Ein Transparentes Datenmodell zur Verarbeitung Verlinkter XML-Dokumente

Wolfgang May & Dimitrio Malheiro

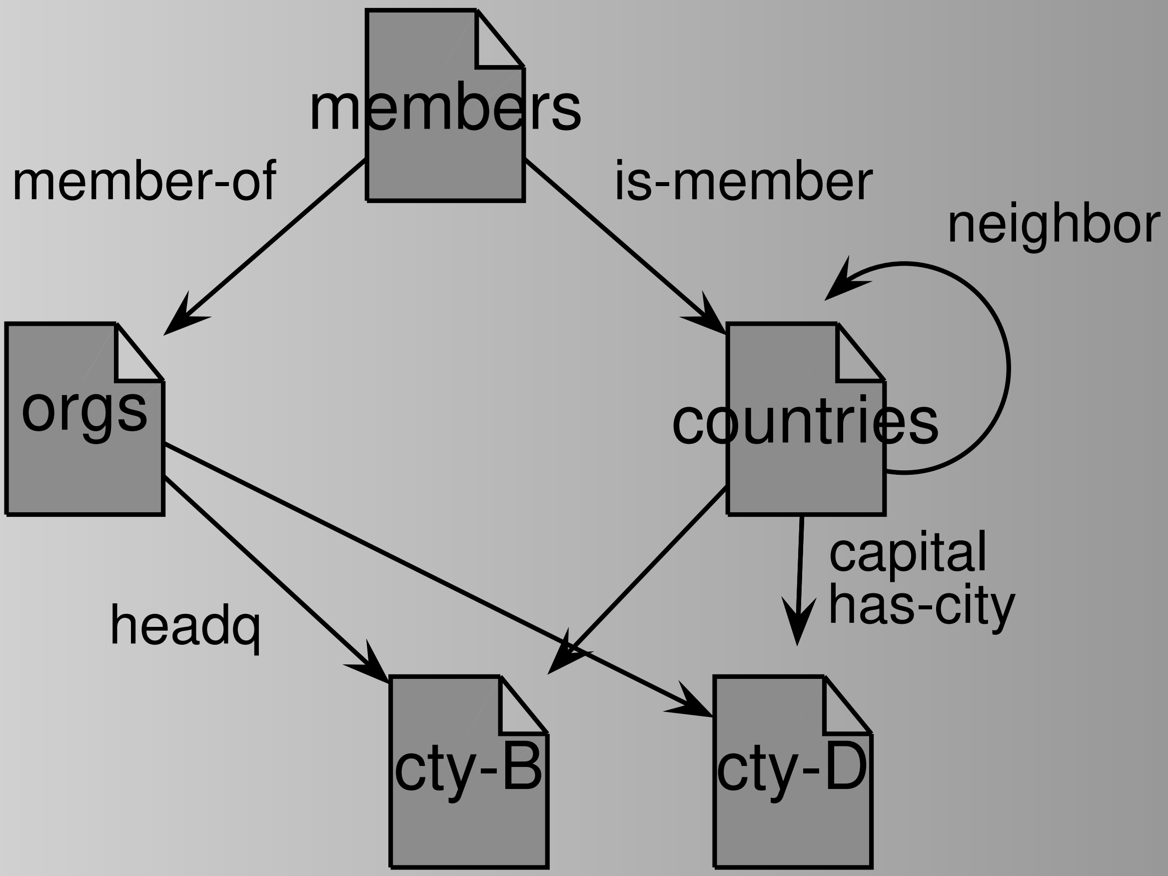
Univ. Freiburg/Göttingen Univ. Freiburg

GI-Fachtagung BTW 2003
Leipzig, 26.2.2003

Situation

- autonome XML-Datenquellen im Web
- Referenzen zwischen Datenquellen

- mondial-countries.xml
- mondial-cities-*car-code*.xml
- mondial-organizations.xml
- mondial-memberships.xml



XLink:Referenzen in XML

- W3C XLink-Standard: Sprache zur Definition von Referenzen zwischen XML-Dokumenten
- Wohin?
  XPointer: *url#xpath-expr*

  <*linkelement* xlink:type="simple" href="*url#xpath-expr*"/>

Simple Links

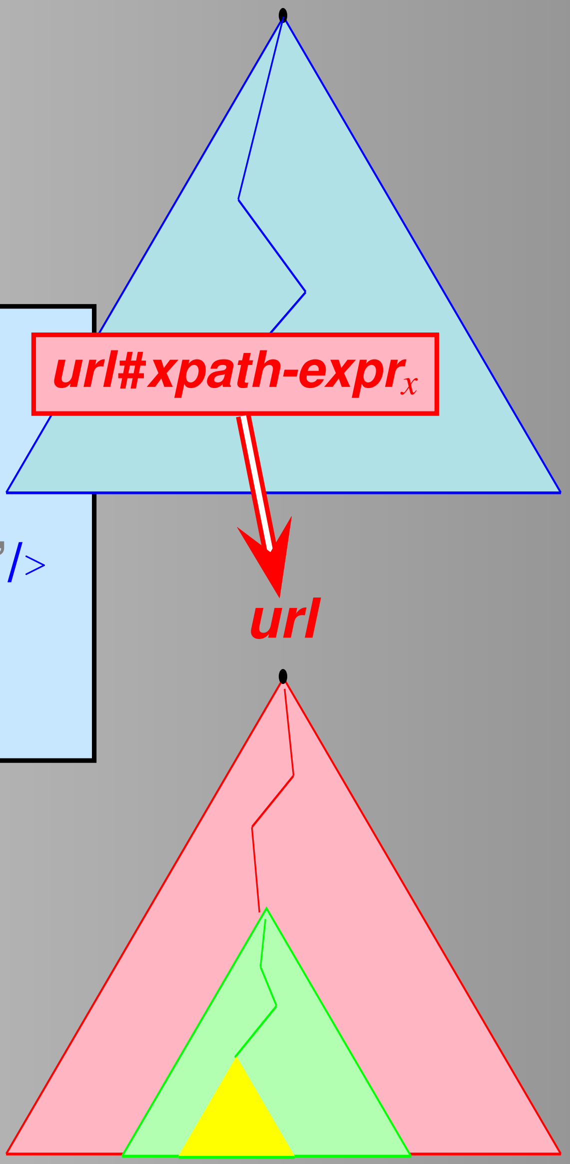
- ähnlich: HTML <A href="...">.

Sitz einer Organisation:

```
<organization abbrev="EU">
   <headq xlink:type="simple"
          xlink:href=
          " file:cities-B.xml#//city[name='Brussels'] "/>
   :
</organization>
```

Anfrage:

```
//organization[@abbrev="EU"]/headq/??/name
```



Anfragen über Links

XPointer und XLink:

- Sprachen um inter-Dokument-Referenzen in XML zu *beschreiben*,
- XLink-Verhaltensspezifikation auf Browsing beschränkt

Bisher nicht spezifiziert

- Semantik von Referenzen im Datenmodell (d.h., XML Query Data Model),
- Formulierung und Auswertung von Anfragen über Referenzen

Sichtweisen

- Anfragender:
  - transparent gegenüber den Benutzern
- eigene Daten referenzieren fremde Datenquellen:
  - Anfragen an die Dokumente müssen weitergeleitet werden
- fremde Daten referenzieren die eigenen Daten:
  - Restrukturierung, Aufteilung
  - Beibehaltung desselben externen Schemas

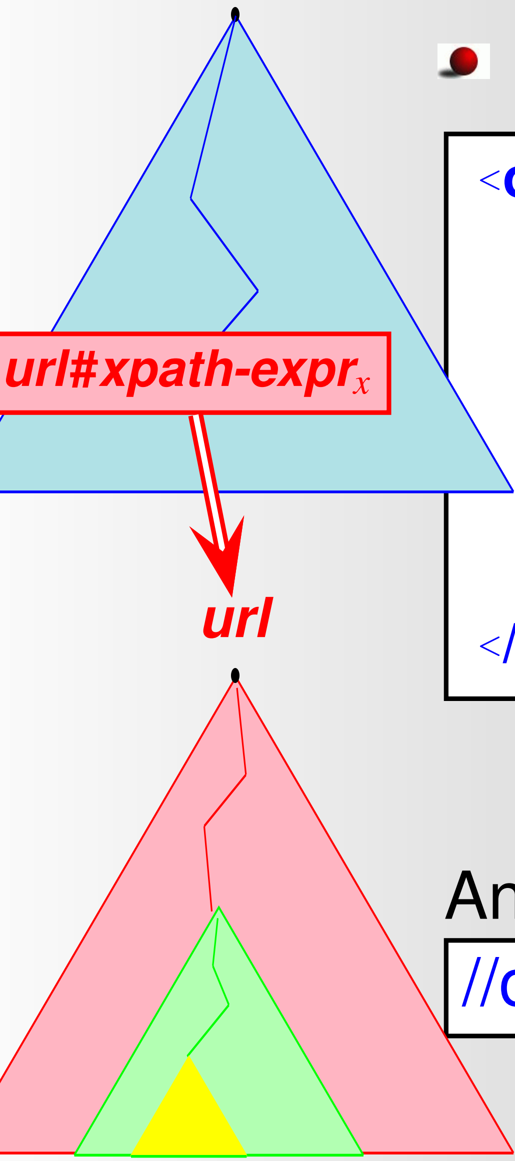
Datenmodelle für Referenzen

- Erweiterung des abstrakten Datenmodells um Referenzen
  - erfordert einen expliziten Navigationsoperator und einen Zwischenschritt

    organization[@abbrev="EU"]/headq/@href$\rightsquigarrow$city/name

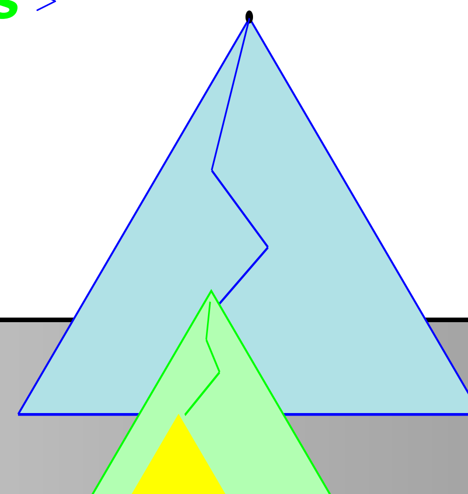
- Logisches Datenmodell mit transparenten Referenzen:
  - Jedes XLink wird als View-Definition gesehen
    - Eingebettete Views als implizite Teilbäume
    - Integration der externen Schemata
    - Anfragen mit üblicher XPath-Syntax/Semantik

Logisches Modell: Transparente Links



- Link-Elemente als *transparent* ansehen

```
<organization abbrev="EU">
   <headq xlink:type="simple"
             href="file:cities-B.xml#//city[name='Brussels']"
             attributes of Brussels >
        contents of Brussels
   </headq>
   :
</organization>
```



Anfrage:

```
//organization[@abbrev="EU"]/headq/name
```

Logisches Datenmodell

- XLink-Spracherweiterung zur Spezifikation des gewünschten logischen Datenmodells:
  dbxlink:transparent-Attribut:

Zusammensetzen der Bäume

- Beitrag der XLink-Elemente
  - Elementhüllen der XLink-Elemente behalten oder weglassen
    keep-element, drop-element
  - in Referenzattribute umwandeln – make-attribute
  - nur Attribute behalten und durchgeben – keep-attributes
- Beitrag der referenzierten Elemente:
  - Elemente als ganzes einfügen – insert-elements
  - nur Elementinhalt (und Attribute) in vorhandene Hülle einfügen – insert-contents
- Namen für from/to bei arc-Elementen

Anwendung des logischen Modells

- transparente Integration von *View-Definitionen* in die XML-Datenbank
- Anfragen an verlinkte XML-Dokumente im Web
- Aufsplitten eines XML-Dokuments in mehrere Dokumente Beibehaltung desselben logischen Schemas.

Realisierung

(Diplomarbeit Dimitrio Malheiro)

- Materialisierung der logischen Instanz:
  - Validierung und Veranschaulichung des Ansatzes
  - Erstellung der logischen Instanz zu einem gegebenen Einstiegsdokument durch ein XSLT-Skript
  - Problem bei unendlicher Rekursionstiefe
  - http://www.informatik.uni-freiburg.de/˜malheiro/demo
- Auswertung von Anfragen gegen das logische Modell
  - Erweiterung des LoPiX-Systems
  - keine Materialisierung
  - Zerlegung der Anfragen
  - on-demand-Zugriff auf externe Quellen
  - http://www.informatik.uni-freiburg.de/˜may/lopix

Auswertungsaspekte

Definition von weiteren Attributen im dbxlink-Namespace zur Auswertung von Anfragen in diesem Modell:

- Wann wird ein Link ausgewertet?
  (Parsing-Zeitpunkt oder Auswertungszeitpunkt)
  dbxlink:actuate-Attribut
- Wo findet die Berechnung statt?
  (lokal oder bei dem referenzierten Server)
  dbxlink:eval-Attribut
- welche (Zwischen)ergebnisse werden gespeichert und wiederverwendet.
  (Dokument, Link-Ergebnis, Anfrageergebnis, Nichts)
  dbxlink:cache-Attribut

Fortsetzung im Rahmen des DFG-Projektes “LinXIS”.
http://www.informatik.uni-freiburg.de/˜may/LinXIS

Fragen ??