XML Data Exchange

Marcelo Arenas
P. Universidad Católica de Chile

Joint work with Leonid Libkin (U. of Toronto)

Data Exchange in Relational Databases

- Data exchange has been extensively studied in the relational world.
  - It has also been implemented: Clio.

- Relational data exchange settings:
  - Source and target schemas: Relational schemas.
  - Relationship between source and target schemas: Source-to-target dependencies.

- Semantics of data exchange has been precisely defined.
  - Algorithms for materializing target instances and for answering queries over the target have been developed.

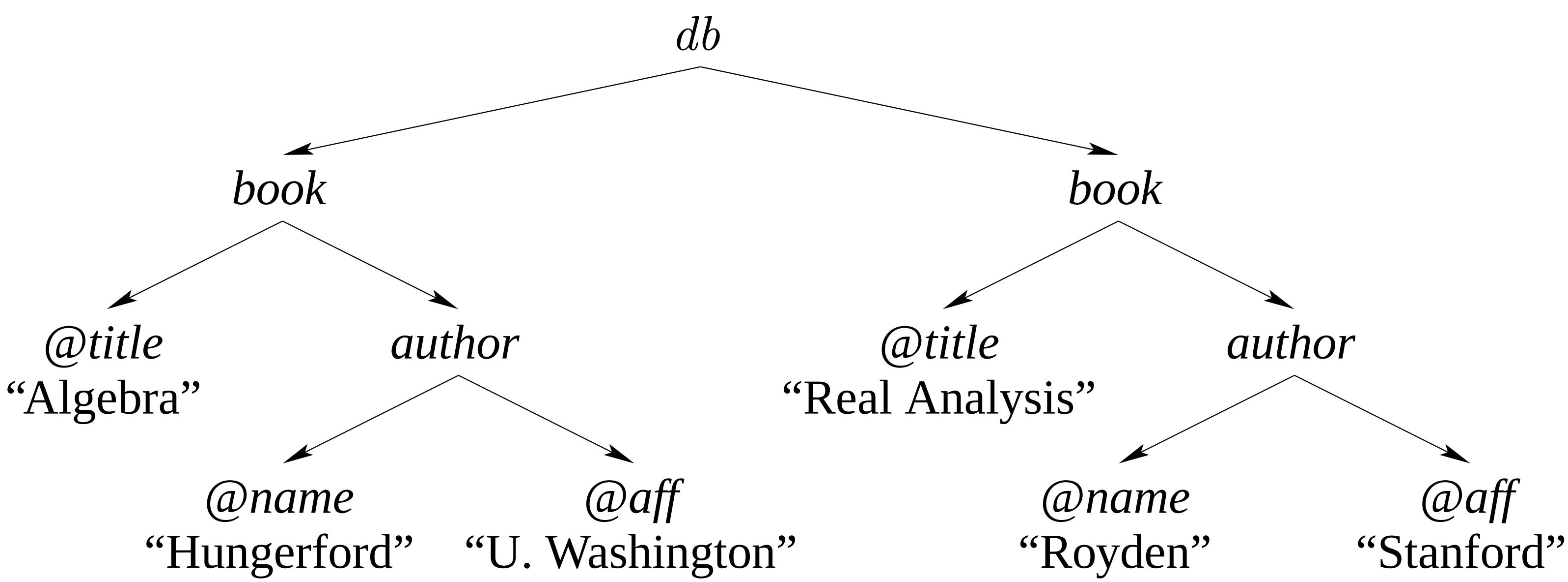
Outline

- XML data exchange settings.
  - XML source-to-target dependencies.
- Query answering in XML data exchange.
- Final remarks.

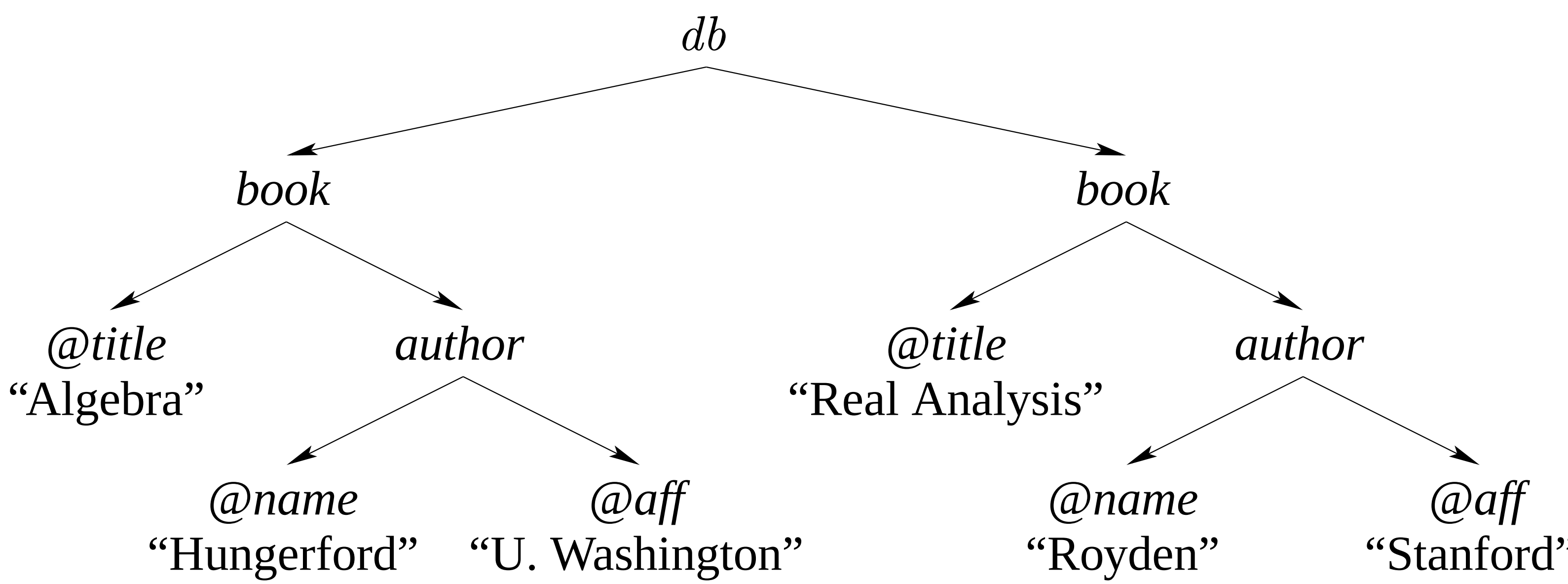
Outline

- XML data exchange settings.
  - XML source-to-target dependencies.
- Query answering in XML data exchange.
- Final remarks.

XML Documents



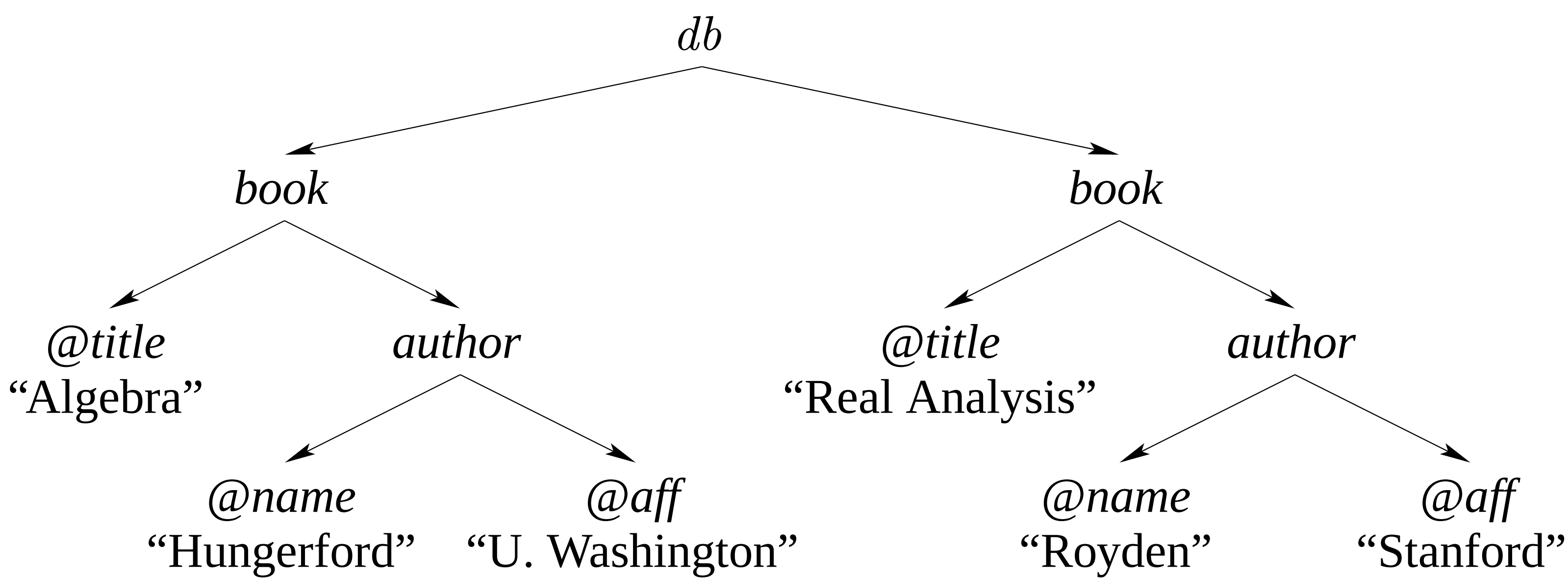
XML Documents



DTD :

$$\begin{array}{rcl} db & \longrightarrow & book^{+} \\ book & \longrightarrow & author^{+} \\ author & \longrightarrow & \varepsilon \end{array}$$

XML Documents



DTD :

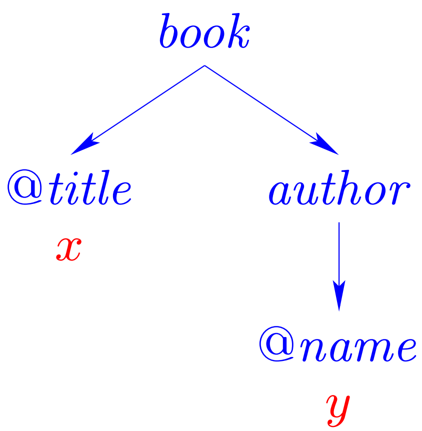
$$\begin{array}{rcl} db & \rightarrow & book^{+} \\ book & \rightarrow & author^{+} \\ author & \rightarrow & \varepsilon \end{array} \qquad \begin{array}{rcl} book & \rightarrow & @title \\ author & \rightarrow & @name, @aff \end{array}$$

XML Data Exchange Settings

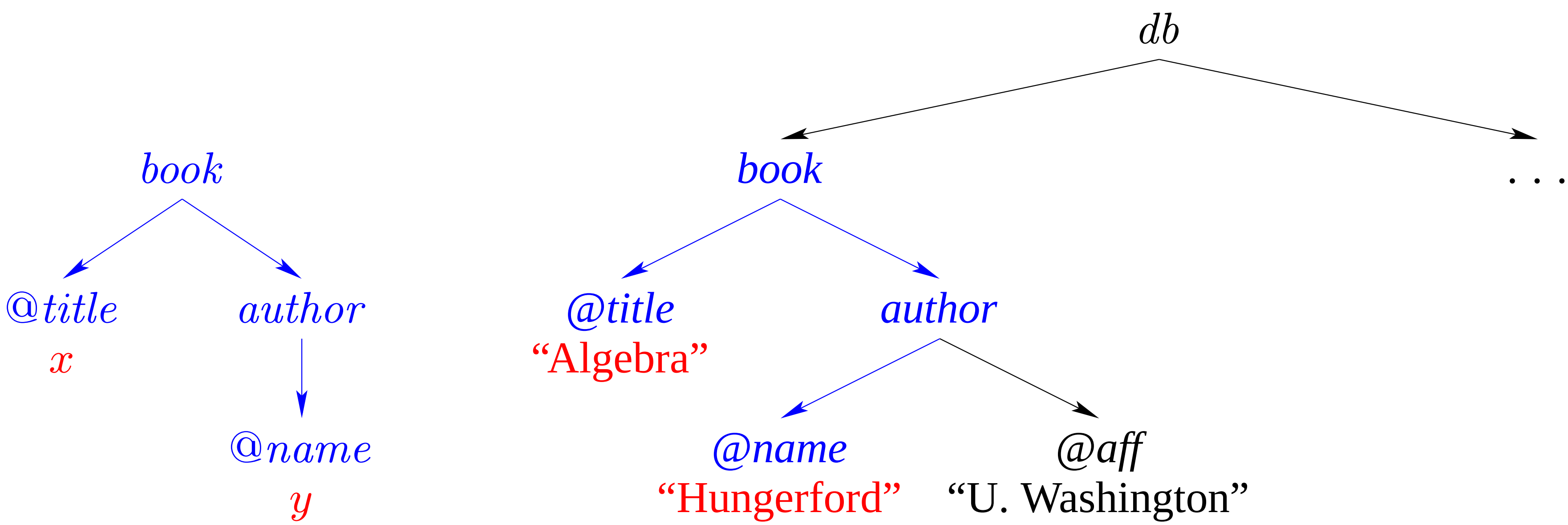
- Source and target schemas are given by DTDs.
- To specify the relationship between the source and the target schemas we use source-to-target dependencies.

  To define these dependencies, we use tree patterns ...

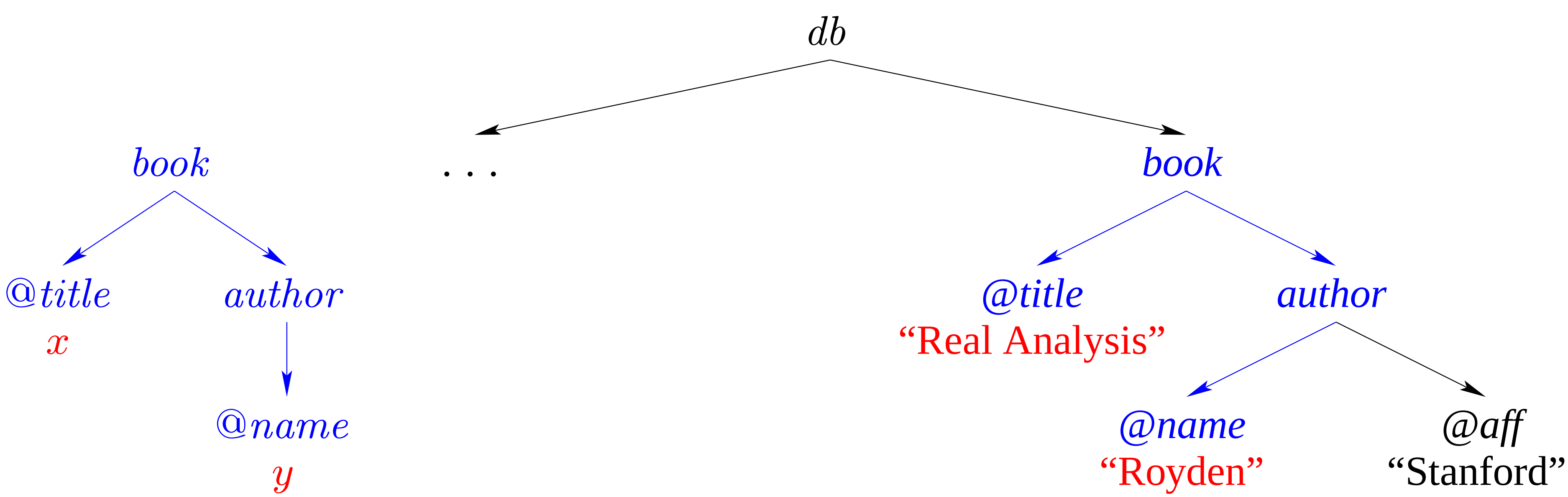
Tree Patterns: Example



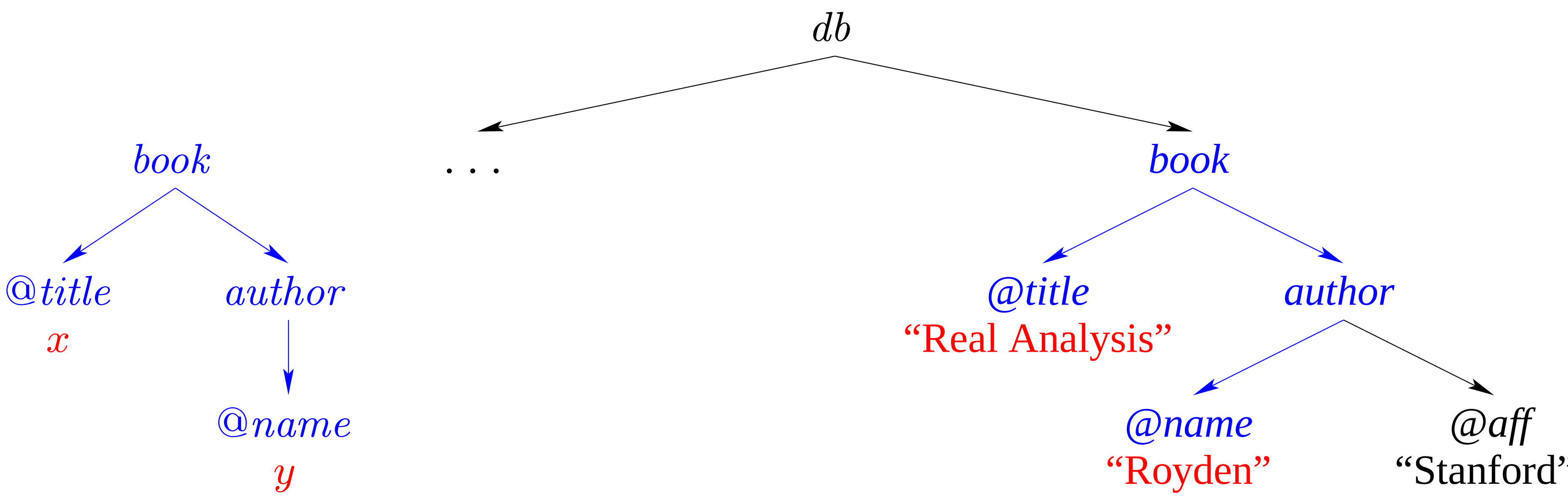
Tree Patterns: Example



Tree Patterns: Example



Tree Patterns: Example



Collect tuples (x, y): (Algebra, Hungerford), (Real Analysis, Royden)

Tree Patterns

- Tree patterns: XPath-like language.
  - Example: $book(@title = x)[author(@name = y)]$

- Language also includes wildcard $\_$ (matching more than one symbol) and descendant operator $//$.

XML Source-to-target Dependencies

Source-to-target dependency (STD):

$$\psi_{\mathbf{T}}(\bar{x}, \bar{z}) \coloneq \varphi_{\mathbf{S}}(\bar{x}, \bar{y}),$$

where $\varphi_{\mathbf{S}}(\bar{x}, \bar{y})$ and $\psi_{\mathbf{T}}(\bar{x}, \bar{z})$ are tree-pattern formulas over the source and target DTDs, resp.

XML Data Exchange Settings

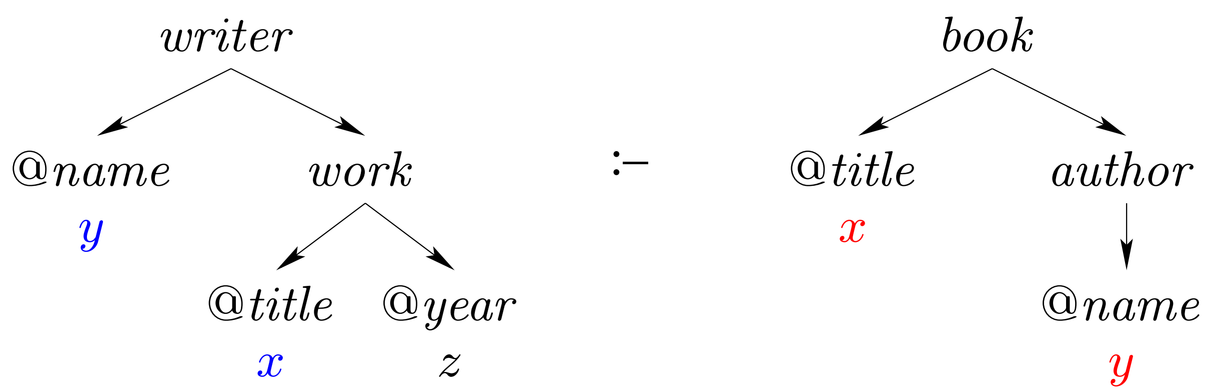
Source DTD:

$$\begin{array}{rcl@{\qquad}rcl} db & \to & book^{+} & & & \\ book & \to & author^{+} & book & \to & @title \\ author & \to & \varepsilon & author & \to & @name, @aff \end{array}$$

Target DTD:

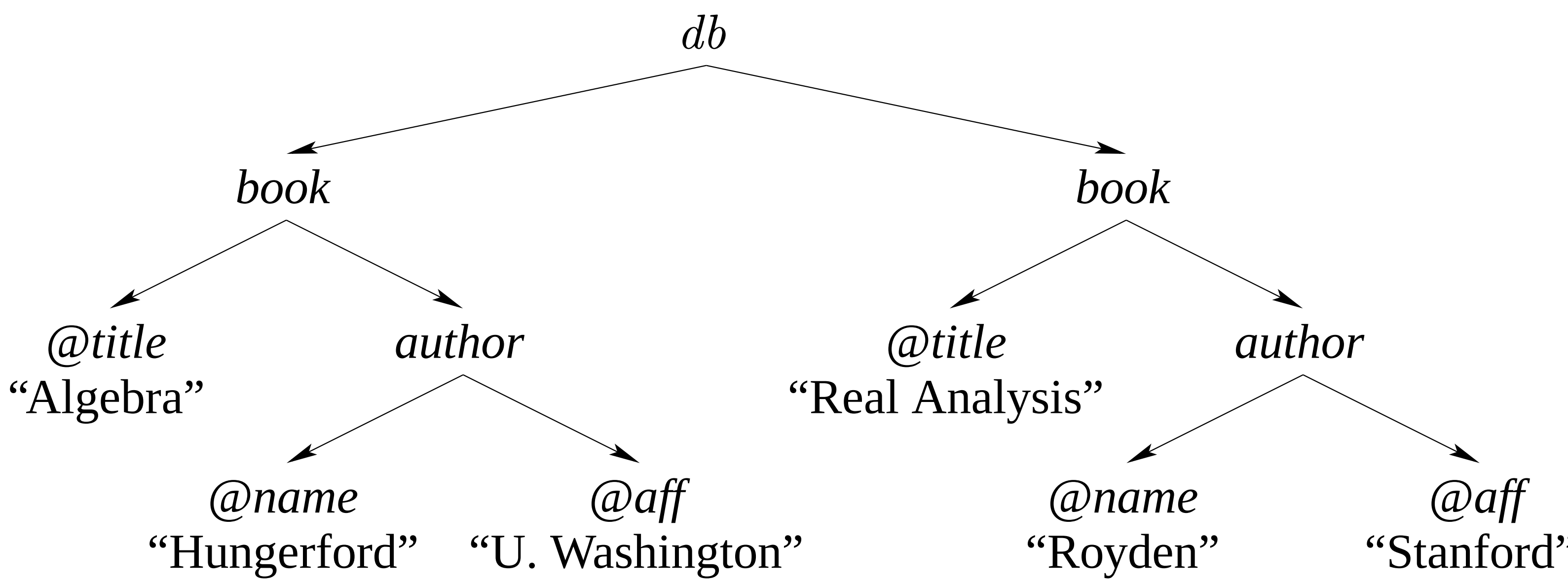
$$\begin{array}{rcl@{\qquad}rcl} bib & \to & writer^{+} & & & \\ writer & \to & work^{+} & writer & \to & @name \\ work & \to & \varepsilon & work & \to & @title, @year \end{array}$$

$\Sigma_{\mathbf{ST}}$:



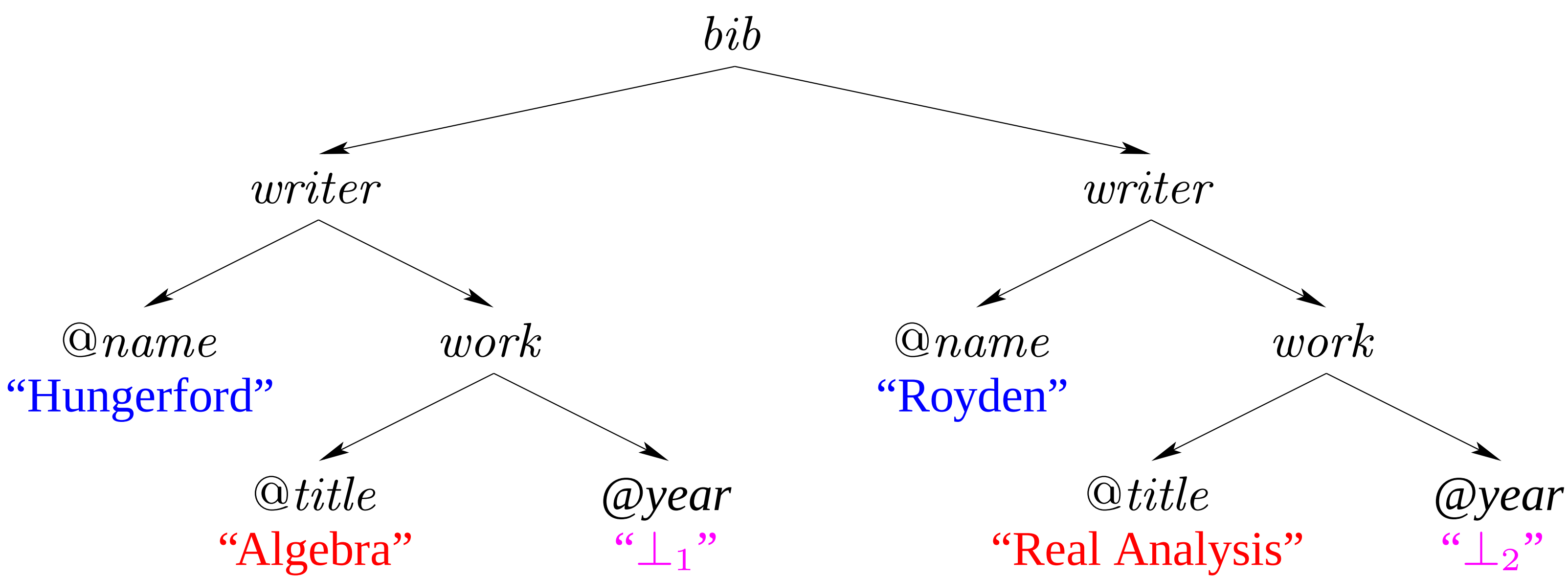
Example: Finding Solutions

Let T be our original tree:



Example: Finding Solutions

A solution for T:



Outline

- XML data exchange settings.
  - XML source-to-target dependencies.
- Query answering in XML data exchange.
- Final remarks.

Query Answering in XML Data Exchange

- Decision to make: What is our query language?

- We start by considering a query language that produces tuples of values.

Conjunctive Tree Queries

- Query language $\mathcal{CTQ}^{//}$ is defined by

$$Q \quad := \quad \varphi \ \mid \ Q \wedge Q \ \mid \ \exists x \, Q,$$

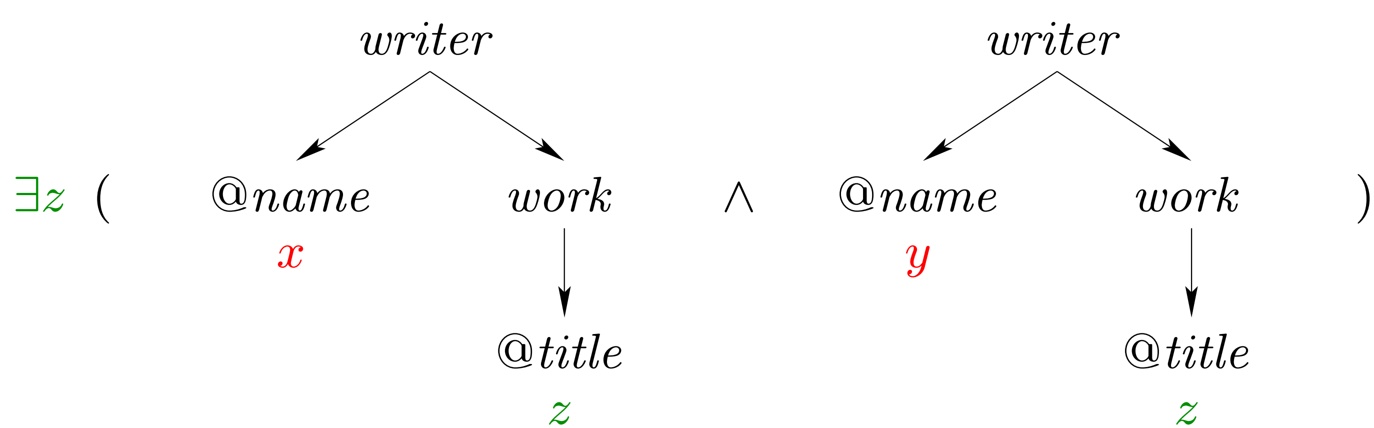
where φ ranges over tree-pattern formulas.

- By disallowing descendant $//$ we obtain restriction $\mathcal{CTQ}$.

Example: Conjunctive Tree Query

List all pairs of authors that have written articles with the same title.

$Q(x, y) \ :=$



Certain Answers Semantics

- Given: A source tree T and a conjunctive tree query Q over the target.

- Answer to Q should represent the answer to this query in the space of solutions for T.

- Certain answers semantics:

$$\underline{\textit{certain}}(Q, T) \quad = \bigcap_{T' \text{ is a solution for } T} Q(T').$$

Computing Certain Answers

We study the following problem.

Given data exchange setting $(D_{\mathbf{S}}, D_{\mathbf{T}}, \Sigma_{\mathbf{ST}})$ and query Q:

PROBLEM:	CERTAIN-ANSWERS(Q).
INPUT:	Tree T conforming to $D_{\mathbf{S}}$ and tuple $\bar{a}$.
QUESTION:	Is $\bar{a} \in$ <u>*certain*</u>(Q, T)?

Computing Certain Answers: General Picture

Theorem For every XML data exchange setting and $\mathcal{CTQ}^{//}$-query Q, $\textsc{Certain-Answers}(Q)$ is in coNP.

Remark: In terms of the size of the document (data complexity).

Theorem There exist an XML data exchange setting and a $\mathcal{CTQ}^{//}$-query Q such that $\textsc{Certain-Answers}(Q)$ is coNP-hard.

We want to find tractable cases ...

Computing Certain Answers: Finding Tractable Cases

- To find tractable cases, we have to concentrate on fully-specified STDs:

  We impose restrictions on tree patterns over target DTDs:

  - no descendant relation $//$; and
  - no wildcard $\_$; and
  - all patterns start at the root.

  No restrictions imposed on tree patterns over source DTDs.

- Subsume non-relational data exchange handled by Clio.

From now on, all STDs are fully-specified.

Computing Certain Answers: Towards a Classification

Given a class $\mathcal{C}$ of regular expressions and a class $\mathcal{Q}$ of queries:

$\mathcal{C}$ is tractable for $\mathcal{Q}$ if for every data exchange setting in which target DTDs only use regular expressions from $\mathcal{C}$ and every $\mathcal{Q}$-query Q, CERTAIN-ANSWERS(Q) is in PTIME.

$\mathcal{C}$ is coNP-complete for $\mathcal{Q}$ if there is a data exchange setting in which target DTDs only use regular expressions from $\mathcal{C}$ and a $\mathcal{Q}$-query Q such that CERTAIN-ANSWERS(Q) is coNP-complete.

Remark (Ladner): If PTIME $\neq$ NP, there are problems in coNP which are neither tractable nor coNP-complete.

Computing Certain Answers: Towards a Classification

- Our classification is based on classes of regular expressions used in target DTDs.

- They must contain the simplest type of regular expressions: $(a + b + c)^*$

- Such classes will be called admissible.

Computing Certain Answers: Dichotomy

Theorem

1) Every admissible class $\mathcal{C}$ of regular expressions is either tractable or coNP-complete for $\mathcal{CTQ}^{//}$.

2) For every tractable class: Given a source tree T, one can compute in PTIME a solution $T^\star$ for T such that

$$\underline{\textit{certain}}(Q,T) \quad = \quad \textit{remove\_null\_tuples}(Q(T^\star)).$$

3) It is decidable whether the regular expressions used in a target DTD belong to a tractable class.

Outline

- XML data exchange settings.
  - XML source-to-target dependencies.
- Query answering in XML data exchange.
- Final remarks.

Final Remarks

- Dichotomy also holds for unions of conjunctive queries.

- Future work:

  - We would like to consider XML query languages that produce XML trees.

    How do we define certain answers?

  - The notion of reasonable solutions needs to be investigated further.