



# Dokumentgrammatiken und die Semantik generischer Dokumentstrukturen

Henning Lobin

Justus-Liebig-Universität Gießen

DFG-Rundespräch „Semistrukturierte Daten“  
21./22.2.2002

# Einleitung



# XML-basierte Informationsmodellierung – Status Quo

- Entwicklung verschiedener Schema-Sprachen in letzter Zeit
- Motivation durch Anwendungserfordernisse
  - ◆ Validierung und Parsing
  - ◆ Editierung und Transformierung
  - ◆ Information Retrieval und Querying
- Motivation auch das Ziel, den formalen Spielraum von Markup-Sprachen besser auszunutzen
- Bedarf nach formaler Modellierung:
  - ◆ Grammatik
  - ◆ Datenstrukturen
  - ◆ Domäne

# Entwicklung neuer Schemasprachen

- Grundlage: SGML/XML-DTDs (ISO/W3C 1986-1998)
- XML Schema (W3C; 1999-2001)

```
<!ELEMENT doc (title, section*)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT section (para+)>
<!ELEMENT para (#PCDATA)>

<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified">
  <xsd:element name="doc">
    <xsd:complexType>
      <xsd:sequence>
        <xsd:element ref="title"/>
        <xsd:element name="section" type="sectionType"
          minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
    </xsd:complexType>
  </xsd:element>
  <xsd:element name="para" type="xsd:string"/>
  <xsd:complexType name="sectionType">
    <xsd:sequence>
      <xsd:element ref="para" maxOccurs="unbounded"/>
    </xsd:sequence>
  </xsd:complexType>
  <xsd:element name="title" type="xsd:string"/>
</xsd:schema>
```

# Entwicklung neuer Schemasprachen

- Grundlage: SGML/XML-DTDs (ISO/W3C 1986-1998)
- XML Schema (W3C; 1999-2001)
- Proprietäre Entwicklungen: XDR, SOX, DSD, ...
- Schematron (Jelliffe 1999-2000)
- RELAX NG (= RELAX + TREX; ISO/OASIS 1999-2000, Murata/Clark 2001)

```
<grammar>
<start>
  <element name="html">
    <zeroOrMore>
      <ref name="section"/>
    </zeroOrMore>
  </element>
</start>
<define name="section">
  <element name="div">
    <attribute name="class">
      <value>section</value>
    </attribute>
    <zeroOrMore>
      <element name="para">
        <text/>
      </element>
    </zeroOrMore>
  </element>
</define>
<define name="subsection">
  <element name="div">
    <attribute name="class">
      <value>subsection</value>
    </attribute>
    <zeroOrMore>
      <element name="para">
        <text/>
      </element>
    </zeroOrMore>
  </element>
</define>
</grammar>
```

# Entwicklung neuer Schemasprachen

- Grundlage: SGML/XML-DTDs (ISO/W3C 1986-1998)
- XML Schema (W3C; 1999-2001)
- Proprietäre Entwicklungen: XDR, SOX, DSD, ...
- Schematron (Jelliffe 1999-2000)
- RELAX NG (= RELAX + TREX; ISO/OASIS 1999-2000, Murata/Clark 2001)

# Ansätze

- Untersuchung der formalen Eigenschaften von Markup-Sprachen: Markup-Sprachen als formale Grammatiken
- Formale Datenmodellierung für Dokumentgrammatiken: Ansätze unter Verwendung von UML, Entity-Relationship-Modellen, Prädikatenlogik und Semantik von Programmiersprachen
- Domänenmodellierung für Dokumentklassen: KI-Methoden (Wissensrepräsentation, Logik, Inferenzsysteme) und proprietäre Ansätze

# 1. Formale Eigenschaften von Markup-Sprachen

# Untersuchung von Schema-Sprachen als formale Grammatiken

- Berstel/Boasson (2000), Lee/Chu (2000), Murata/Lee/Mani (2000)
- XML-basierte Markup-Sprachen gehören zur Klasse der regulären Baum-Sprachen bzw. –Grammatiken (Murata 1998)
  - ◆ Kontextfreie Grammatik:  $X \rightarrow Y_1 \dots Y_n$
  - ◆ Baum-Grammatik:  $X \rightarrow \langle a \rangle Y_1 \dots Y_n \langle /a \rangle$
- drei Einschränkungen mit absteigender Mächtigkeit:
  - ◆ Local Tree Grammars (z.B. DTD):  $X \rightarrow \langle X \rangle Y_1 \dots Y_n \langle /X \rangle$
  - ◆ Single-Type Grammars (z.B. XML Schema)
  - ◆ Restrained Competition Grammars
- Schema-Sprache für Regular Tree Grammars: z.B. RELAX
 

```

<!ELEMENT      section      (PARA-IN-SECTION) * >
<!ELEMENT      footnote     (PARA-IN-FOOTNOTE) * >
<!RULE  PARA-IN-SECTION    para    (#PCDATA|footnote) * >
<!RULE  PARA-IN-FOOTNOTE  para    (#PCDATA) >
      
```

## 2. Formale Datenmodellierung



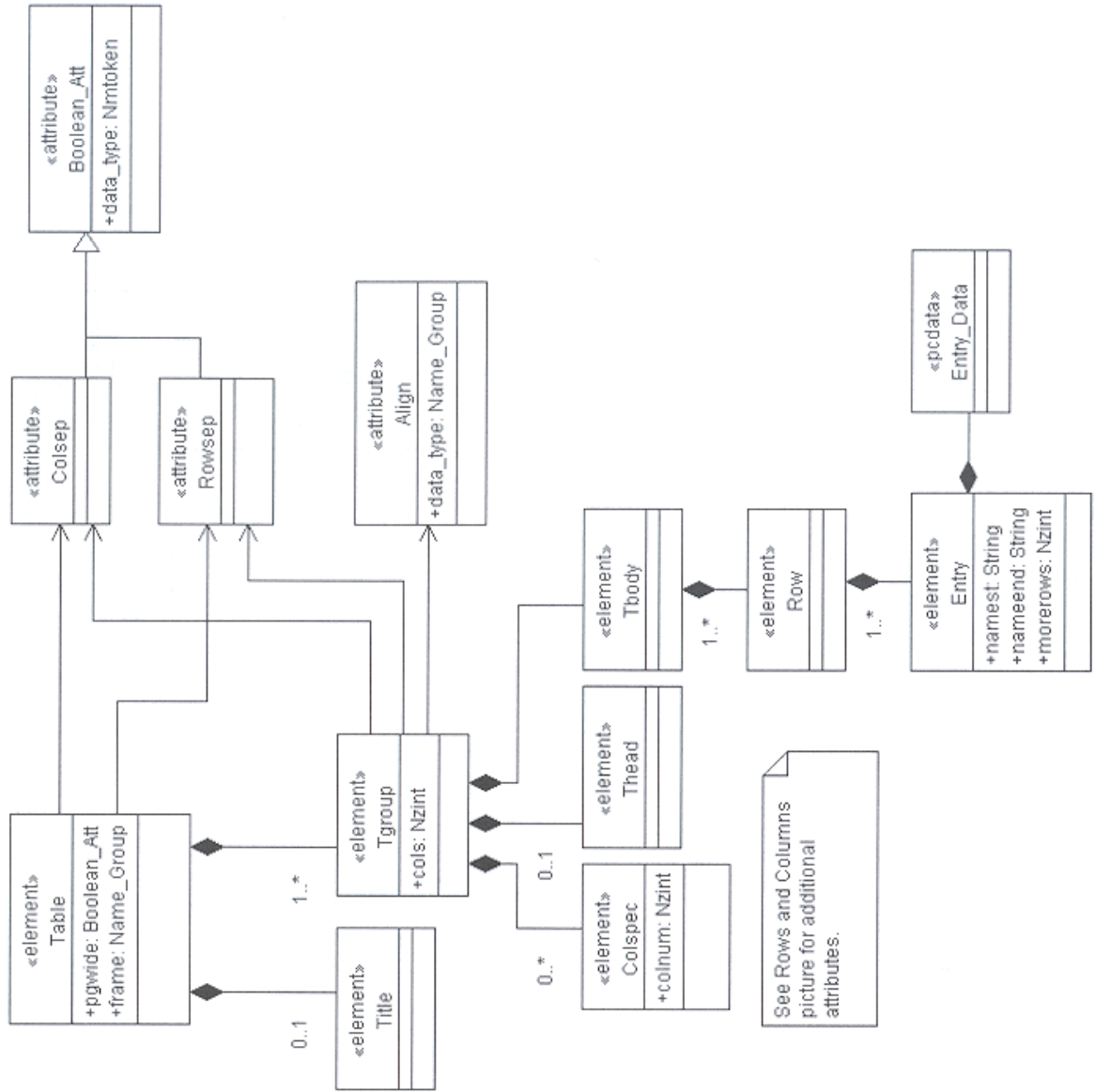
# Formale Datenmodellierung – Idee

- Bezugnahme auf etablierte Modelle der formalen Datenmodellierung
- Motivation: durch formale Datenmodellierung sollen die in der grammatischen Datenmodellierung impliziten Abhängigkeiten explizit gemacht werden
- Schaffung weitergehender Retrieval- und Transformationsprozesse unterhalb der Ebene der semantischen Interpretation der Daten.
- das in der Dokument-Grammatik kodierte Strukturwissen wird präzisiert, wobei z.B. von Reihenfolgebeziehungen abstrahiert wird.



# Formale Datenmodellierung – Ansätze

- Prädikatenlogik (Ramalho et al. 1999)
- UML (Kimber/Heintz 2001)





# Formale Datenmodellierung – Ansätze

- Prädikatenlogik (Ramalho et al. 1999)
- UML (Kimber/Heintz 2001)
- Entity-Relationship-Modellierung  
(Mani/Lee/Muntz 2001)
- Semantik von Programmiersprachen  
(Brown et al. 2001)

# 3. Domänenmodellierung

# Domänenmodellierung – Idee

- Kombination der Dokumentgrammatik mit einem formalen Modell des Gegenstandsbereichs
- Motivation: das in der Dokument-Grammatik ausschnittsweise enthaltene Wissen soll möglichst umfassend modelliert werden
- Berücksichtigung auch solcher Aspekte, die in der Grammatik überhaupt nicht ausgedrückt werden
- Domänenmodell ist unabhängig von einer bestimmten Dokumentgrammatik



# Domänenmodellierung – Ansätze

- Frame-basierte Modelle (z.B. Welty/Ide 1999)
- Prädikatenlogik/Prolog (Sperberg-McQueen/Huitfeldt/Renear 2001)

```
node([1], element(tei2)).
...
node([1, 5, 2], element(p)).
node([1, 5, 2, 1], element(del)).
node([1, 5, 2, 1, 1], element('I')).
node([1, 5, 2, 1, 2], element('t')).
...
attr([1], lang, 'de').
...

infer(Property, Loc) :- node(Loc, element(Property)).
infer(Property, Loc) :-
    node(Anc, element(Property))
    descendant(Loc, Anc).

?- infer(Property, [1, 5, 2]).
Property = p ->;
Property = lang('de') ->;
no
?-
```



# Domänenmodellierung – Ansätze

- Frame-basierte Modelle (z.B. Welty/Ide 1999)
- Prädikatenlogik/Prolog (Sperberg-McQueen/Huitfeldt/Renear 2001)
- Schema Adjuncts (Vorthmann/Robie 2001)
- Ontologien (z.B. Erdmann/Studer 2001)
- **Semantische Netze**

# Domänenmodellierung – Anwendungsmöglichkeiten

- **Erstellung**
  - semantische Validierung
  - Unterstützung bei der Daten-Editierung
  - Textkategorisierung, Textparsing
  
- **Bearbeitung**
  - automatische Transformation
  
- **Retrieval**
  - Auffinden impliziter Datenstrukturen
  - Erkennung semantisch validierbarer Datenstrukturen

# Das Projekt „Semantik generischer Dokumentstrukturen“

## DFG-Projekt: „Semantik generischer Dokumentstrukturen“

- Teil der Forschergruppe „Texttechnologische Informationsmodellierung“ gemeinsam mit Bielefeld, Tübingen und Dortmund (2001-2004)
- Fokussierung Dokument-orientierter Verwendung von XML (vs. Daten-orientierte Verwendung)
- Systematische Untersuchung formaler semantischer Spezifikationstechniken von Dokumentgrammatiken
- Inventar von semantischen Spezifikationen für Standard-Dokumenttypen
- Übertragung von Verfahren und Konzepten aus der linguistischen Grammatiktheorie

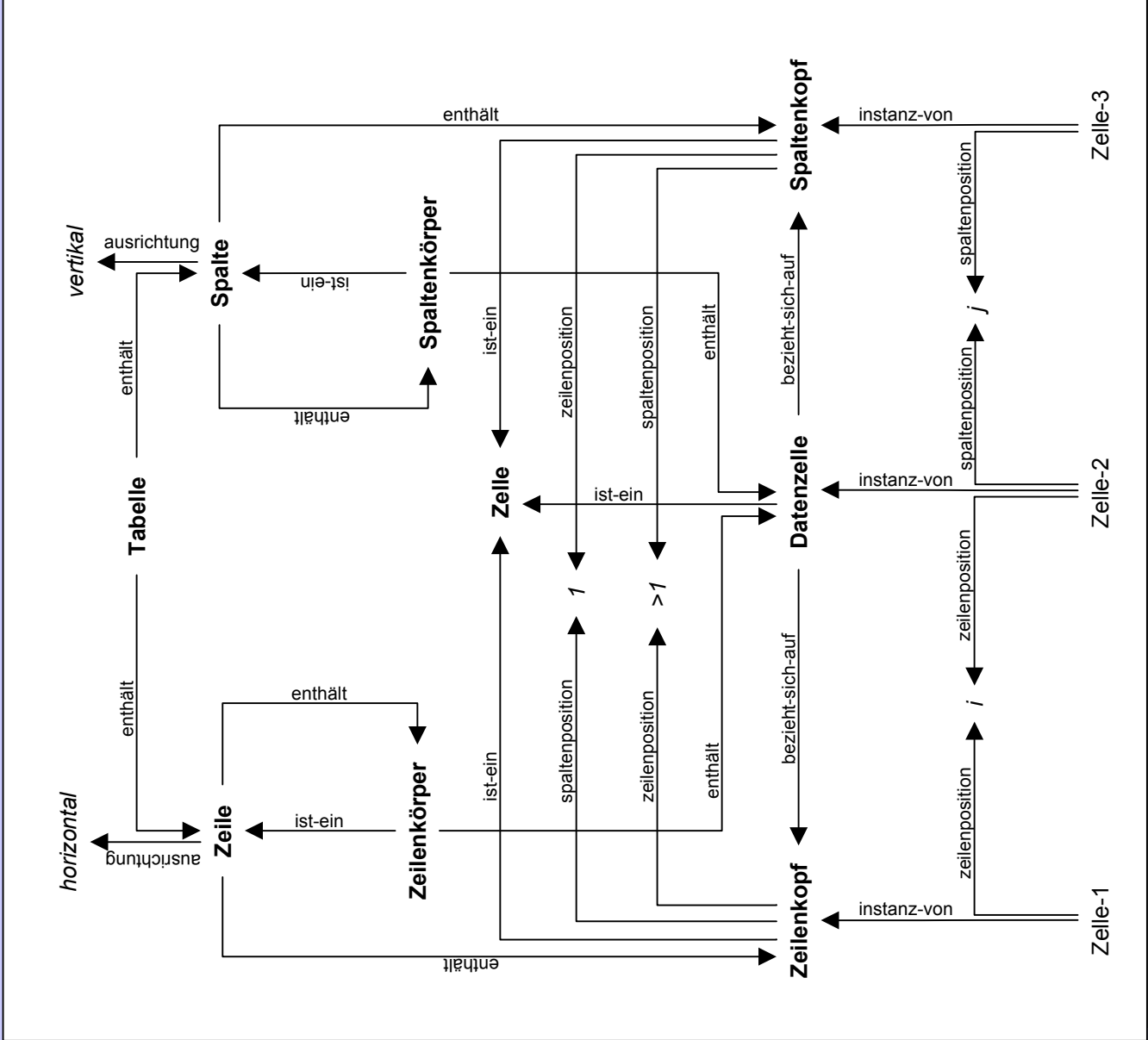
# Problem: Implizite Bedeutungen

	<b>Hamburg</b>	<b>Berlin</b>	<b>München</b>
<b>Bielefeld</b>	253,6	384,4	593,5
<b>Gießen</b>	442,0	468,2	426,8
<b>Herrsching</b>	774,0	614,1	47,0

```
<table>
<tr>
  <td></td>
  <td>Hamburg</td>
  <td>Berlin</td>
  <td>München</td>
</tr>
<tr>
  <td>Bielefeld</td>
  <td>253,6</td>
  <td>384,4</td>
  <td>593,5</td>
</tr>
<tr>
  <td>Gießen</td>
  <td>442,0</td>
  <td>468,2</td>
  <td>426,8</td>
</tr>
<tr>
  <td>Herrsching</td>
  <td>774,0</td>
  <td>614,1</td>
  <td>47,0</td>
</tr>
</table>
```

tr: table row  
td: table data

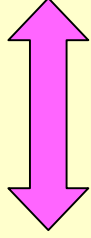
# Beschreibungsansatz: Semantische Netze



# Verbindung von Semantischen Netzwerken und XML-Strukturen

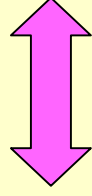
## Dokument-Grammatik

```
<!ELEMENT table (tr*)>
<!ELEMENT tr (td*)>
<!ELEMENT td (#PCDATA)>
```

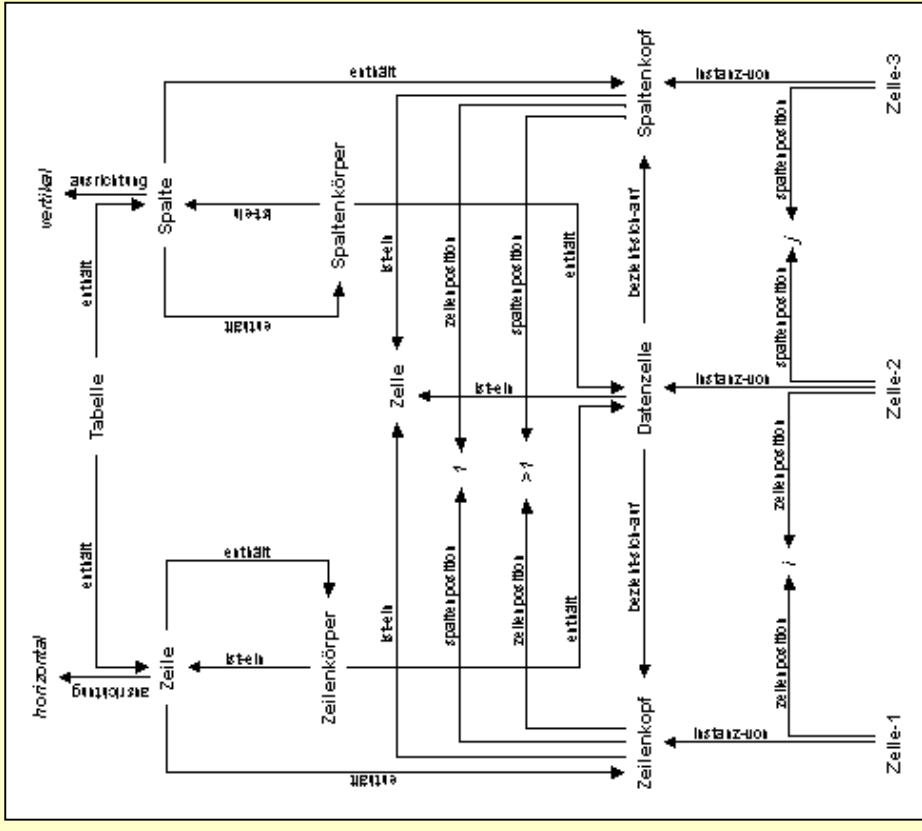


## Dokument-Instanz

```
<table>
<tr>
<td></td><td>Hamburg</td><td>Berlin</td>
</tr>
<tr>
<td>Bielefeld</td><td>253,6</td><td>384,4</td>
</tr>
<tr>
<td>Giessen</td><td>442,0</td><td>468,2</td>
</tr>
</table>
```



## Semantisches Modell der Dokument-Struktur



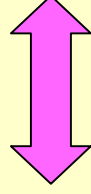
# Adressierung der Dokumentinstanzen

**Tabelle:** table  
**Zelle:** table/tr/td  
**Datenzelle allgemein:** table/tr[position()>1 and position()=<last()>]  
**Datenzelle speziell:** table/tr[position()=i]/td[position()=j]  
**Zeilenkopf:** table/tr[position()=i]/td[position()=1]  
**Spaltenkopf:** table/tr[position()=1]/td[position()=j]  
**Zeile:** table/tr[position()=i]/td  
**Spalte:** table/tr/td[position()=j]  
**Zeilenkörper:** table/tr[position()=i]/td[position()>1 and position()=<last()>]  
**Spaltenkörper:** table/tr[position()>1 and position()=<last()>]/td[position()=j]

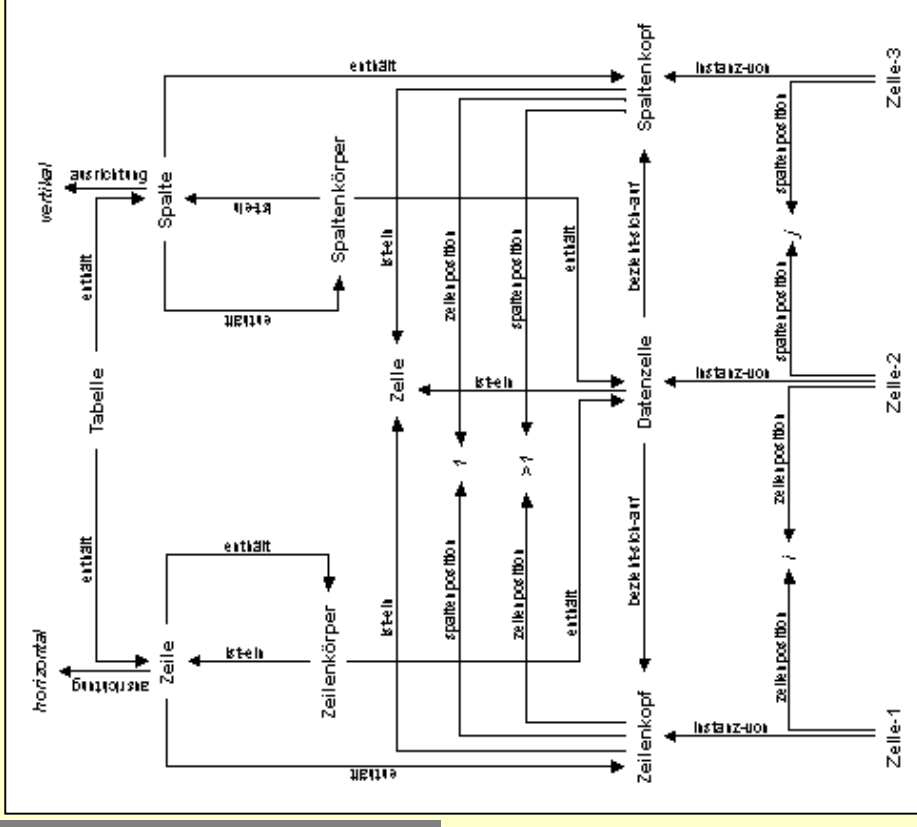
## Dokument-Instanz

```

<table>
<tr>
<td></td><td>Hamburg</td><td>Berlin</td>
</tr>
<tr>
<td>Bielefeld</td><td>253,6</td><td>384,4</td>
</tr>
<tr>
<td>Giessen</td><td>442,0</td><td>468,2</td>
</tr>
</table>
    
```



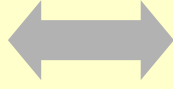
## Semantisches Modell der Dokument-Struktur



# Verbindung von Semantischen Netzwerken und XML-Strukturen

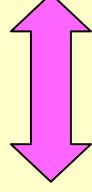
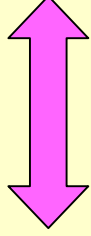
## Dokument-Grammatik

```
<!ELEMENT table (tr*)>
<!ELEMENT tr (td*)>
<!ELEMENT td (#PCDATA)>
```

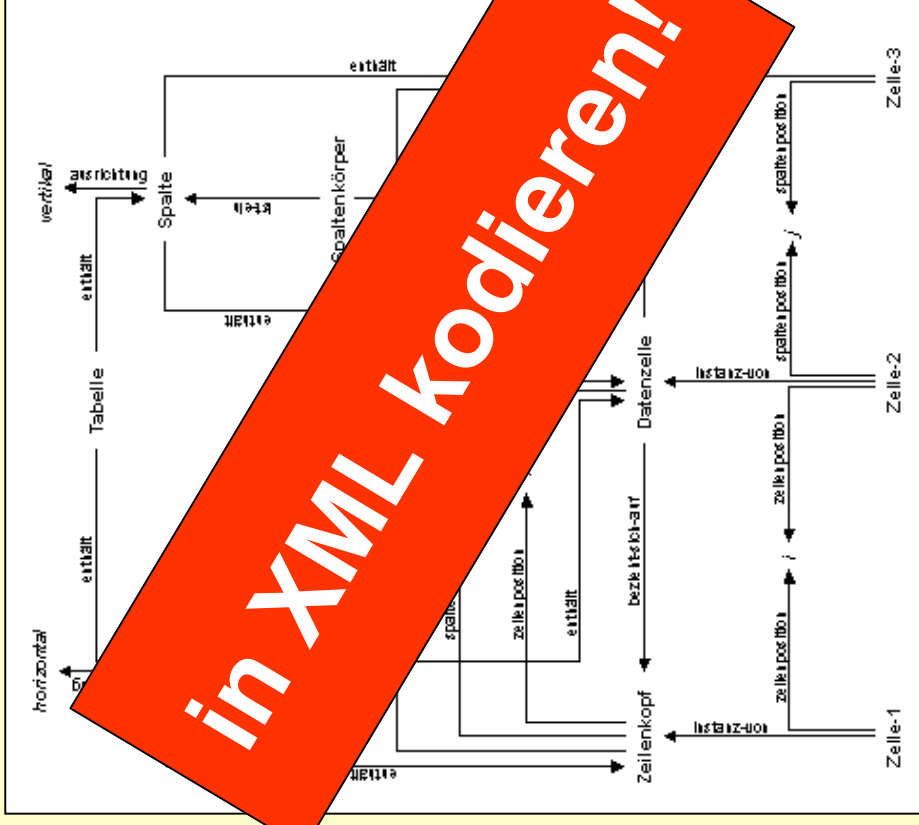


## Dokument-Instanz

```
<table>
<tr>
<td></td><td>Hamburg</td><td>Berlin</td>
</tr>
<tr>
<td>Bielefeld</td><td>253,6</td><td>384,4</td>
</tr>
<tr>
<td>Giessen</td><td>442,0</td><td>468,2</td>
</tr>
</table>
```

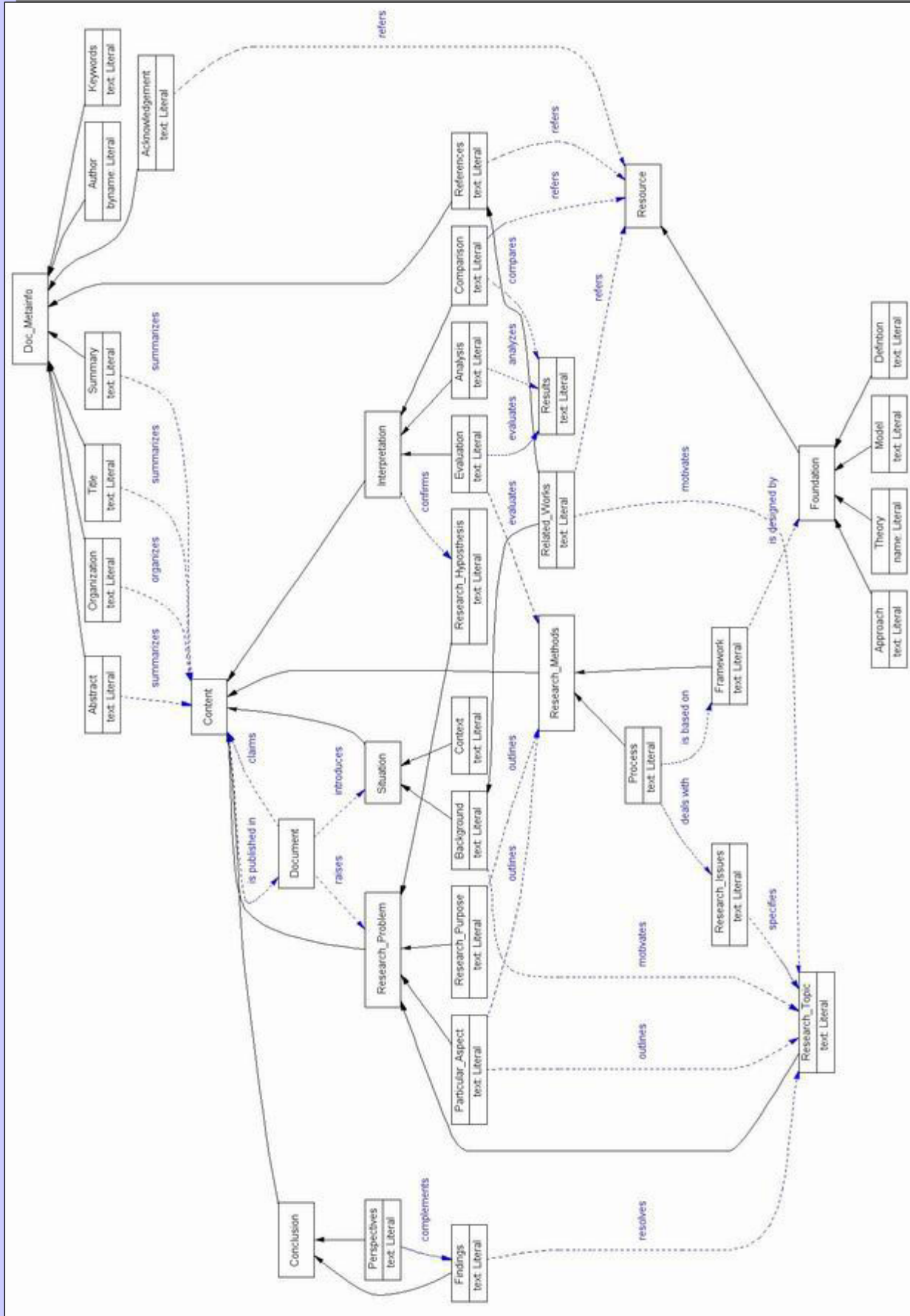


## Semantisches Modell der Dokument-Struktur



**in XML kodieren!**

**Exemplarische  
Modellierungsdomäne:  
Wissenschaftliche Papiere**



# Externe Annotation

```
<article>
<preface>
...
</preface>
<chapter>
  <title>presentation
  schema</title>
  <p>...</p>
  ...
</chapter>
...
</article>
```

*XPath*

*XPointer*

```
<?XML version="1.0"?>
<rdf:RDF
  xmlns:rdf=".....">
  <rdf:Description
    about="...##/child::p[2]">
  <rs:support about="...../...html">
  </rdf:Description about=...>
  <rs:refer about="...../...html">
  </rdf:description>
</rdf:RDF>
```

# Semantische Annotation im Schema-Dokument

```
<schema xmlns="http://www.w3.org/1999/XMLSchema"
  targetNamespace="http://purl.orgdc/elements/1.1"
  ...
  <annotation>
    <documentation>Draft XML schema for Dublin Core Element set
    </documentation>
  </annotation>
  <simpleType name="title"
    xx:semantics="http://http:purl.org/elements/1.1/cdmes.rdf#title">
    <restriction base="string"/>
  </simpleType>
  ...
</schema>
```

# Fazit

- Formale Modellierung erschließt der maschinellen Bearbeitung von XML-Dokumenten Bereiche, die bislang nur menschlichen Bearbeitern zugänglich waren.
- Automatisierte Transformationsprozesse befreien die Informationsmodellierung von der engen Bindung an bestimmte Auszeichnungsvokabularien
- Schemata können zur Optimierung von Anfragen an XML-Dokumente beitragen
- Die Entwicklung entsprechender Tools befindet sich erst in den Anfängen

# Literatur

- Berstel, Jean und Luc Boasson (2000): "Formal Properties of XML Grammars and Languages".
- Brown, Allen, Matthew Fuchs, Jonathan Robie und Philip Wadler (2001): "MSL. A model for W3C XML Schema". In *Proc. of WWW10*.
- Jelliffe, Rick (2000): *Schematron*. <http://www.ascc.net/xml/resource/schematron>.
- Kimber, Eliot und John Heintz (2001): "Using UML to define XML document types". In *Markup Languages 2/3*, 295-320.
- Lee, Dongwon und Wesley W. Chu (2000): *Comparative Analysis of Six XML Schema Languages*. <http://www.cs.ucla.edu/~dongwon/paper>.
- Lobin, Henning (2001): "Netzwerkorientierte Modellierung der Semantik von Dokumentgrammatiken". In Lobin (Hrsg.), *Sprach- und Texttechnologie in digitalen Medien*. Gießen: GLDV, 141-150
- Mani, Murali, Dongwon Lee und Richard R. Muntz (2001): "Semantic Data Modeling using XML Schemas".
- Murata, Makoto (1998): "Data Model for Document Transformation and Assembly". In *PODDP 1998*.
- Murata, Makoto, Dongwon Lee und Murali Mani (2000): "Taxonomy of XML Schema Languages using Formal Language Theory". In *Proc. of Extreme Markup Languages 2000*.
- Ramalho, José Carlos, Jorge Gustavo Rocha, José João Almeida und Pedro Henriques (1999): "SGML documents: Where does quality go?". In *Markup Languages 1/1*, 75-90.
- [RELAX NG] (2001): *RELAX NG Specification. Committee Specification 11 August 2001*. OASIS – Organization for the Advancement of Structured Information Standards.
- Sperberg-McQueen, Michael, Claus Huitfeld und Allen Renear (2001): "Meaning and interpretation of markup". In *Markup Languages 2/3*, 215-234.
- Vorthman, Scott und Jonathan Robie (2001): "Beyond Schemas". In *Markup Languages 2/3*, 281-294.
- Welty, Christopher und Nancy Die (1999): "Using the Right Tools: Enhancing Retrieval from Marked-up Documents". In *Computers and the Humanities 33*, 59-84.
- [XML] (1998): *Extensible Markup Language (XML) Version 1.0*. World Wide Web Consortium. [Recommendation]
- [XML Schema] (2001): XML Schema. World Wide Web Consortium. [drei Teile, Recommendation]