(Modal) Logics for Semistructured Data (bis)

Stéphane Demri

Laboratoire Spécification et Vérification
CNRS UMR 8643 & INRIA Futurs Proj. SECSI & ENS de Cachan

Plan of the talk

1. Semistructured data (SSD).
2. Reasoning tasks.
3. Logical languages.
4. A detailed example: path constraints.
5. Comparing type constraints.
6. Miscellaneous.

Why modal logics?

“ Being ’modal’ is neither a merit nor a fault, in itself; it is merely a difference. Modality makes it easier to describe just [...] whereas it makes it more difficult to describe [...] “

(Cardelli & Ghelli 01)

What to expect from ML for SSD?

1. To encode naturally reasoning tasks for SSD.

2. To get new decidability/complexity results.

3. To use modal theorem provers to solve problems on SSD.

4. To compare the expressivity of querying languages and schema languages with those of (extended) modal logics.

5. To design and study modal logics with new features.

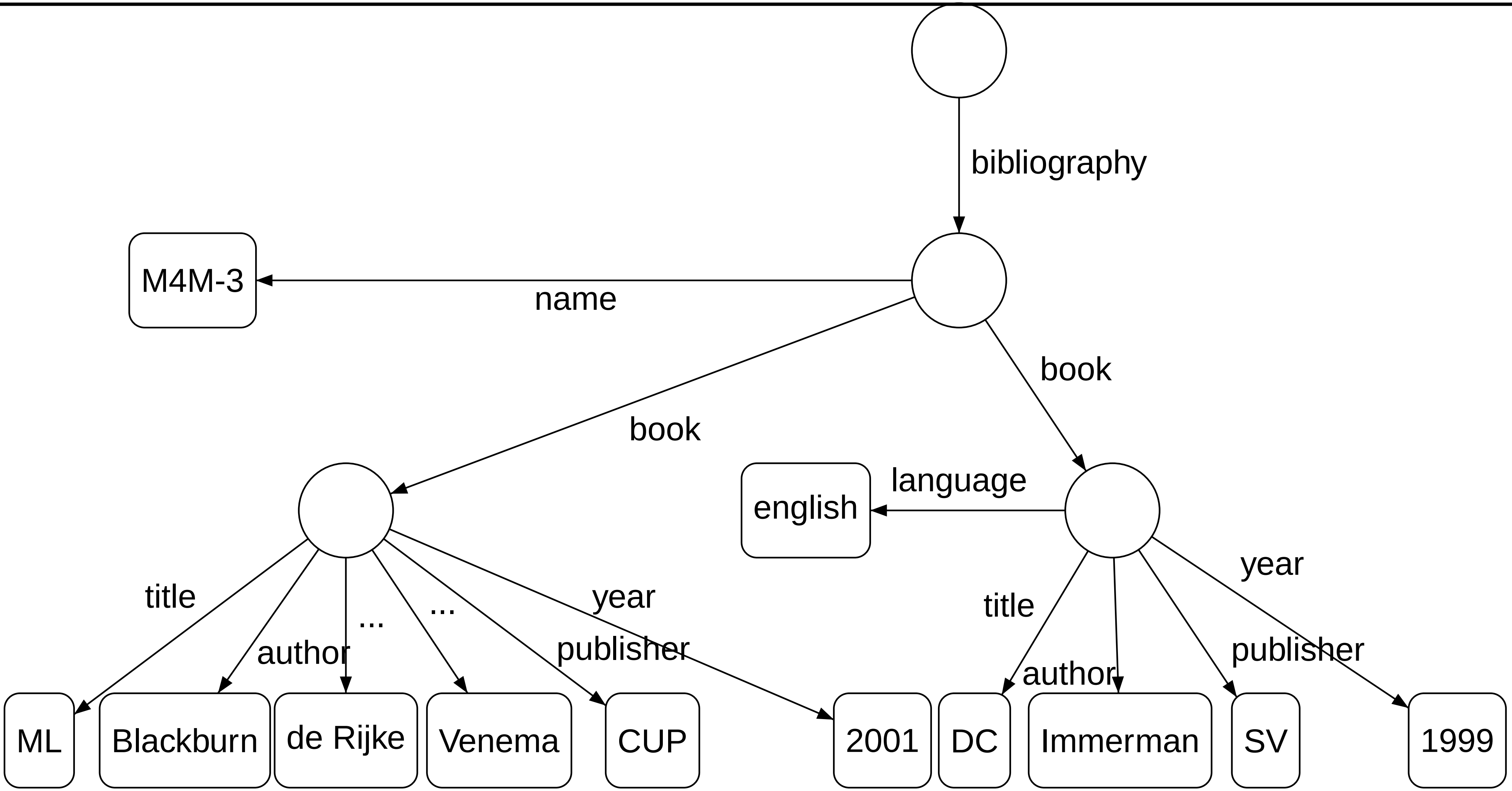
Semistructured data

- Relaxation of classical relational model.
- Schema-less (but need for delineating the meaningful data).
- Self-describing (this is controversial !!).
- Best described by a rooted edge labeled graph.
- Examples:
  - XML (eXtended Markup Language) documents.
  - Web pages with hypertext links.
    (pages are nodes, hyperlinks are labeled edges).

An XML document

```
<bibliography name="M4M-3">
<book>
<title> Modal Logic </title>
<author> Blackburn </author>
<author> de Rijke </author>
<author> Venema </author>
<publisher> Cambridge University Press </publisher>
<year> 2001 </year>
</book>
<book language = "english">
<title> Descriptive Complexity </title>
<author> Immerman </author>
<publisher> Springer-Verlag </publisher>
<year> 1999 </year>
</book>
</bibliography>
```

Tree representation



Diversity of models

Great variety of models/graphs for semistructured data.

- Labels on edges, on nodes.
- Trees vs rooted connected graphs.
- Ordered vs unordered trees.
- Ranked vs unranked trees.

Reasoning tasks parameterized by the models for SSD.

Reasoning tasks

- Querying (model-checking)
  - Integrity constraints.
    E.g., path constraints $(a \cdot b)^* \subseteq (c \cup e)$.
  - Type constraints.
    E.g., membership problem for regular tree languages.
- Comparing constraints (validity)
  - Emptiness problem for a Boolean expression built over constraints.
    E.g., implication of path constraints $a \subseteq b \models a \cdot c \subseteq b \cdot c$,
    equivalence between tree automata.
  - Comparing integrity constraints given type constraints.

The modal approach

- To provide a structured method for implementing querying languages by taking advantage of known techniques for (hybrid) modal logics.
- To encode reasoning tasks for SSD into problems for modal logics, e.g. the comparison of constraints encoded as a validity problem.
- A reasonable expectation: complexity of validity comparable to the complexity of the problem on constraints.
- Additional requirement: modal encoding of W3C standards.

W3C standards

Additional requirement: modal encoding of W3C standards

- World Wide Web Consortium (W3C) develops specifications, guidelines, etc for the Web.
- W3C standards
  - Document Object Model (DOM).
  - eXtensible Markup Language (XML).
  - XML Path language XPath for addressing part of an XML document.
  - XML Schema.
- XQuery, eXtensible Style Language Transformations (XSLT) for which XPath is a building block.

Pioneering works

- Schemes subsumption encoded into a hybrid modal logic (Alechina 97).

- Schemes subsumption encoded into a description logic (Calvanese et al 98).

- Equivalence of Document Type Definitions (DTDs) for XML documents encoded into a PDL-like description logic with
  - qualified number restrictions,
  - well-foundedness operator.

  (Calvanese et al 99).

Other works

- Tree logic based on modal ambient logic expressing path formulas (Cardelli & Ghelli 01).
- Comparison of XPath fragments and Computational Tree Logic CTL (Miklau & Suciu 02), (Gottlob & Koch 02).
- Implication of path constraints encoded into a fragment of Converse PDL with nominals (Alechina et al 03).
- DTD with well-typed references encoded into a hybrid modal logic with binder $\downarrow$, fragment of FO + TC (Bidoit et al 03).
- Path constraints encoded into fragments of hybrid modal logics (Franceschet & de Rijke 03).
- XPath queries and equivalence problem encoded into PDL over finite node labelled ordered trees (Marx 03).

Querying: standard logical languages

Why not to use standard logical languages?

- Monadic second order logic.
- First-order logic.
- Modal μ-calculus.
- Modal logics (PDL).

Complexity of model-checking

MSO	PSPACE-complete
μ-calculus	in NP $\cap$ co-NP
FO	PSPACE-complete
PDL	P-complete

Theorem. (Courcelle 90) Model checking MSO formulae for graphs of bounded tree-width is in time linear in the size of the graph.

Tree-width measures how close are graphs to being trees.

XPath: a serious competitor

- XPath expression establishes a relation between
  - a context node,
  - a node in the answer set.
- Core XPath: large, practical fragment of XPath.
- MC for the W3C standard Core XPath is only P-complete. (Gottlob et al 03).

What is a modal language for SSD?

(assuming we know what is a modal language.)

- To be between FO and MSO? To be hybrid?
- To encode reasoning tasks for W3C standards?
- To have tractable model-checking problem?
- To have a decidable satisfiability problem?
- to be a declared spatial logic for querying graphs (between MSO and FO), see e.g. (Cardelli et al 02).

A detailed example: path constraints

- Integrity constraints for SSD from (Abiteboul & Vianu 97).
- Interests of regular path expressions:
  - They give semantical information on the data.
  - They are used for query optimization.
- Regular path expressions:

$$p ::= a \in L \mid \epsilon \mid p + p \mid p^* \mid p; p \mid \sharp.$$

- $\sharp$: wildcard.
- Simple path expressions: $p ::= a \mid \epsilon \mid p; p$.

Models

- Rooted edge labeled connected graphs:
  - (XML) Documents with pointers (id/idref attributes).
  - Web pages with hyperlinks.
- L-structure: $G = \langle V, rt, (R_a)_{a \in L} \rangle$.
- deterministic vs non-deterministic structures. Deterministic models more appropriate for Web pages.

Transition relations

$$\begin{array}{rcl}
tr(a) & = & R_a \text{ for } a \in L \\
tr(p^*) & \text{is} & \text{the reflexive transitive closure of } tr(p) \\
tr(\#) & = & \bigcup_{a \in L} R_a \\
tr(\epsilon) & = & \{\langle u, u\rangle : u \in V\} \\
tr(p_1\,;p_2) & = & \{\langle u, v\rangle : \exists z\,(tr(p_1)(u,z) \wedge tr(p_2)(z,v))\} \\
tr(p_1 + p_2) & = & tr(p_1) \cup tr(p_2)
\end{array}$$

Path constraints

- Forward constraint:
  $G \models p \subseteq_f q$ iff $tr(p)(rt) \subseteq tr(q)(rt)$.

- Backward constraint:
  $G \models p \subseteq_b q$ iff $tr(p)(rt) \subseteq tr(q)(rt)^{-1}$.

- Standard path constraints: $p \subseteq q$.

- Lollipop path constraint:
  $G \models r \rightsquigarrow p \subseteq q$ iff for every $x \in tr(r)(rt)$,
  $\langle V, x, (R_a)_{a \in L} \rangle \models p \subseteq q$.

Problems

- Query evaluation problem for a class C of path constraints:

  instance: a finite L-structure G and a constraint c in C;

  question: $G \models c$?

- Containment problem for a class C of path constraints:

  instance: constraints $c_1, \ldots, c_{n+1}$, $n \geq 0$, in C;

  question: is it the case that for every L-structure G, $G \models c_1$ and $\ldots$ and $G \models c_n$ imply $G \models c_{n+1}$?
  (if so, we write $c_1, \ldots, c_n \to c_{n+1}$.)

- Variants: containment problem over finite and/or deterministic structures.

Decidable problems

Theorem. (Abiteboul & Vianu 97) The [resp. finite] containment problem for forward constraints with simple path expressions is in PTIME.

Theorem. (Buneman et al 98 bis) The [resp. finite] containment problem for forward constraints with simple path expressions over deterministic structures is decidable in linear-space.

Theorem. (Buneman et al 98 bis) The containment problem for lollipop constraints with simple path expressions over deterministic structures is decidable in linear-space.

Undecidable problems

Theorem. (Buneman et al 98) The containment problem for lollipop constraints with simple path expressions is undecidable.

Theorem. (Buneman et al 98 bis) The containment problem for lollipop constraints over deterministic structures is undecidable even if L contains only two letters.

PDL^{path}

A PDL-like logic to encode problems on standard path constraints.

- $\varphi ::= \top \mid \bot \mid \ root \ \mid \neg\varphi \mid \varphi \wedge \varphi \mid [t]\varphi \mid \langle t \rangle \varphi$.
- t path expressions possibly including converse $^{-1}$.
- no propositional variables, a unique nominal $root$.
- Models: rooted edge labelled (connected) graphs G.
- Satisfiability/validity problem (at the root).

Translation of constraints

- $c = p \subseteq_f q$, $\varphi_c = [p]\langle(q)^{-1}\rangle root$.
- $c = p \subseteq_b q$, $\varphi_c = [p]\langle q\rangle root$.
- $G \models c$ iff $G \models \varphi_c$.
- c_1, …, c_{n+1} standard path constraints.
  $c_1, \ldots, c_n \to c_{n+1}$ iff $(\varphi_{c_1} \wedge \cdots \wedge \varphi_{c_n}) \Rightarrow \varphi_{c_{n+1}}$ is PDL^{path} valid.

Translation of problems

Lemma. Let C be either the full class of L-structures or the class of deterministic L-structures.

1. The query evaluation problem for standard path constraints is LOGSPACE reducible to the model checking problem for PDL^{path}.

2. The containment problem for forward constraints restricted to L-structures in C is LOGSPACE reducible to the validity problem for PDL^{path} restricted to L-structures in C.

3. The containment problem for backward constraints restricted to L-structures in C is LOGSPACE reducible to the validity problem for PDL^{path} without converse and restricted to L-structures in C.

Results for PDL$^{\mathrm{path}}$ *(Alechina et al 03)*

Theorem. The model checking problem for PDL^{path} is P-complete.

Theorem. The satisfiability and validity problems for PDL^{path} are decidable in EXPTIME.
(translation into CPDL with nominals)

Theorem. The satisfiability problem for PDL^{path} is EXPTIME-hard whenever $|L| \geq 1$.
(reduction from global satisfiability problem for K with spy-point technique)

Corollary. The minimal tense logic K_t augmented with a single nominal but without proposition letters has an EXPTIME-hard satisfiability problem.

Path problems (Alechina et al 03)

Theorem. The query evaluation problem for the class of path constraints is NLOGSPACE-complete both for deterministic and non-deterministic graphs.

Theorem. The containment problem for forward constraints is in EXPTIME, while it is at least PSPACE-hard if $|L| \geq 2$.

Theorem. The containment problem for backward constraints is in EXPTIME, while it is at least PSPACE-hard if $|L| \geq 3$.

Lemma. The containment problem for backward constraints restricted to deterministic L-structures for finite sets of labels L is in EXPTIME.

Open problems

- Decidability status of satisfiability for PDL^{path} and CPDL with nominals over deterministic structures.
- Decidability status of containment problem for forward constraints over deterministic structures.
- Complexity of containment problem for forward constraints and for backward constraints.
  We know PSPACE lower bound and EXPTIME upper bound.
  Close relationship with prefix rewriting (Debarbieux et al 03).

Typing mechanisms for XML

Typing mechanisms:

- Graph Schemas (Buneman et al 97).
- XML Document Type Definitions (DTDs).
- XML Schema (approved by W3C).
- Tree automata.
- DTD with well-typed references, see e.g. (Bidoit et al 03).

Expressive power:

- DTD and XML Schema less expressive than regular tree languages.
- Equivalence between tree automata and MSO.

A DTD

```
<!DOCTYPE bibliography [
 <!ELEMENT bibliography (book)*>
 <!ELEMENT book         (title, (author)+,
                         publisher?, year>
 <!ELEMENT title        (#PCDATA)>
 <!ELEMENT author       (#PCDATA)>
 <!ELEMENT publisher    (#PCDATA)>
 <!ELEMENT year         (#PCDATA)>
]>
```

Declaration of attributes:

```
<!ATTLIS bibliography name     CDATA #IMPLIES>
<!ATTLIS book         language CDATA #IMPLIES>
```

Comparing type constraints

$L(T)$: language of structures (trees) with type T.

- Inclusion: $L(T_1) \subseteq L(T_2)$?
- Equivalence: $L(T_1) = L(T_2)$?
- Emptiness of intersection: $L(T_1) \cap L(T_2) = \emptyset$?
- Implication: $L(T_1) \cap \ldots \cap L(T_n) \subseteq L(T_{n+1})$?
- Existence of a lot more variants.

Variants

- Analogous problems with integrity constraints instead of type constraints.
  E.g., containment problem for standard path constraints.
- Analogous problems relativized to finite structures.
- Analogous problems relativized to structures satisfying integrity constraints, see e.g. (Buneman et al 03).
- Equivalence or root equivalence of (XPath) queries, see e.g. (Marx 03).

The modal approach

(Calvanese et al 99)

- XML documents as labeled unranked ordered finite trees.
- DTDs T_1, T_2, $\mathcal{R}$ equivalence relations between tags.
  (to abstract similar tags.)
- $T_1 \sqsubseteq_{\mathcal{R}} T_2$ iff $\varphi(T_1, T_2, \mathcal{R})$ is DL valid.
- DL variant of Repeat Automata DPDL.

Theorem. (Calvanese et al 99) Equivalence of DTDs is in EXPTIME.

Presburger constraints

- Presburger arithmetic decidable in 2EXPSPACE.
- Presburger tree automata recognizing finite unordered/ordered unranked trees with Presburger constraints on the number of children (Seidl et al 03).
- Presburger MSO logic over unordered trees is decidable (Seidl et al 03).
- Similar automata for XML documents studied in (Lugiez & Dal Zilio 03).

Presburger Modal Logic

- Example of modal logic with new features motivated by reasoning tasks for SSD.
- $\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid (\langle a_1 \rangle_{x_1} \varphi_1, \ldots, \langle a_n \rangle_{x_n} \varphi_n)_{A(x_1,\ldots,x_n)}$.
- $A(x_1, \ldots, x_n)$: formula of Presburger arithmetic with free variables $x_1, \ldots, x_n$ expressing a constraint on the numbers of a_i-successors.
- $\Diamond_{\geq 3}\, \varphi \approx (\Diamond_x\, \varphi)_{x \geq 3}$ (graded modal logics).

Semantics

- Models: finite branching LTSs.

- $\sharp_{a,\varphi}(w) = |\{w' : (w, w') \in R_a, \mathcal{M}, w' \models \varphi\}|$.

- $\mathcal{M}, w \models (\langle a_1 \rangle_{x_1} \varphi_1, \ldots, \langle a_n \rangle_{x_n} \varphi_n)_{A(x_1,\ldots,x_n)}$ iff
  $x_1 \leftarrow \sharp_{a_1,\varphi_1}(w), \ldots, x_n \leftarrow \sharp_{a_n,\varphi_n}(w) \models_{\mathrm{Presburger}} A(x_1, \ldots, x_n)$.

Complexity issues

- What are the PSPACE fragments of Presburger Modal Logic? E.g. graded modal logics with binary encoding of integers (Tobies 00).

- What are the EXPTIME fragments of Presburger Modal Logic + fixed point operators? E.g. graded μ-calculus (Kupferman et al 02).

Perspectives

- To improve the impact of modal logics for SSD.
  E.g., to demonstrate the practical effects of the modal approach.
- To extend the modal approach for richer specification languages.
  E.g., to handle DTD with well-typed references.
- To compare the ML approach with the automata-based approach for SSD.
  E.g., is there a place for two concurrent structured frameworks?
- To design and study new modal logics inspired by SSD.
  E.g., to study complexity issues for fragments of Presburger Modal Logic.

(Abiteboul & Vianu 97)

```
@inProceedings{Abiteboul&Vianu97,
   author =    {S. Abiteboul and V. Vianu},
   title =     {Regular path queries with
                constraints},
   booktitle = {PODS’97},
   pages = "122--133",
   year =      1997,
}
```

(Alechina 97)

```
@techreport{Alechina97,
    author = "N. Alechina",
    title = "Semi-structured information: a modal
              logic approach",
    number = "CSR-97-08",
    month = "August",
    year = "1997",
    institution = "School of Computer Science,
                   The University of Birmingham"
}
```

(Alechina et al 03)

```
@Article{Alechina&Demri&DeRijke03,
  author =      "N. Alechina and S. Demri
                 and M. de Rijke",
  title =       "A modal perspective on
                 path constraints",
  journal =     JLC,
  year =        "2003",
  note =        "To appear"
}
```

(Bidoit et al 03)

```
@InProceedings{Bidoit&Cerrito&Thion03,
  author =         {N. Bidoit and S. Cerrito and
                    V. Thion},
  title =          "Un premier pas vers la
                    modélisation des données
                    semi-structurées par la logique
                    multi-modale hybride",
  booktitle = {19èmes Journées des Bases de
               Données Avancées (BDA 2003),
               Toulouse,  France},
  year =           {2003},
  month =          {October}
  note =           {To appear},
}
```

(Buneman et al 97)

```
@InProceedings{Bunemanetal97,
  author =         {P. Buneman and S. Davidson and
                    M. Fernandez and D. Suciu},
  title =          {Adding structure to
                    unstructured data},
  booktitle = {6th International Conference on
               Database Theory (ICDT’97)},
  pages =          {336--350},
  year =           {1997},
}
```

(Buneman et al 98)

```
@inProceedings{Buneman&Fan&Weinstein98,
   author =    {P. Buneman and W. Fan and
                S. Weinstein},
   title =     {Path constraints on semistructured
                and structured data},
   booktitle = {PODS’98},
   page = "129--138",
   year =      1998,
}
```

(Buneman et al 98 bis)

```
@techReport{Buneman&Fan&Weinstein98b,
   author =      {P. Buneman and W. Fan and
                  S. Weinstein},
   title =       {Path constraints on
                  deterministic graphs},
   number =      {Technical Report MS-CIS-98-33},
   institution = {LINCS, CIS, UPenn},
   year =        1998,
}
```

(Buneman et al 03)

```
@Article{Buneman&Fan&Weinstein03,
  author =         {P. Buneman and W. Fan and
                    S. Weinstein},
  title =          {Interaction between Path
                    and Type Constraints},
  journal =        TOCL,
  year =           {2003},
  note =           {to appear. Long version
                    of a paper in PODS’99},
}
```

(Calvanese et al 98)

```
@InProceedings{Calvanese&DeGiacomo&Lenzerini98,
  author =       {D. Calvanese and G. De Giacomo
                  and M. Lenzerini},
  title =        {What can knowledge representation
                  do for semi-structured data},
  booktitle = {Fifteenth National Conference on
               Artificial Intelligence and Tenth
               Innovative Applications of Artificial
               Intelligence Conference (AAAI/IAAI),
               Madison, Wisconsin},
  pages =        {205--210},
  year =         {1998},
  publisher = {AAAI Press - MIT Press},
}
```

(Calvanese et al 99)

```
@Article{Calvanese&DeGiacomo&Lenzerini99,
  author =      "D. Calvanese and G. De Giacomo and
                 M. Lenzerini",
  title =       "{R}epresenting and {R}easoning on
                 {XML} Documents: A {D}escription
                 {L}ogic {A}pproach",
  journal =     JLC,
  year =        "1999",
  volume =      "9",
  number =      "3",
  pages =       "295--318",
}
```

(Cardelli & Ghelli 01)

```
@InProceedings{Cardelli&Ghelli01,
  author =          {L. Cardelli and G. Ghelli},
  title =           {A query language based on the
                     ambient logic},
  booktitle = {ESOP’01},
  pages =           {1--22},
  year =            {2001},
  editor =          {D. Sands},
  volume =          {2028},
  series =          LNCS,
  publisher = "Springer, Berlin",
}
```

(Cardelli et al 02)

```
@InProceedings{Cardelli&Gardner&Ghelli02,
  author =        {L. Cardelli and Ph. Gardner and
                   G. Ghelli},
  title =         {A Spatial Logic for Querying
                   Graphs},
  booktitle =     {ICALP’02, Malaga, Spain},
  pages =         {597--610},
  year =          {2002},
  volume =        {2380},
  series =        LNCS,
  publisher = {Springer, Berlin},
```

(Courcelle 90)

```
@InProceedings{Courcelle90,
  author =        "B. Courcelle",
  title =         "Graph rewriting: An algebraic and
                   logic approach",
  booktitle =     "Handbook of Theoretical Computer
                   Science, Volume B, Formal models
                   and semantics",
  year =          "1990",
  editor =        "J. Van Leeuwen",
  pages =         "193--242",
  publisher =     "Elsevier",
}
```

(Dal Zilio & Lugiez 03)

```
@InProceedings{DalZilio&Lugiez03,
  author =       {S. Dal Zilio and D. Lugiez},
  title =        "{XML} Schema, Tree Logic and
                  Sheaves Automata",
  booktitle = {RTA 2003},
  pages =        {246--263},
  year =         {2003},
  editor =       {R. Nieuwenhuis},
  volume =       {2706},
  series =       LNCS,
  publisher = {Springer, Berlin},
}
```

(Debarbieux et al 03)

```
@InProceedings{Debarbieuxetal03,
  author =       {D. Debarbieux and Y. Roos and
                  S. Tison and Y. Andr{\’e} and
                  A.-C. Caron},
  title =        {Path Rewriting in
                  Semistructured Data},
  booktitle = {4th International Conference on
               Words (WORDS 2003), Turku, Finland},
  year =         {2003},
  note =         {Published in a TUCS report},
}
```

(Miklau & Suciu 02)

```
@inproceedings{Miklau&Suciu02,
        author      =   "G. Miklau and D. Suciu",
        title       =   "Containment and Equivalence
                         for an XPath Fragment",
        booktitle  =    "PODS’02",
        pages       =   "65--76",
        year        =   "2002"
}
```

(Gottlob & Koch 02)

```
@InProceedings{Gottlob&Koch02,
  author =      "G. Gottlob and C. Koch",
  title =       {Monadic Queries over
                 Tree-Structured Data},
  booktitle =   "LICS’02",
  year =        "2002",
  pages =       "189--202",
  publisher =   "IEEE Computer Society"
}
```

(Gottlob et al 03)

```
@InProceedings{Gottlob&Koch&Pichler03,
  author =       "G. Gottlob and C. Koch and
                  R. Pichler",
  title =        {The complexity of XPath
                  query evaluation},
  booktitle = {PODS’03},
  pages =        {179--190},
  year =         {2003},
}
```

(Kupferman et al 02)

```
@inproceedings{Kupferman&Sattler&Vardi02,
  author      =   "O. Kupferman and
                   U. Sattler and
                   M.Y. Vardi",
  title       =   "The Complexity of the
                   Graded $\mu$-Calculus",
  booktitle  =    "CADE-18",
  series      =   LNCS,
  volume      =   "2392",
  pages       =   "423--437",
  editor      =   "A. Voronkov",
  publisher   =   "Springer, Berlin",
  year        =   "2002"
}
```

(Seidl et al 03)

```
@InProceedings{Seidl&Schwentick&Muscholl03,
  author =       {H. Seidl and Th. Schwentick and
                  A. Muscholl},
  title =        {Numerical Document Queries},
  booktitle = {PODS 2003, San Diego, CA},
  pages =        {155--166},
  year =         {2003},
  publisher = {ACM},
}
```

(Tobies 00)

```
@Article{Tobies00,
  author =      "S. Tobies",
  title =       "{PSPACE} Reasoning for Graded
                 Modal Logics",
  journal =     JLC,
  year =        "2000",
  volume =      "10",
  pages =       "1--22",
}
```