

Constraints-Preserving Inlining: XPRESS Approach

Dongwon Lee

Dept. of Computer Science

UCLA

Introduction



- Dramatic increase of web data (HTML, XML)
- XML gains popularity as a candidate
- Needs to manage XML data arise
- Two approaches
 - Building customized DB
 - Using RDB as underlying engine
- **XPRESS** (Xml Processing and Relaxation in rElational Storage System) project @ UCLA/CSD



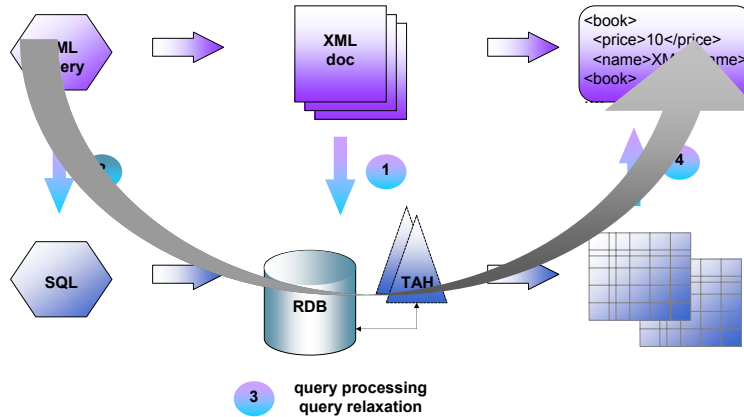
XPRESS: Architecture



Query

Database

Result



October 11

ER 2000

3

XPRESS: Pros & Cons



■ Pros

- Present market is dominated by RDB; impractical to abandon RDB to adopt XML
- Mature RDB techniques (OLAP, Data Warehousing, etc) can be reused
- Support various types of input/output

■ Cons

- High conversion costs btw. XML and Relational models
- Incorrect or incomplete conversion

October 11

ER 2000

4

Related Works



- STORED @ AT&T Labs [SIGMOD98]
- Inlining @ U. Wisconsin [VLDB99]
- INRIA [IEEE Eng. Bulletin 00]

- XML-DBMS, IBM DB2 XML Extender, Informix Web DataBlade, Oracle 8i, Sybase Adaptive Server Enterprise 12.0, ...

- Existing mapping methods
 - Structure-oriented; ignoring constraints
 - Require programming/human intervention

Background: XML & DTD



- Not a single, predefined markup language (e.g. HTML): it's a meta-language by W3C
 - Start and end tags (`<name>...</name>`)
- Two XML schema languages from W3C
 - DTD (Document Type Definition), XML-Schema
- **DTD**: formal description about the *structures* and *constraints* of the XML document
 - XML vs. DTD \approx SQL vs. DDL
 - Element vs. attribute (DTD) \approx table vs. column (RDB) \approx class vs. attribute (OO)

Document Type Definition (DTD)



- *Element* (ordered) & *attribute* (unordered)

```
<!ELEMENT elm-name elm-type>  
<!ATTLIST elm-name att-name att-type att-option>
```

- **Element type** <elm-type>
 - String type (#PCDATA)
 - 0 or 1 (?), 0 or more (*), 1 or many (+)
 - Sequential (.), choice (|)
- **Attribute type** <att-type>
 - String type (CDATA)
 - identifier (ID), foreign key (IDREF, IDREFS)
- **Attribute option** <att-option>
 - #IMPLIED, #REQUIRED

October 11

ER 2000

7

Example



- A *paper* element has a unique *id*, one *title*, one or many *authors*, and zero or many referenced *papers*:

```
<!ELEMENT paper (title,author+)>  
<!ATTLIST paper pid ID #REQUIRED  
                ref IDREFS #IMPLIED>  
  
<paper pid="p10" ref="p1 p3">  
  <title>...</title>  
  <author>A</author><author>B</author>  
</paper>
```

October 11

ER 2000

8

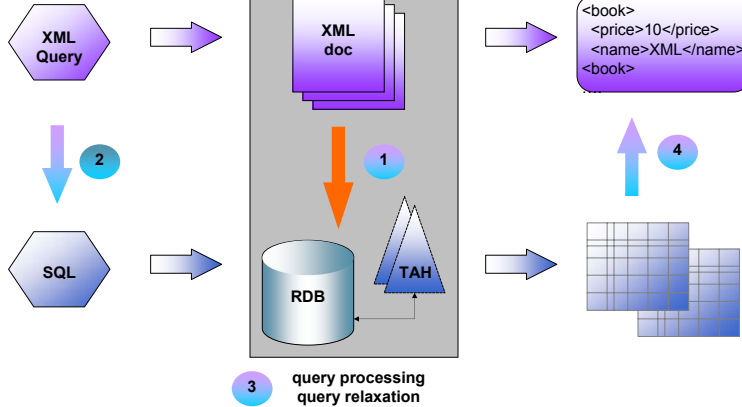
XML to Relational Mapping



Query

Database

Result



October 11

ER 2000

9

Difficulties



- No 1-to-1 mapping
- Set (a^* , $(b+|c?)$), Recursion

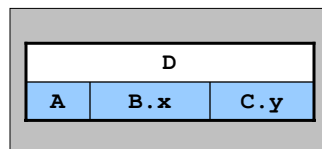
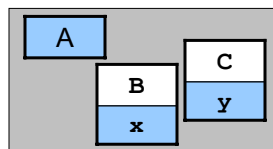
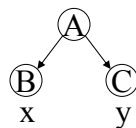
```
<list name="A"><item>1</item><item>2</item> </list>
<list name="B"><item>3</item><item>4</item> </list>
```

- Fragmentation & inlining

<!ELEMENT A (B|C)>

1. Fragmented: 3 tables A, B, C
2. Inlined: 1 table D

list	
name	item
A	1,2
B	3,4



October 11

ER 2000

10

CPI Algorithm



- Uses structure-oriented translation algorithm (e.g., hybrid inlining algorithm [VLDB 99]) as basis
- Preserves **constraints** during the translation
- Convert DTD to a digraph with annotated edge types
- Identify **top nodes**: 1) source, 2) child of *,+ node, 3) recursive node with indegree>1, ...
- Map top nodes *T* to **table *t*** (to avoid non-1NF); map leaf nodes reachable from *T* to **column *c* of *t*** via inlining unless *T* is another top node
- New columns for bookkeeping
 - `fk_key`, `parent_elm`, `root_elm`, `ordinal`, ...

Constraints in DTD



■ Domain constraint

```
<!ATTLIST author gender (M|F) #REQUIRED
                married (Y|N) #IMPLIED>
```

■ Cardinality constraint

```
<!ELEMENT book (title,author+,ref*,price?)>
```

- 1-to-{0,1}: at most 1 (`price`)
- 1-to-{1}: must be 1 (`title`)
- 1-to-{0,...}: any occurrence (`ref`)
- 1-to-{1,...}: at least 1 (`author`)

Constraints in DTD (cont.)



■ Inclusion Dependency (ID)

```
<!ELEMENT person (name, (email|phone)?)>
<!ATTLIST person id ID #REQUIRED>
<!ELEMENT contact EMPTY>
<!ATTLIST contact aid IDREF #REQUIRED>
<!ELEMENT editor (person*)>
<!ATTLIST editor eid IDREFS #IMPLIED>
```

aid \subseteq id, eid \subseteq id

Constraints in DTD (cont.)



■ Equality-Generating Dependency (EGD)

- Values in one columns require values in other columns be *equal*
- In XML, EGD is disguised as “Singleton” property
- When an element instance x of type X satisfies the singleton property towards its sub-element instances y_1 and y_2 of type Y , y_1 and y_2 must be equal
- 1-to- $\{0, 1\}$ and 1-to- $\{1\}$ cardinality cases

$X \rightarrow X.Y$

Constraints in DTD (cont.)



- Tuple-Generating Dependency (TGD)
 - Require some tuples of a certain form be **present**
 - In XML, TGD is disguised as “Not-Nullness” property
 - Child property ($P \rightarrow C$): Every element of type P must have at least one child element of type C
 - 1-to- $\{1\}$ and 1-to- $\{1, \dots\}$ cardinality cases
 - Parent property ($C \rightarrow P$): Every element of type C must have a parent element of type P
 - Only can be enforced by semantic knowledge since any proper element can be a root w/o parent

October 11

ER 2000

15

Conference.dtd



```
<!ELEMENT conf (title,date,editor?,paper*)>
<!ATTLIST conf id ID #REQUIRED>
<!ELEMENT title (#PCDATA)>
<!ELEMENT date EMPTY>
<!ATTLIST date year CDATA #REQUIRED mon CDATA #REQUIRED
day CDATA #IMPLIED>
<!ELEMENT editor (person*)>
<!ATTLIST editor eids IDREFS #IMPLIED>
<!ELEMENT paper (title,contact?,author,cite?)>
<!ATTLIST paper id ID #REQUIRED>
<!ELEMENT contact EMPTY>
<!ATTLIST contact aid IDREF #REQUIRED>
<!ELEMENT author (person+)>
<!ATTLIST author id ID #REQUIRED>
<!ELEMENT person (name,(email|phone)?)>
<!ATTLIST person id ID #REQUIRED>
<!ELEMENT name EMPTY>
<!ATTLIST name fn CDATA #IMPLIED ln CDATA #REQUIRED>
<!ELEMENT email (#PCDATA)>
<!ELEMENT cite (paper*)>
<!ATTLIST cite id ID #REQUIRED format (ACM|IEEE) #IMPLIED>
```

October 11

ER 2000

16

Conference.xml



```
<conf id="er99">
  <title>Int'l Conference on Conceptual Modeling (ER)</title>
  <date> <year>1999</year> <mon>May</mon> <day>20</day> </date>
  <editor ids="sheth bossy">
    <person id="klavans">
      <name fn="Judith" ln="Klavans"/><email>kla@columbia.edu</email>
    </person> </editor>
  <paper id="p1">
    <title>Indexing Model for Structured...</title>
    <contact aid="dao"/>
    <author><person id="dao"><name fn="Tuong" ln="Dao"/>
      </person></author>
  </paper>
  <paper id="p2">
    <title>Logical Information Modeling of...</title>
    <contact aid="shah"/>
    <author>
      <person id="shah"><name fn="Kshitij" ln="Shah"/></person>
      <person id="sheth">
        <name fn="Amit" ln="Sheth"/><email>amit@cs.uga.edu</email>
      </person>
    </author>
  </paper>
</conf>
```

October 11

ER 2000

17

Conference.xml (cont.)



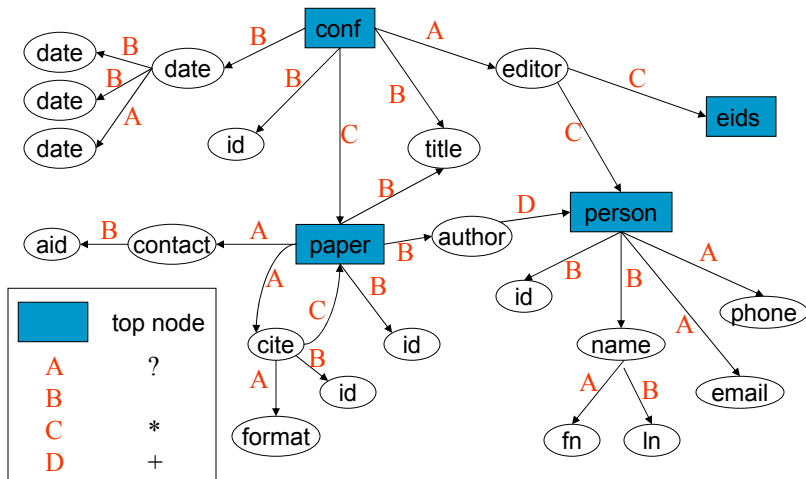
```
<cite id="c100" format="ACM">
  <paper id="p3">
    <title>Making Sense of Scientific Info...</title>
    <author>
      <person id="bossy">
        <name fn="Marcia" ln="Bossy"/><phone>391.4337</phone>
      </person>
    </author>
  </paper>
</cite>
</paper>
</conf>
<paper id="p7">
  <title>Constraints-preserving Transformation...</title>
  <contact aid="lee"/>
  <author>
    <person id="lee">
      <name fn="Dongwon" ln="Lee"/><email>dongwon@cs.ucla.edu</email>
    </person> </author>
  <cite id="c200" format="IEEE"/>
</paper>
```

October 11

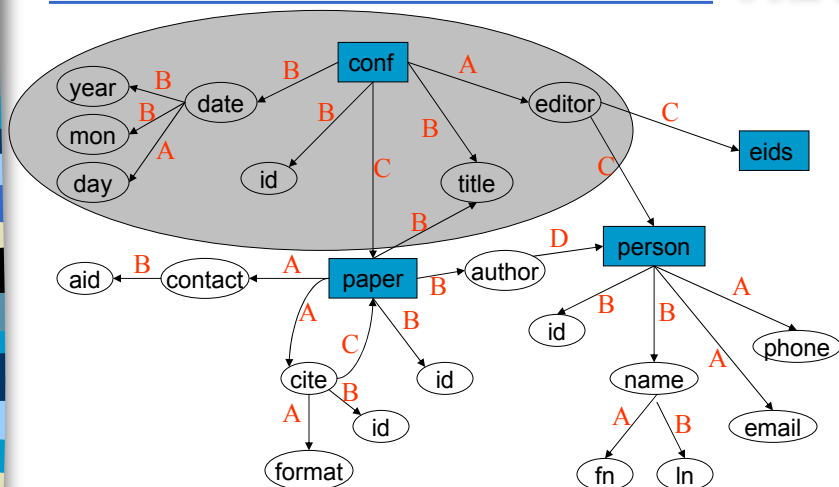
ER 2000

18

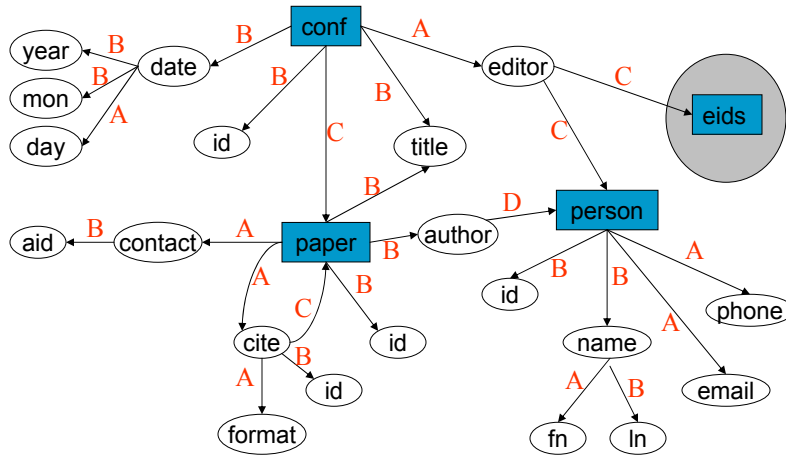
Annotated DTD Graph



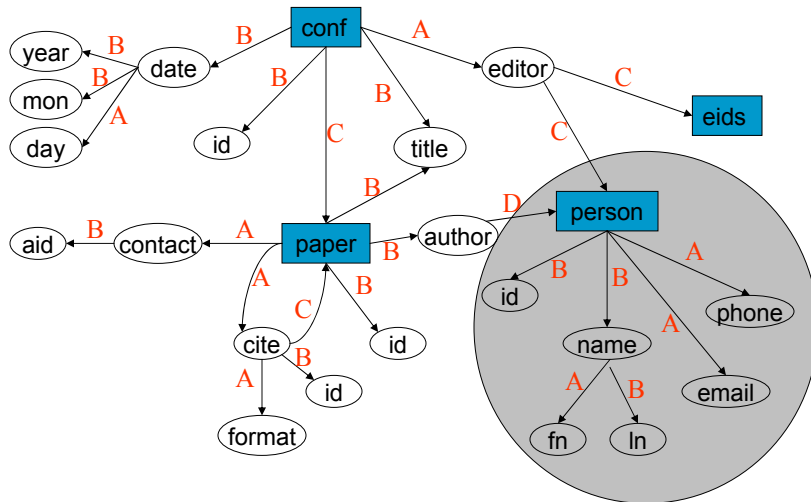
CPI Steps



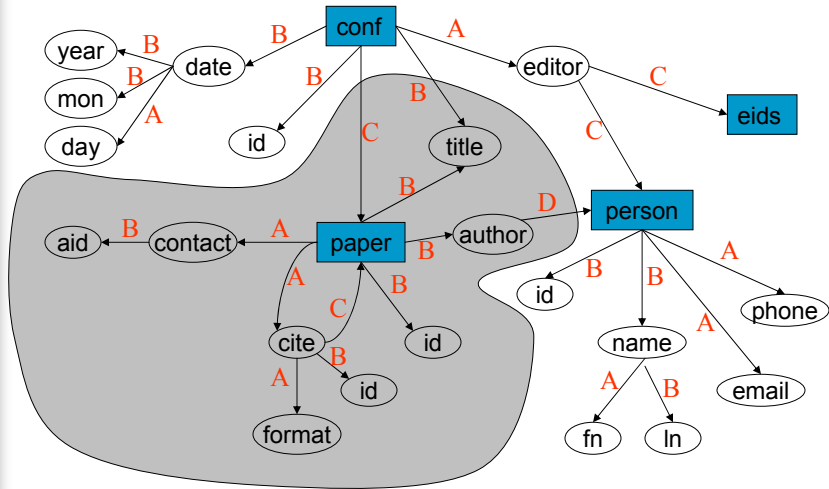
CPI Steps (cont.)



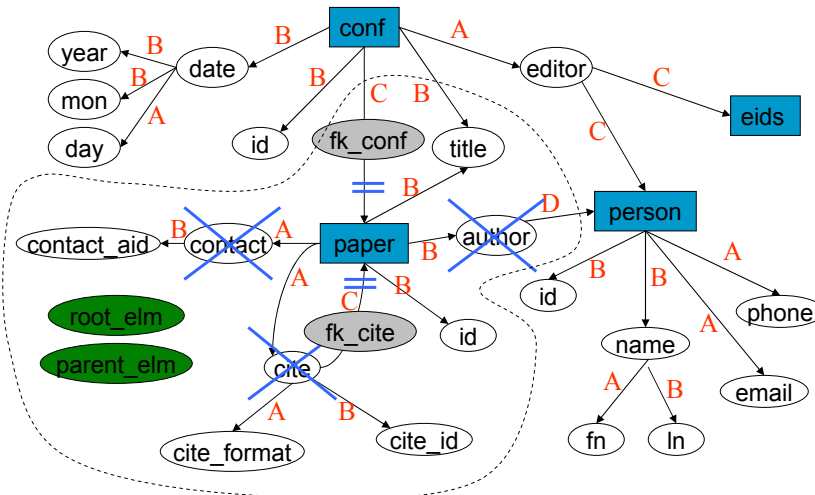
CPI Steps (cont.)



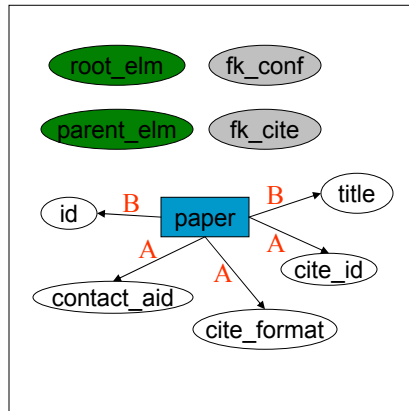
CPI Steps (cont.)



CPI Steps (cont.)



CPI Steps (cont.)



- `id, title` cannot be NULL
- `cite_id, contact_aid, cite_format` can be NULL
- `fk_conf` is a FK to `conf`
- `fk_cite` is included in `cite_id` (i.e., `fk_cite` \subseteq `cite_id`)
- `id` is a PK
- `cite_id` is UNIQUE

October 11

ER 2000

25

Paper Table after CPI



id	root_elm	parent_elm	fk_conf	fk_cite
p1	conf	conf	er99	
p2	conf	conf	er99	
p3	conf	cite		c100
p7	paper			

title	contact_aid	cite_id	cite_format
Indexing...	dao		
Logical...	shah	c100	ACM
Making...			
Constraints...	lee	c200	IEEE

October 11

ER 2000

26

Schema after CPI



```
CREATE TABLE paper (  
  id          NUMBER          NOT NULL,  
  title       VARCHAR(50)     NOT NULL,  
  contact_aid NUMBER,  
  cite_id     NUMBER,  
  cite_format VARCHAR(50)  
             CHECK (VALUE IN ("ACM", "IEEE")),  
  root_elm   VARCHAR(20)     NOT NULL,  
  parent_elm VARCHAR(20),  
  
  fk_cite    VARCHAR(20)  
             CHECK (fk_cite IN (SELECT cite_id FROM paper)),  
  fk_conf    VARCHAR(20),  
  PRIMARY KEY (id),  
  UNIQUE (cite_id),  
  FOREIGN KEY (fk_conf) REFERENCES conf(id),  
  FOREIGN KEY (contact_aid) REFERENCES person(id)  
);
```

October 11

ER 2000

27

Experimentation



- DTD from OASIS site
 - 12 DTDs in different domain: play (Shakespeare), MusicML (music), Xbel (bookmark), tstmt (religious)...
- DBLP data
 - <http://www.cobase.cs.ucla.edu/pub/dblp/>
 - DB-related conferences (79,547), journals (60,963) & books (1,045) XML files (60MB)
 - Each XML file size < 5K bytes
 - Mapped to 10 tables, total 509,392 tuples

October 11

ER 2000

28

Summary



- CPI (Constraints-preserving Inlining) algorithm is presented
 - Work with structure-oriented mapping algorithms
 - Find and preserve constraints during the mapping



<http://www.cobase.cs.ucla.edu/projects/xpress/>