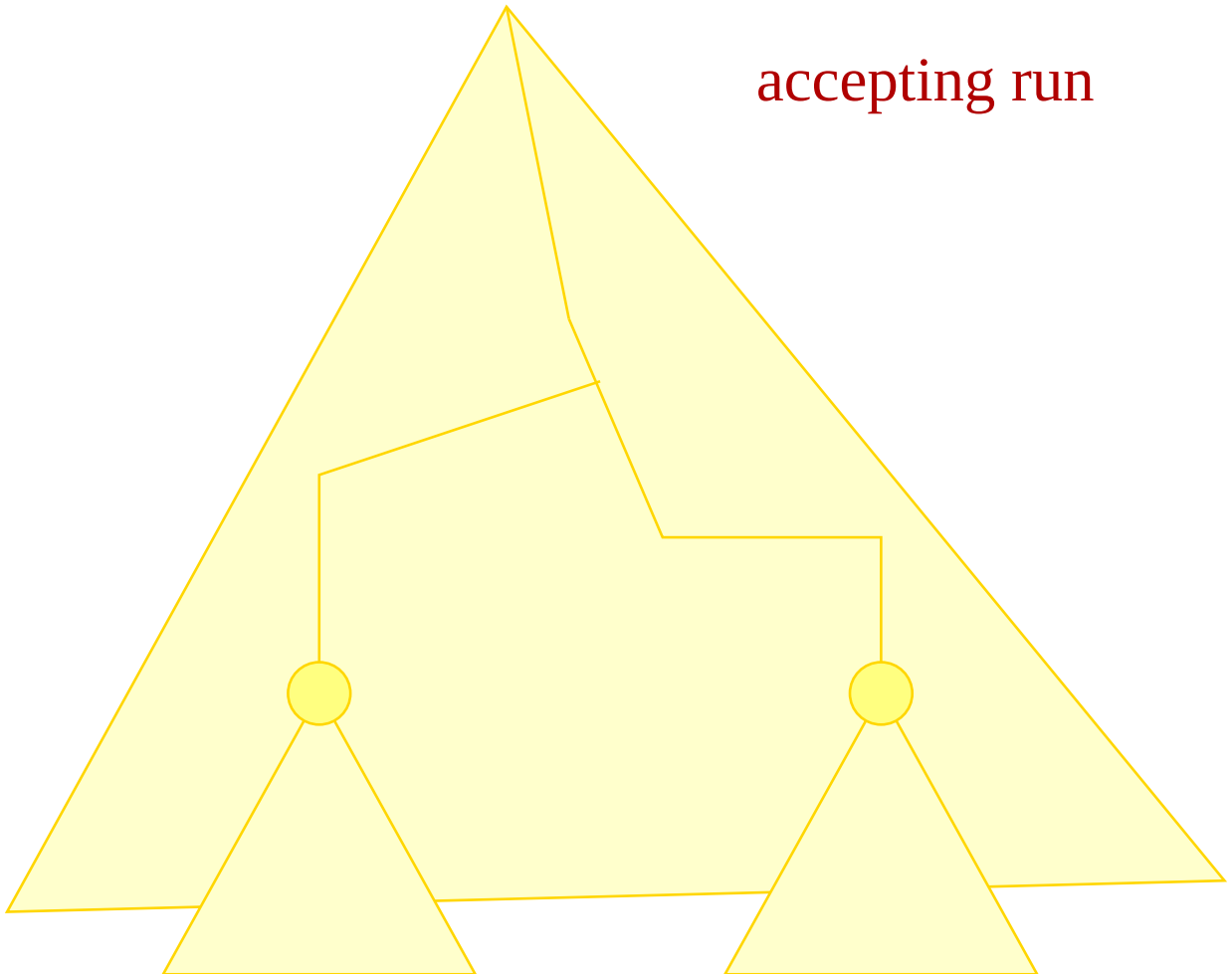
Membership $t \in \mathcal{L}(A)$ can be checked in time $O(|t||A|)$.

Presburger Fixpoint Query:

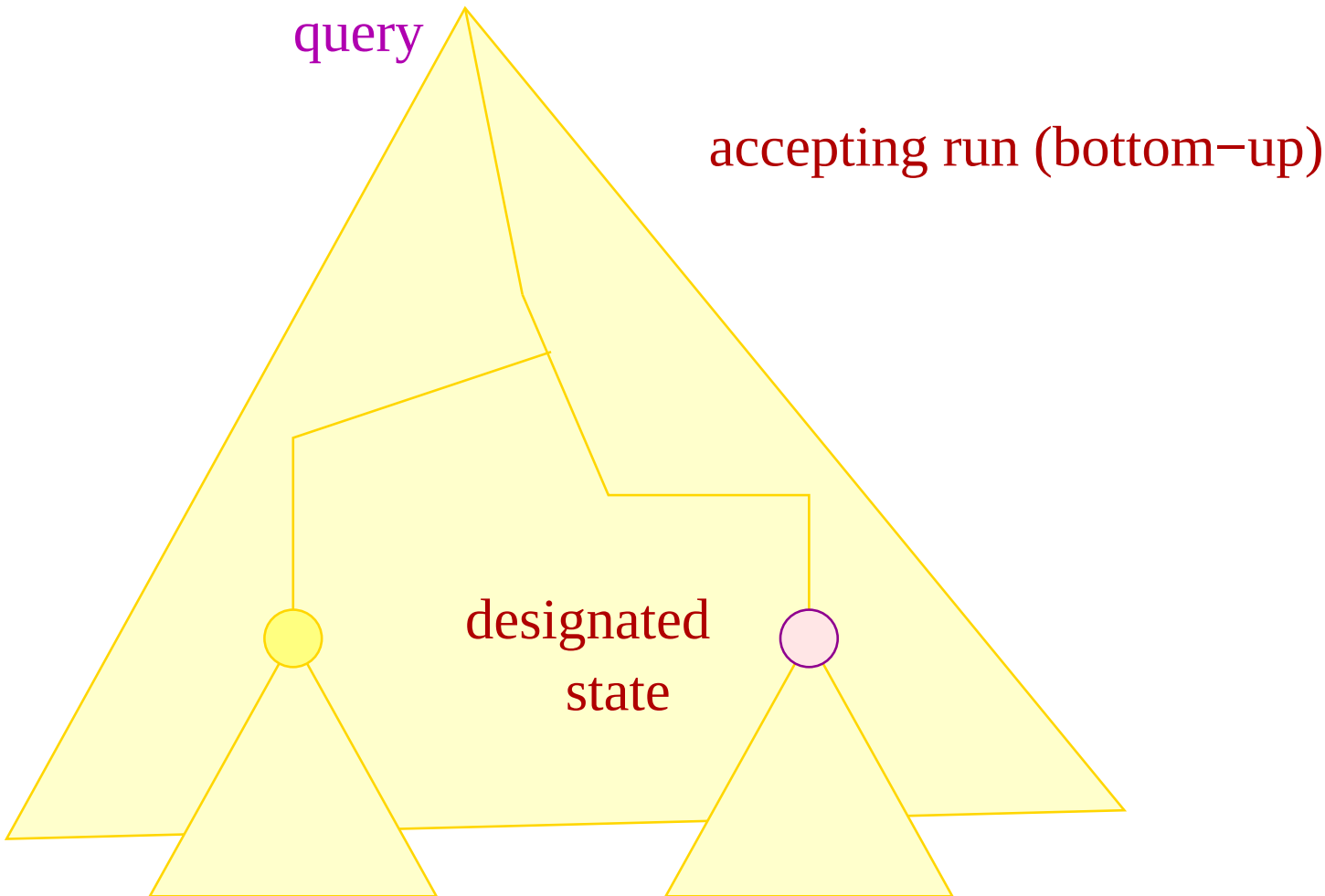
$\bullet \wedge \text{user} \wedge \mu x.\, (*\langle \_ x \_ \rangle \vee \text{jazz}\langle \#\text{review} > 2 \rangle)$

marker $\bullet$ for requested subdocument

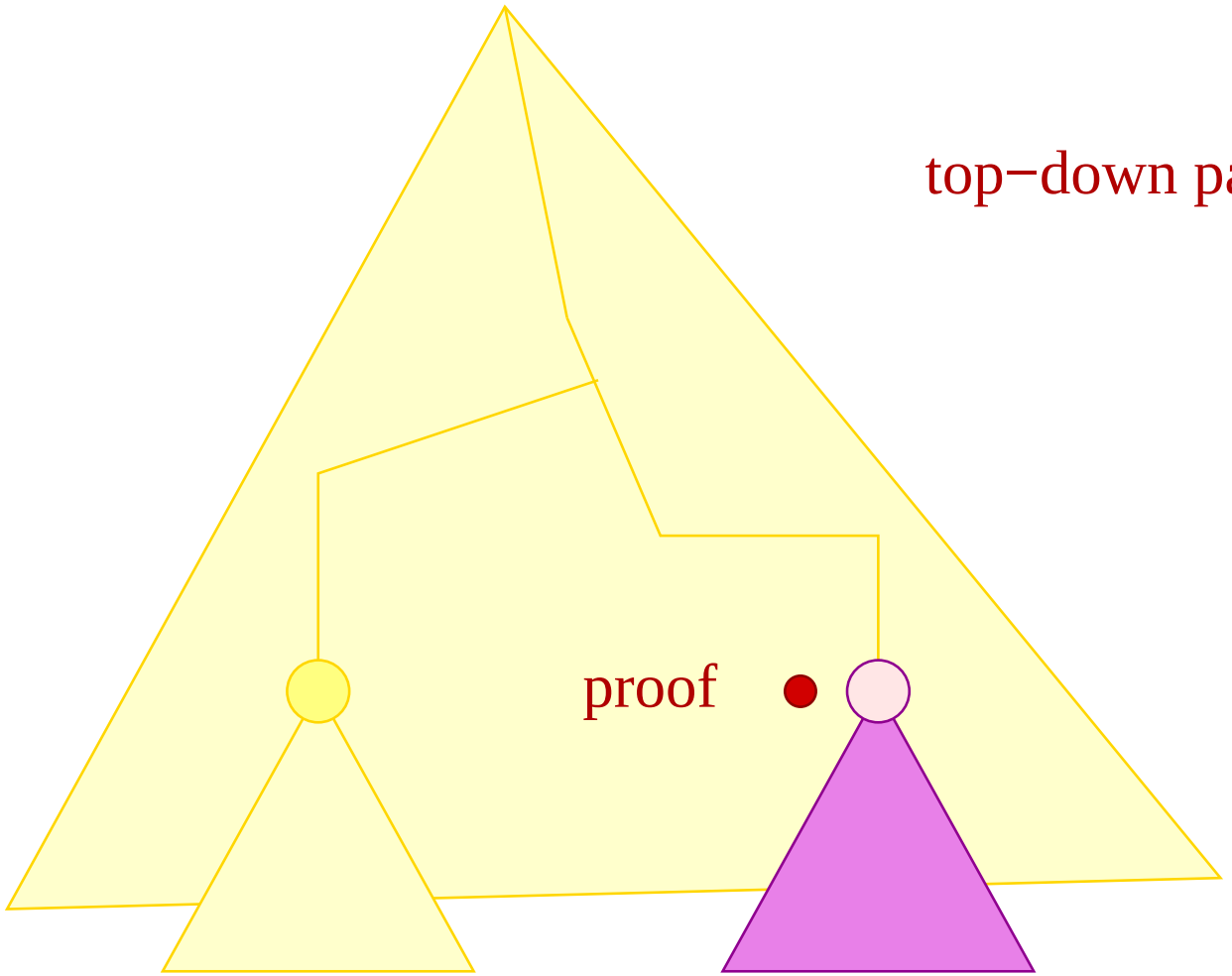
Automata Query:



Automata Query:



Automata Query:



Theorem

A Presburger fixpoint query ϕ over a tree t can be answered in time $O(|t||\phi|^2)$.

Conclusion

→ Linear-time algorithm for computing the Parikh-image of an NFA.

Conclusion

- → Linear-time algorithm for computing the Parikh-image of an NFA.
- → The technique can be generalized to reason about documents containing explicit natural numbers.

Conclusion

→ Linear-time algorithm for computing the Parikh-image of an NFA.

→ The technique can be generalized to reason about documents containing explicit natural numbers.

→ Identify practical formalism for Presburger queries which supports:

- intuitive specifications;
- efficient compilation into automata.

Parikh images of regular sets

Theorem For an NFA A (or regular expression) an existential Presburger formula of linear size can be constructed for $\mathcal{L}(A)$.

Parikh images of regular sets

Theorem For an NFA A (or regular expression) an existential Presburger formula of linear size can be constructed for $\mathcal{L}(A)$.

Proof.

Variables $x_{p,a,q}$ for each transition (p, a, q)

Formula describing connected flow = Parikh image of $w \in \mathcal{L}(A)$