Autoren: E. Bertoni, G. Guerrini, M. Mesiti, I. Rivara, C. Tavella

MEASURING THE STRUCTURAL SIMILARITY AMONG XML DOCUMENTS AND DTDS

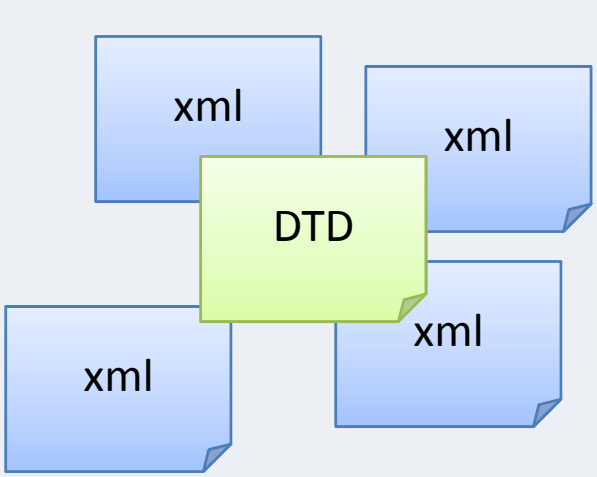
<Paper 1>

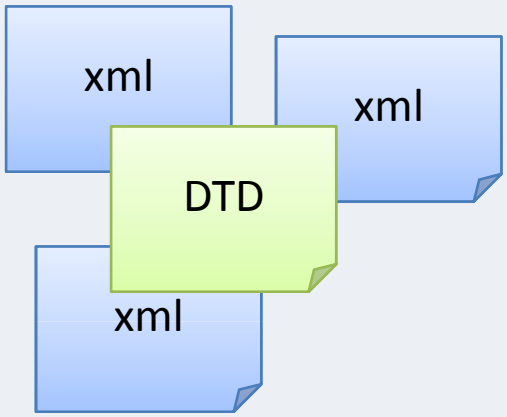
Volker Grabsch und Christine Janischek

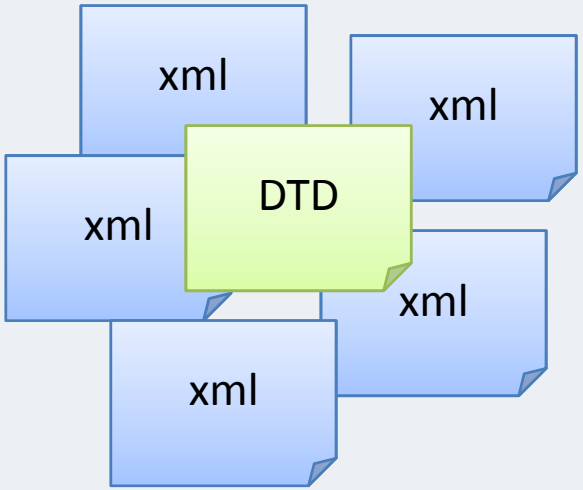
Übersicht

1. Einführung
2. Grundbegriffe
3. Algorithmen
4. Schwachstelle

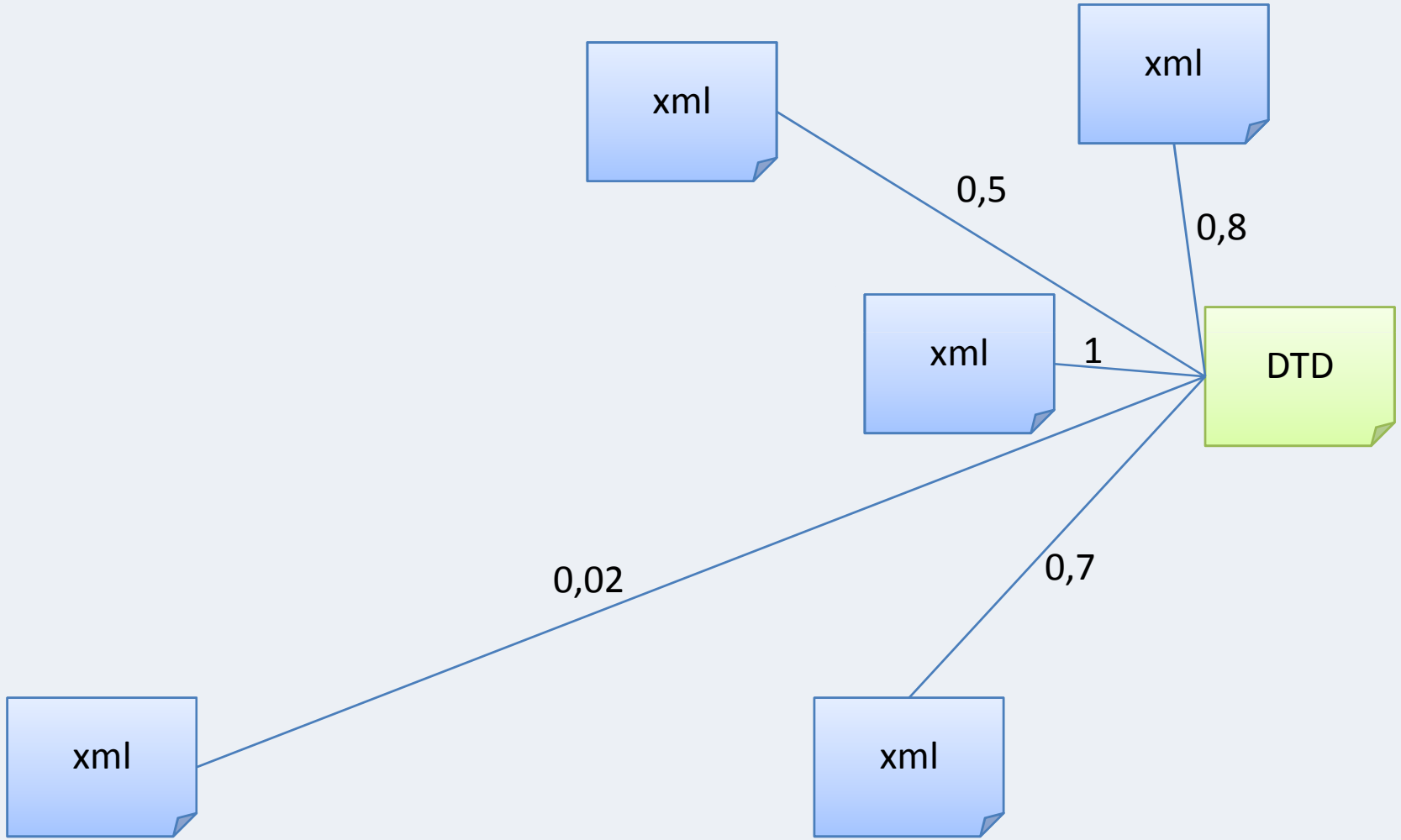
Welches Dokument gehört zu welcher DTD?







Ähnlichkeitsmaß zwischen Dokument und DTD



Vereinfachungen

- Ignorieren der Tag-Reihenfolge (DTD)
- Ignorieren der Textknoten (datenzentriert)

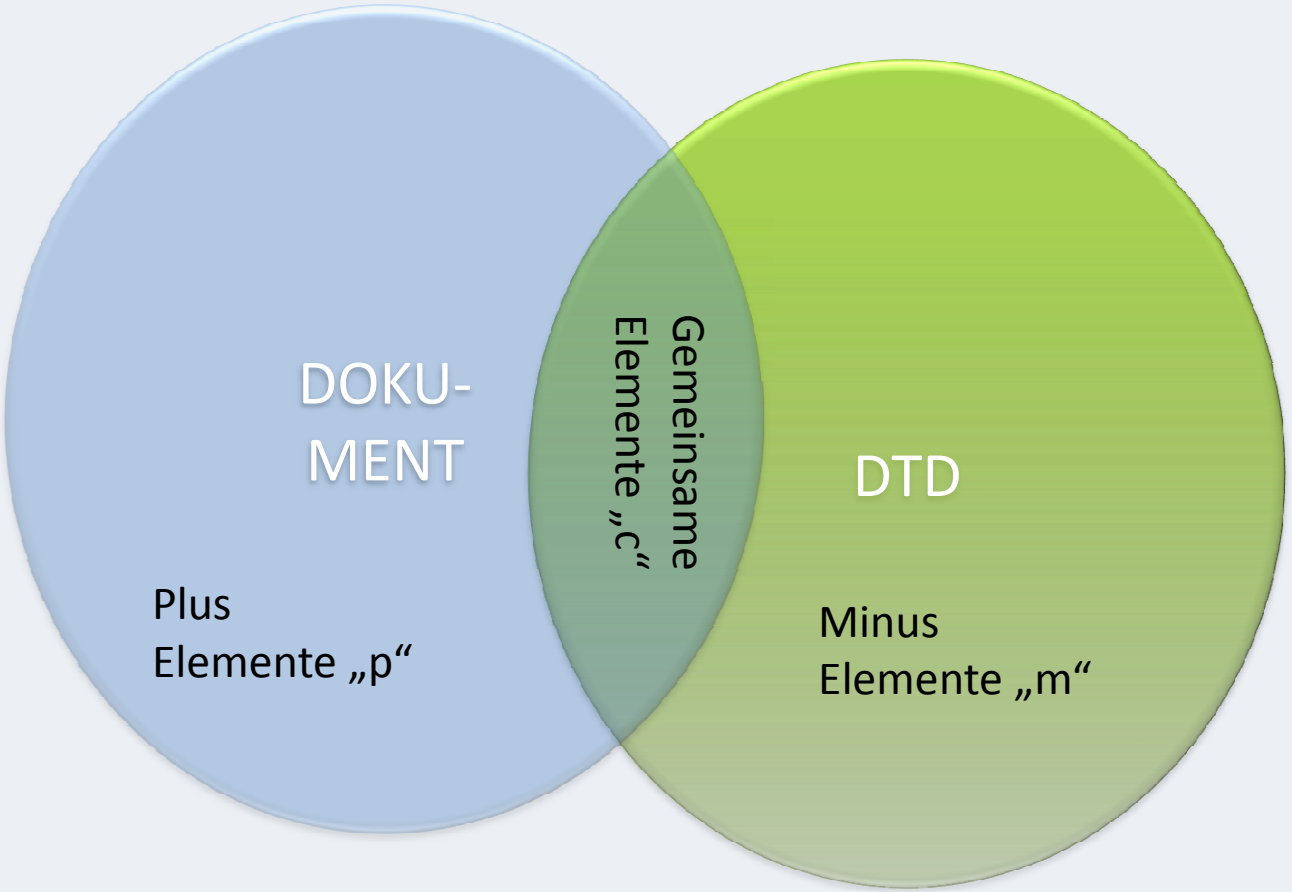
Unkritische Vereinfachungen:

- Verzicht auf den „+“-Operator in der DTD
- Ignorieren der Attributknoten
- Gleiche Semantik bedeutet gleicher Tagname

Übersicht

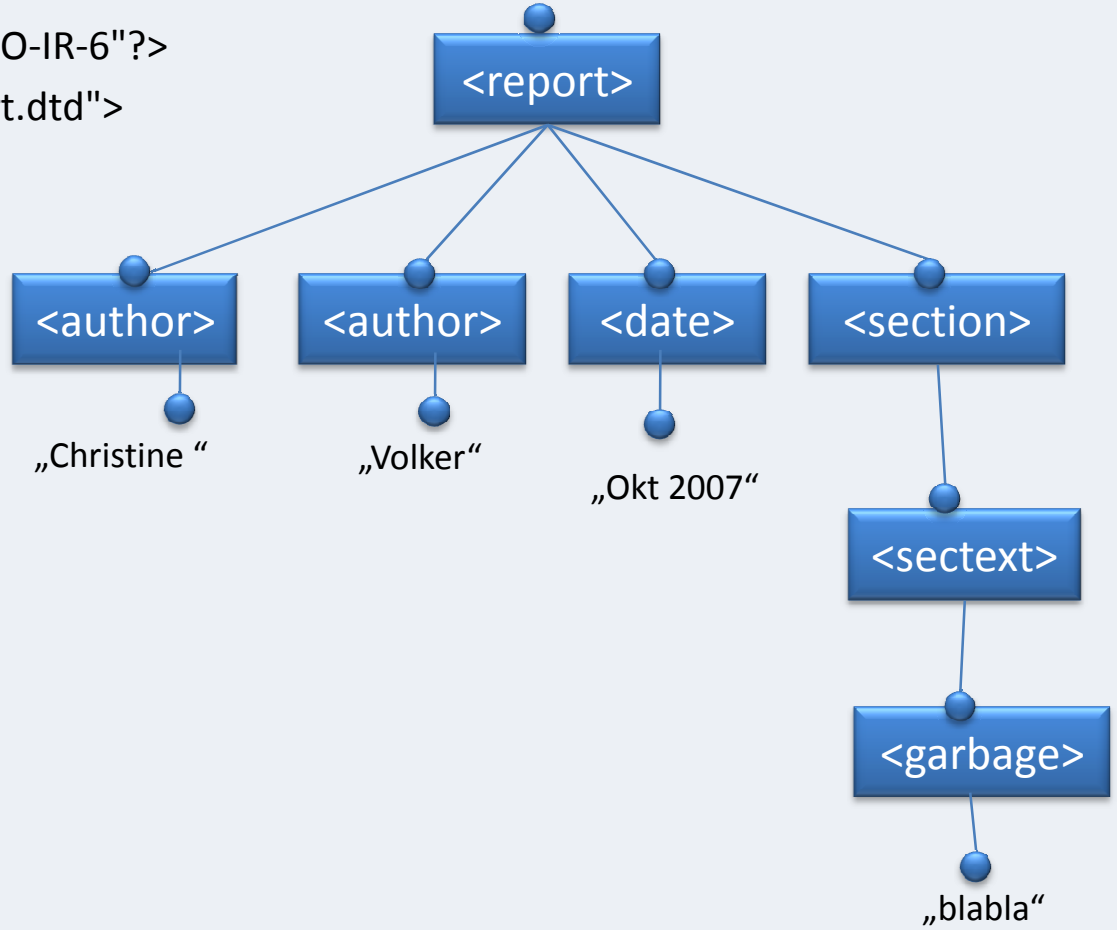
1. Einführung
2. Grundbegriffe
3. Algorithmen
4. Schwachstelle

Gemeinsame, Plus-, und Minus-Elemente



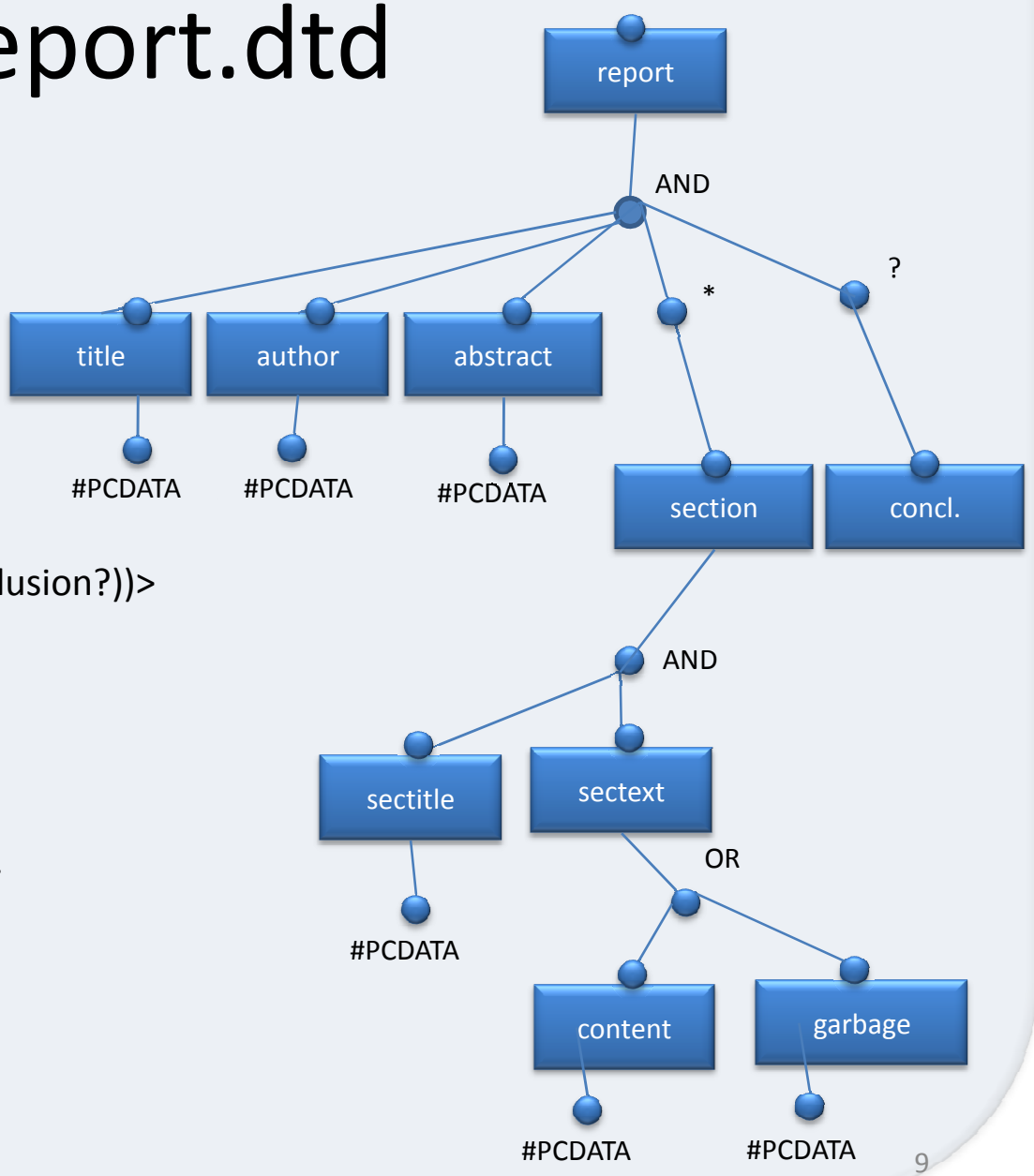
Report.xml

```
<?xml version="1.0" encoding="ISO-IR-6"?>
<!DOCTYPE report SYSTEM "report.dtd">
<report>
 <title>Titel</title>
 <author>Christine</author>
 <author>Volker</author>
  <date>Okt 2007</date>
  <section>
   <sectext>
     <garbage>blabla</garbage>
   </sectext>
  </section>
</report>
```



Report.dtd

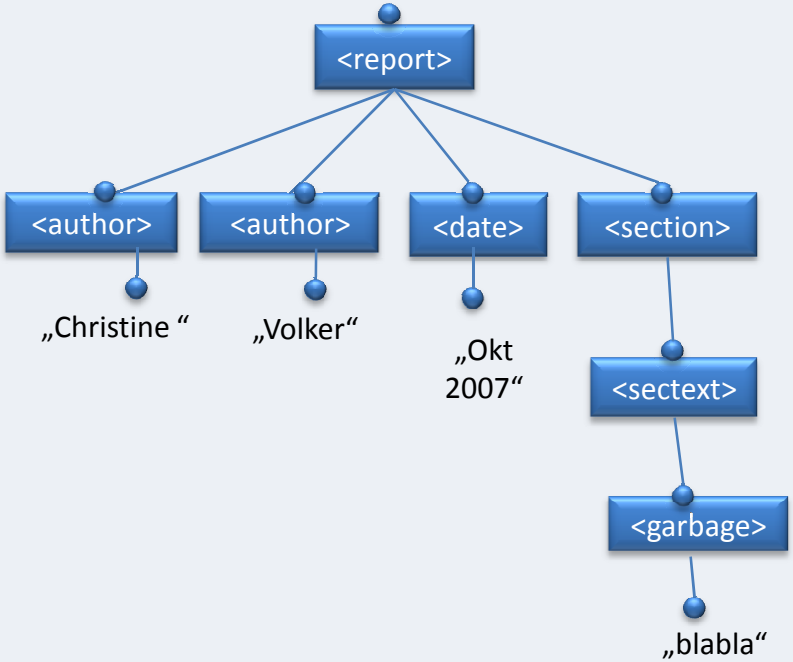
```
<!-- XML DTD "report.dtd" -->

<!ELEMENT report
     (title,author,abstract,(section*),(conclusion?))>
 <!ELEMENT title (#PCDATA)>
 <!ELEMENT author (#PCDATA)>
 <!ELEMENT abstract (#PCDATA)>
 <!ELEMENT section (sectitle,sectext)>
  <!ELEMENT sectitle (#PCDATA)>
  <!ELEMENT sectext (content|garbage)>
     <!ELEMENT content (#PCDATA)>
     <!ELEMENT garbage (#PCDATA)>
 <!ELEMENT conclusion (#PCDATA)>
```

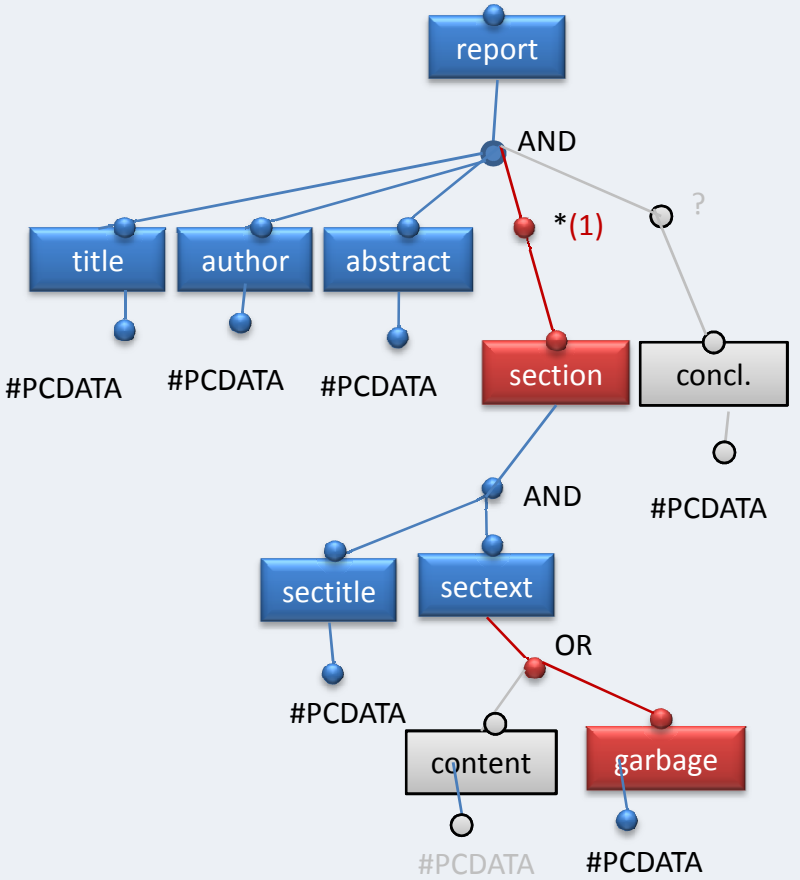


Übersicht

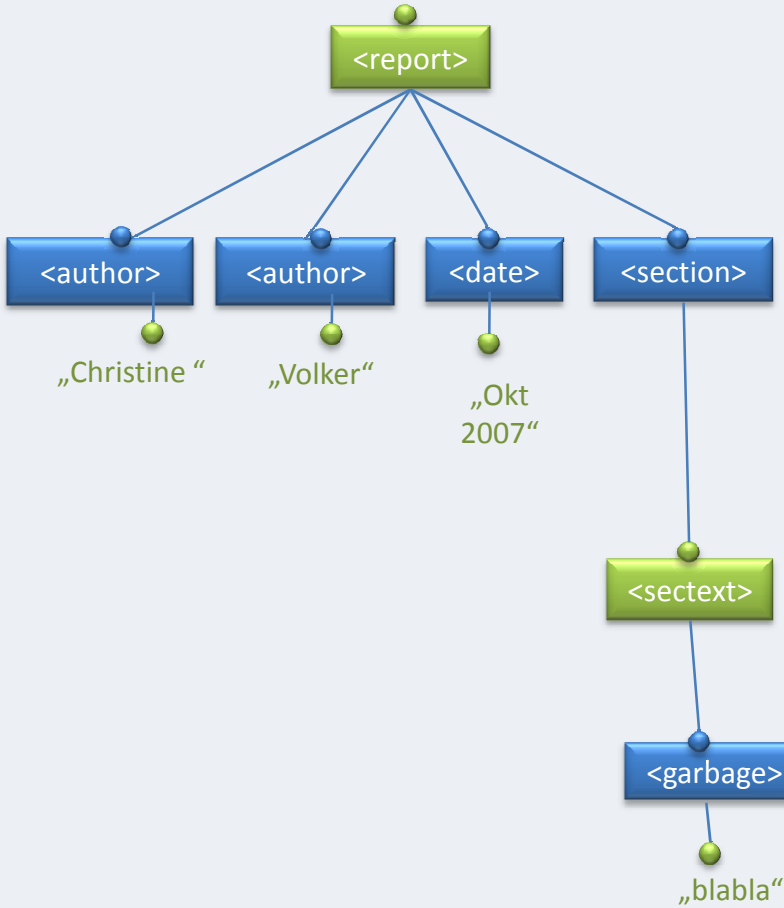
1. Einführung
2. Grundbegriffe
3. Algorithmen
4. Schwachstelle

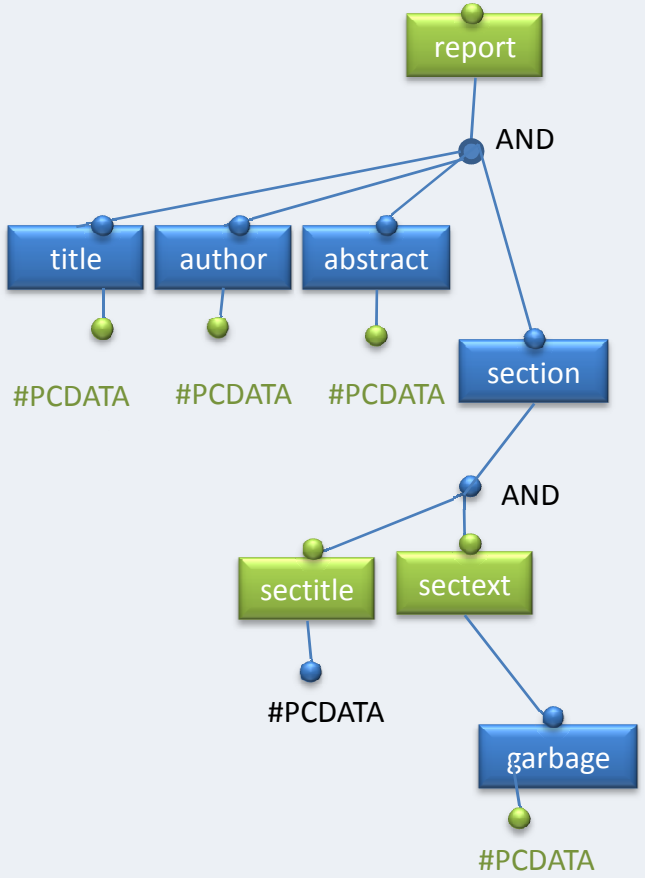
Auswahl der Instanz-DTD





Ähnlichkeit zur Instanz-DTD





(p, m, c)
(0,0,1)
(2,2,2)
(2,3,2)
(0,1,1)
(0,0,1)

Ähnlichkeitsmaß

(p, m, c)	Level Weight (γ=2)	W · (p,m,c)
(0,0,1)	16	(0,0,16)
(2,2,2)	8	(16,16,16)
(2,3,2)	4	(8,12,8)
(0,1,1)	2	(0,2,2)
(0,0,1)	1	(0,0,1)
		(24,30,43)

α,β (≥ 0) sind die Gewichtung der Plus- und Minus-Elemente.

$$\varepsilon((p, m, c)) = \frac{c}{\alpha * p + c + \beta * m}$$

Konvention: $\alpha = \beta = 1$

$\varepsilon((24,30,43)) = 0,44$

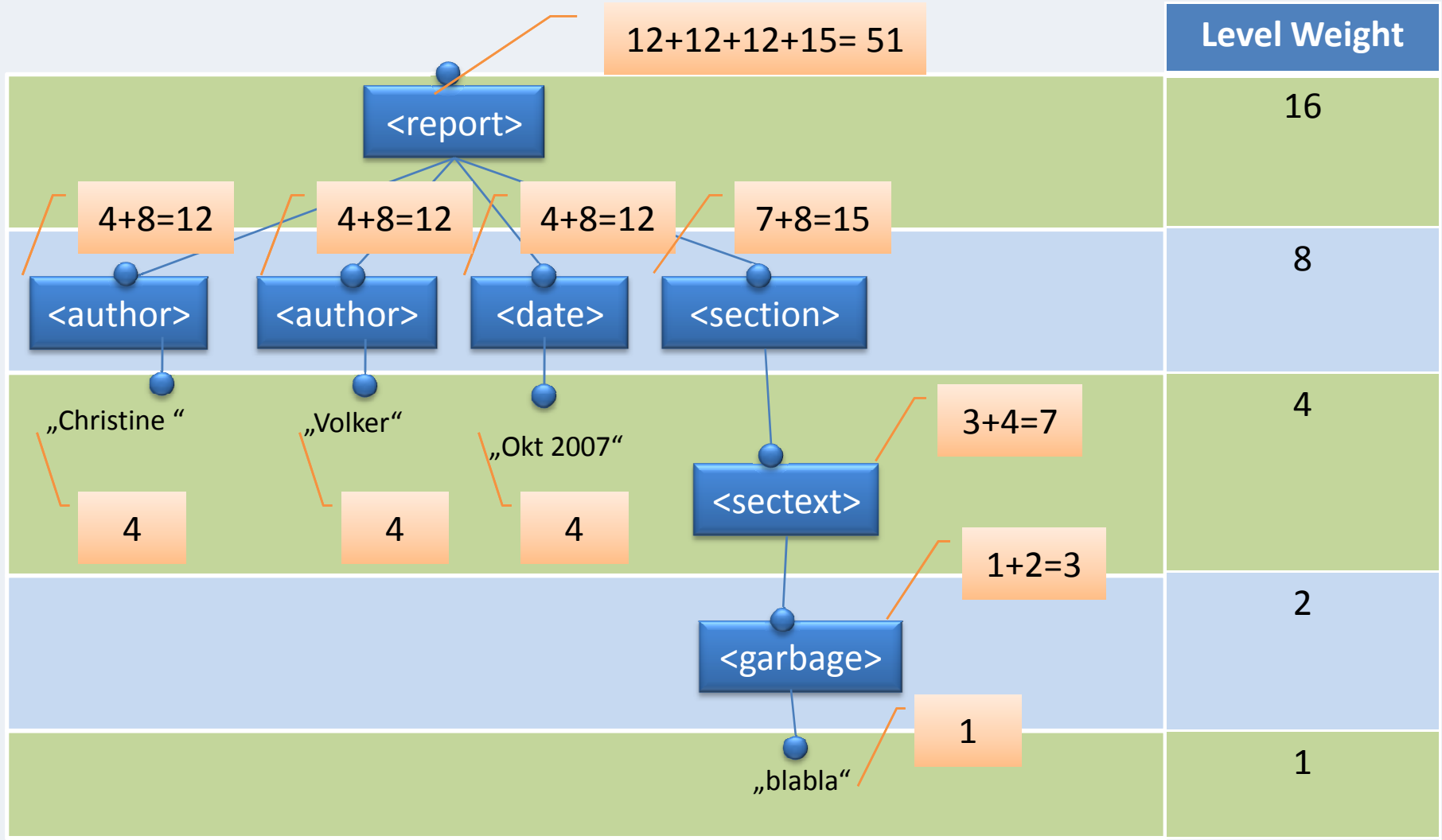
ε liegt immer zwischen 0 und 1.

ε = 1 gilt bei Dokumenten, die mit der DTD bis auf die Tag-Reihenfolge übereinstimmen.

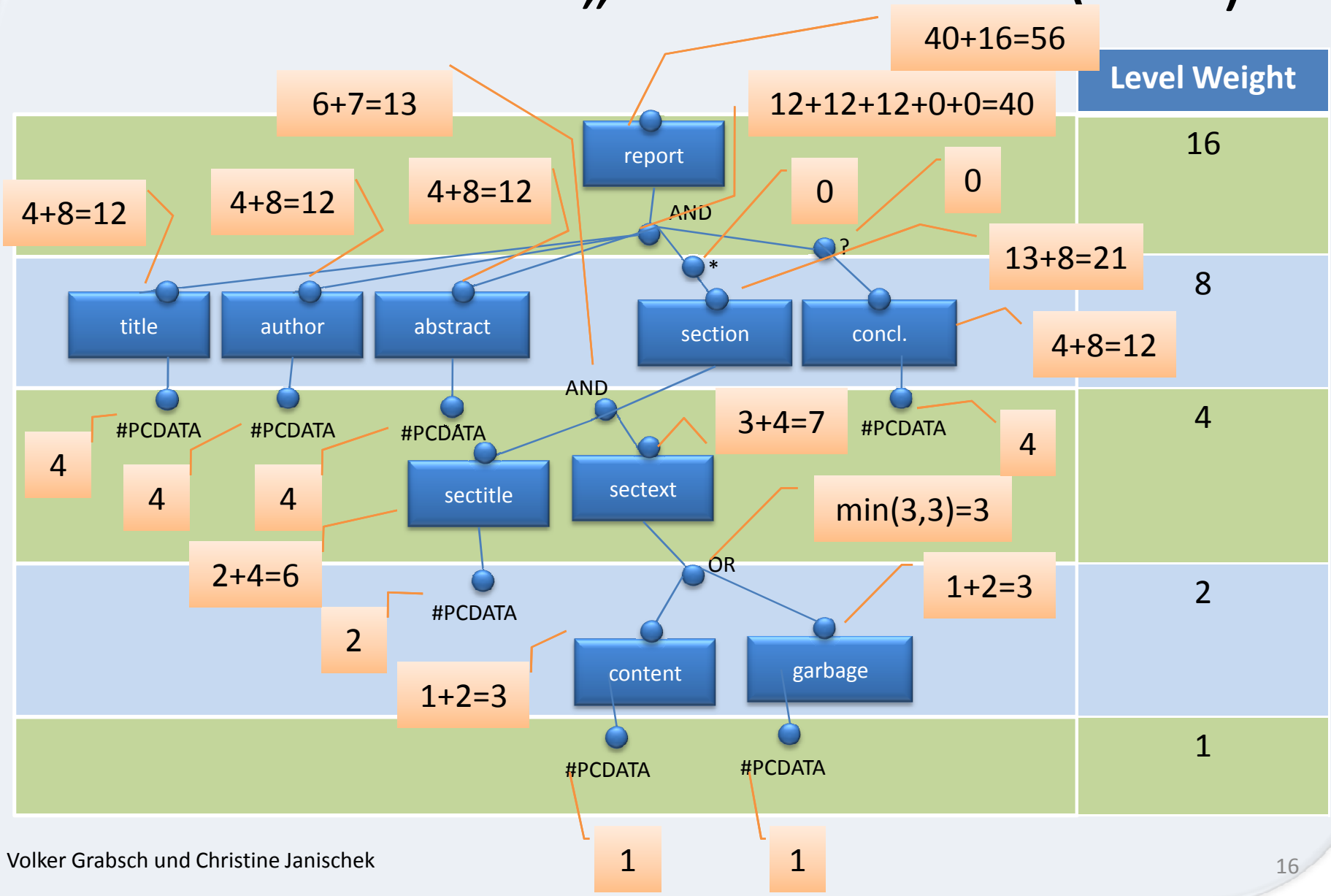
Gewicht eines Teilbaums (W)

- In der „Weight function“ wird die Anzahl der Elemente mit dem jeweiligen Level multipliziert und dann kummuliert.
- Wurzelnahe Elemente sind mehr Wert als entfernte Elemente.
- DTD: Optionale oder wiederholende Elemente werden dabei nicht berücksichtigt
- DTD: Bei „OR“ wird der Teilbaum mit dem kleinsten Gewicht ausgewählt.

Gewicht eines „p“-Teilbaums (XML)



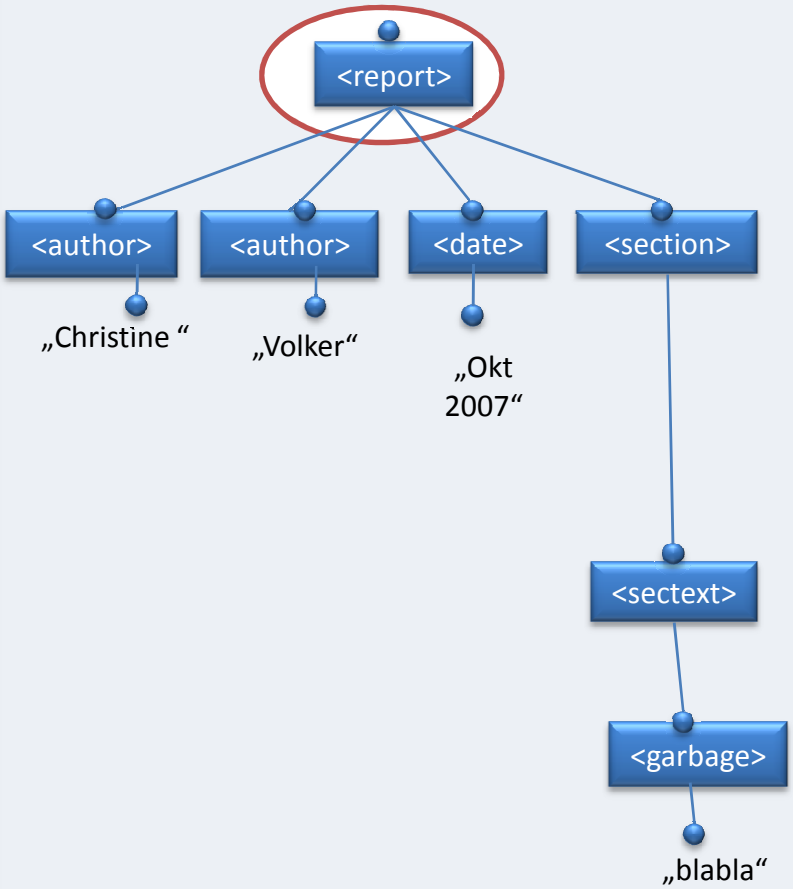
Gewicht eines „m“-Teilbaums (DTD)

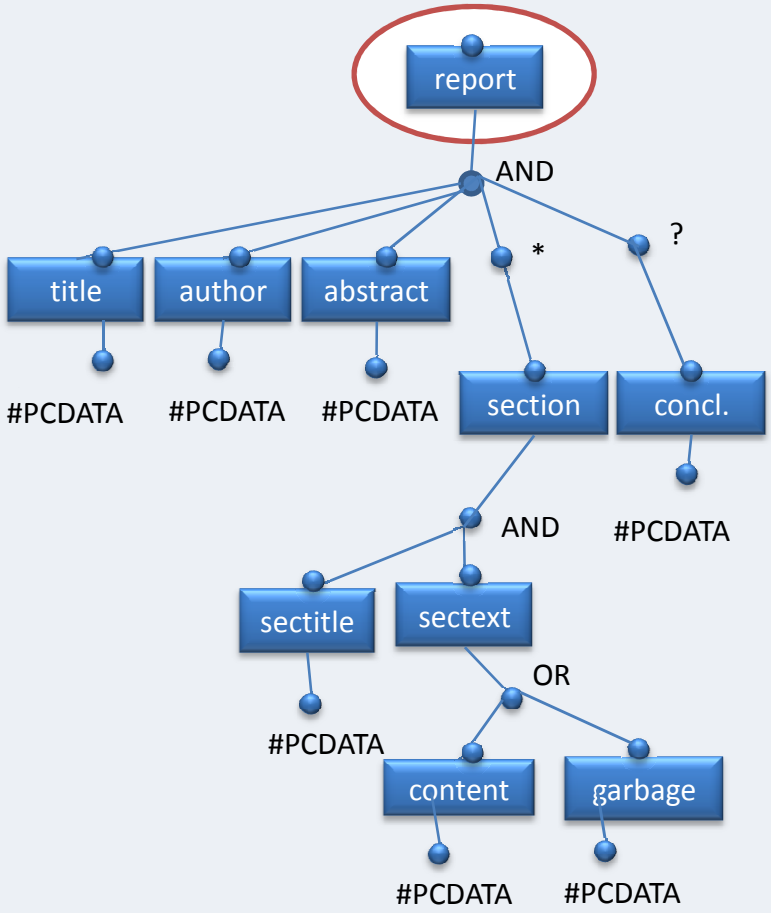


Matching Algorithm (M)

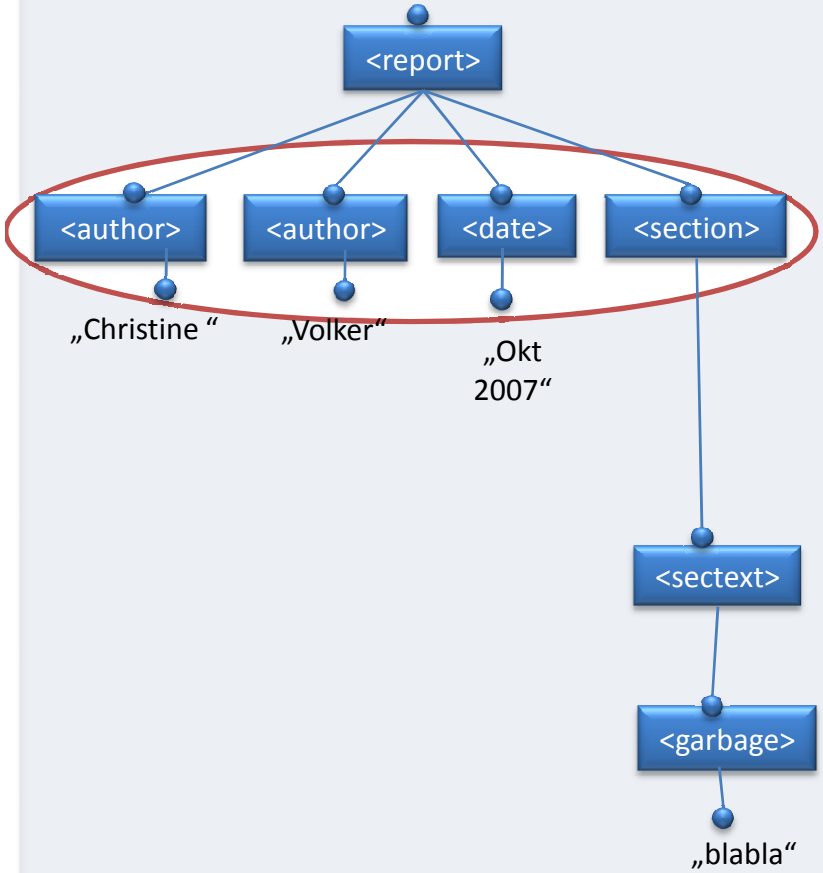
- Naive Lösung: „generate and test“ für alle Instanz-DTDs – nicht praktikabel.
- Besserer Weg:
  - Phase 1: Vom Wurzelelement (root) abwärts , werden die Teile des Baumes rekursiv abgearbeitet.
  - Phase 2: Bei der Rückkehr des rekursiven Aufrufs (backtracking) werden die Alternativen ausgewertet und die Beste ausgewählt.

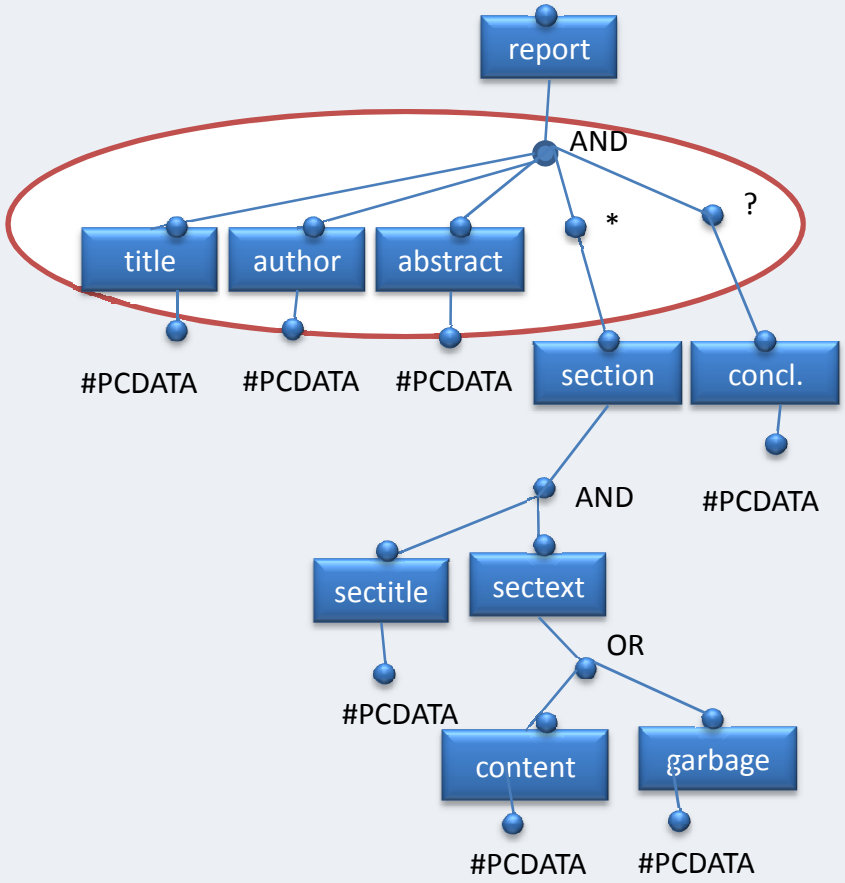
Matching Algorithm (M)



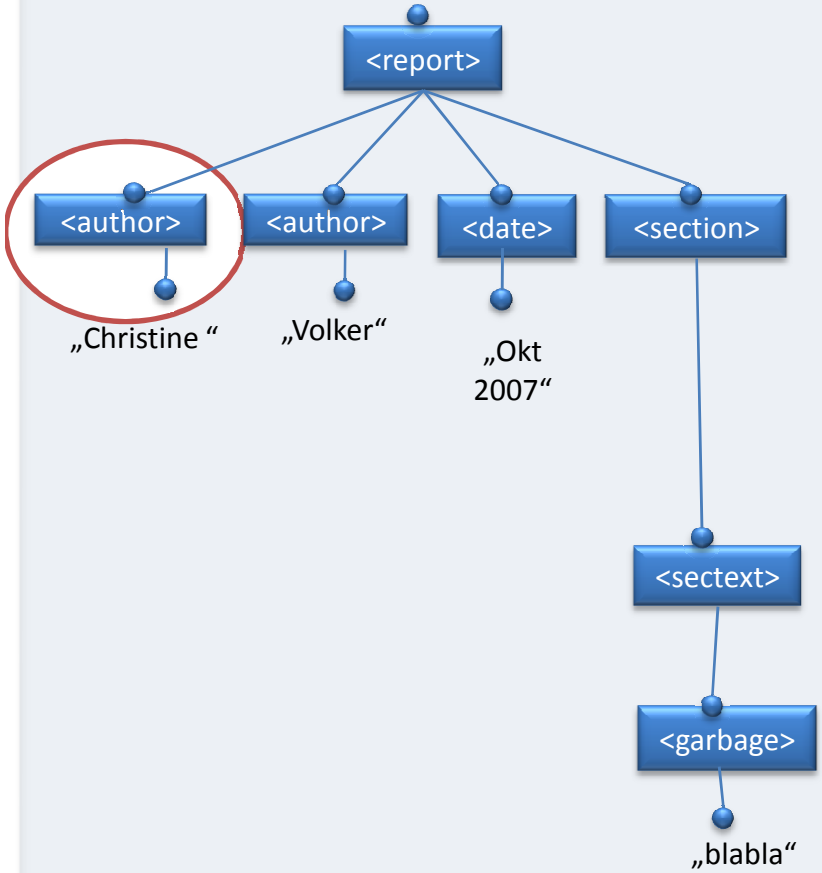


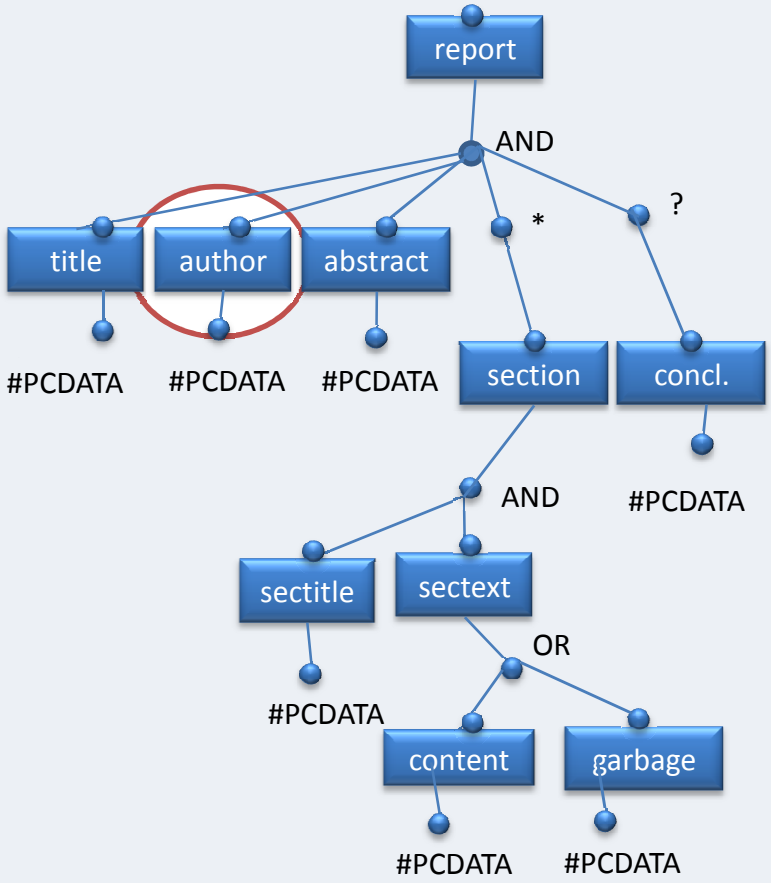
Matching Algorithm (M)



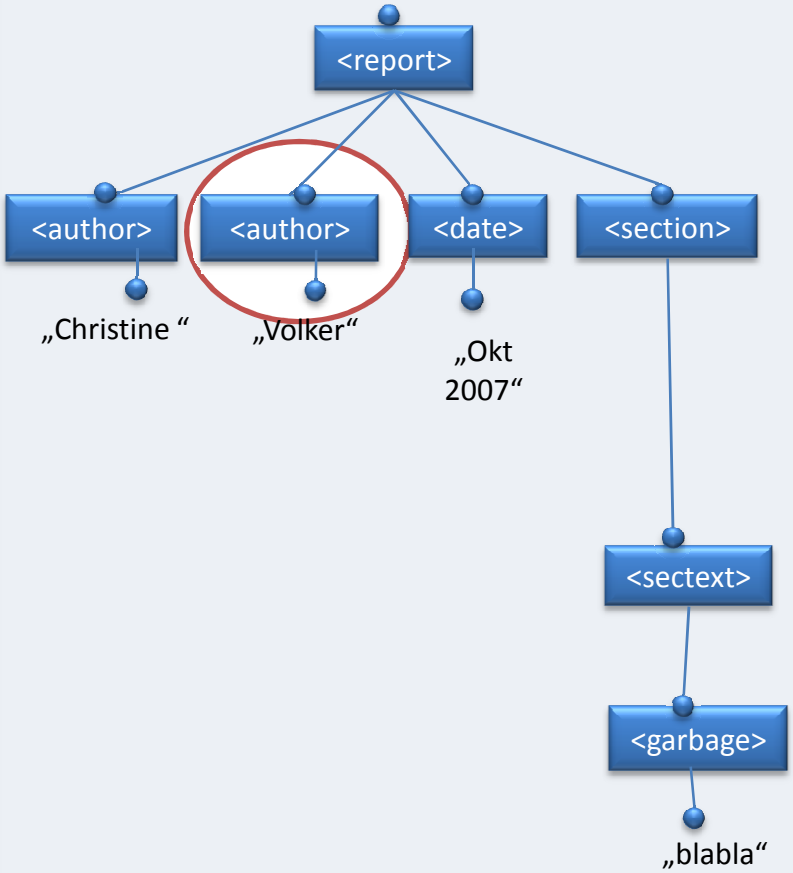


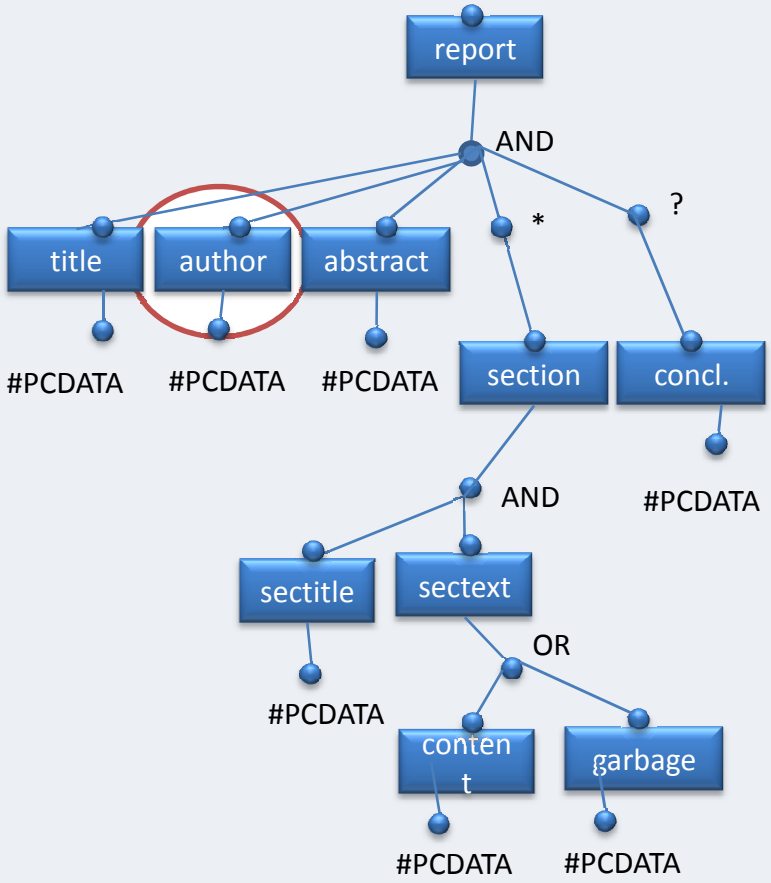
Matching Algorithm (M)



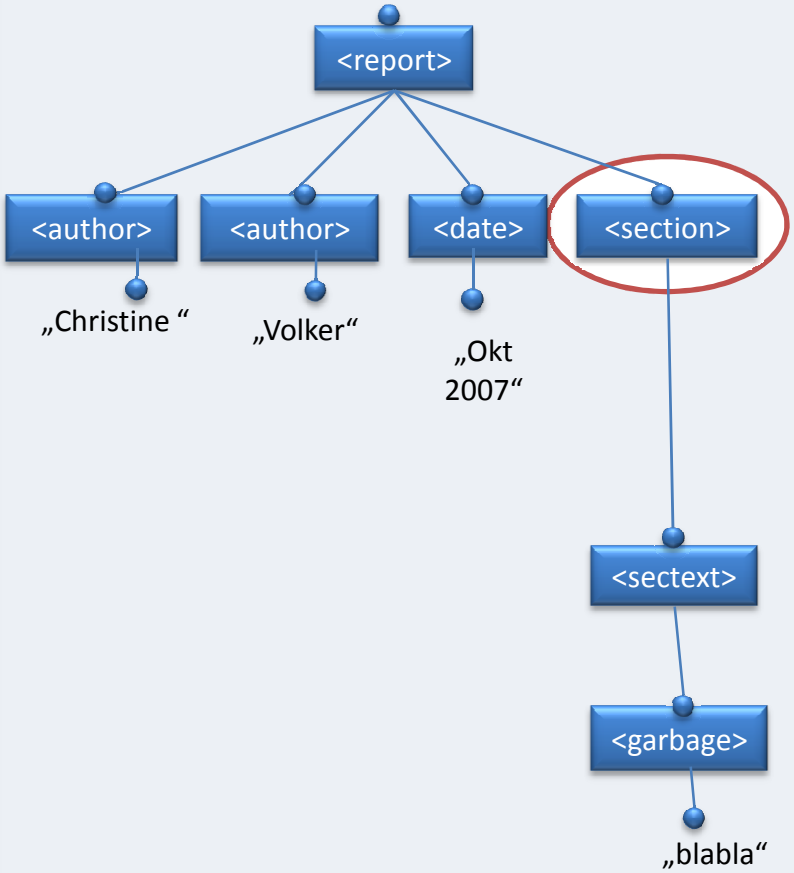


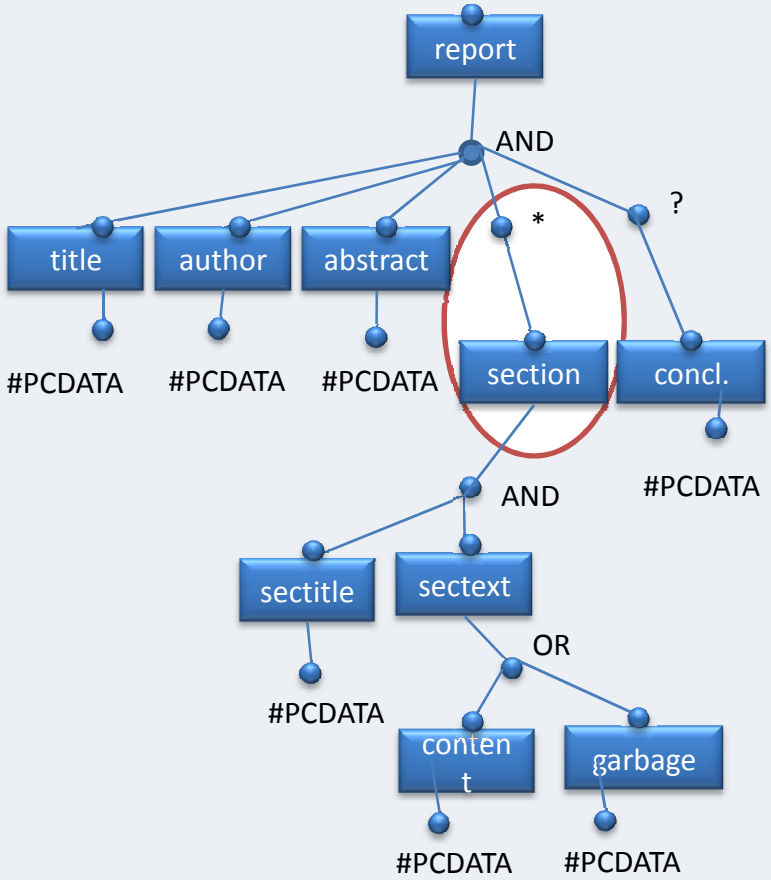
Matching Algorithm (M)



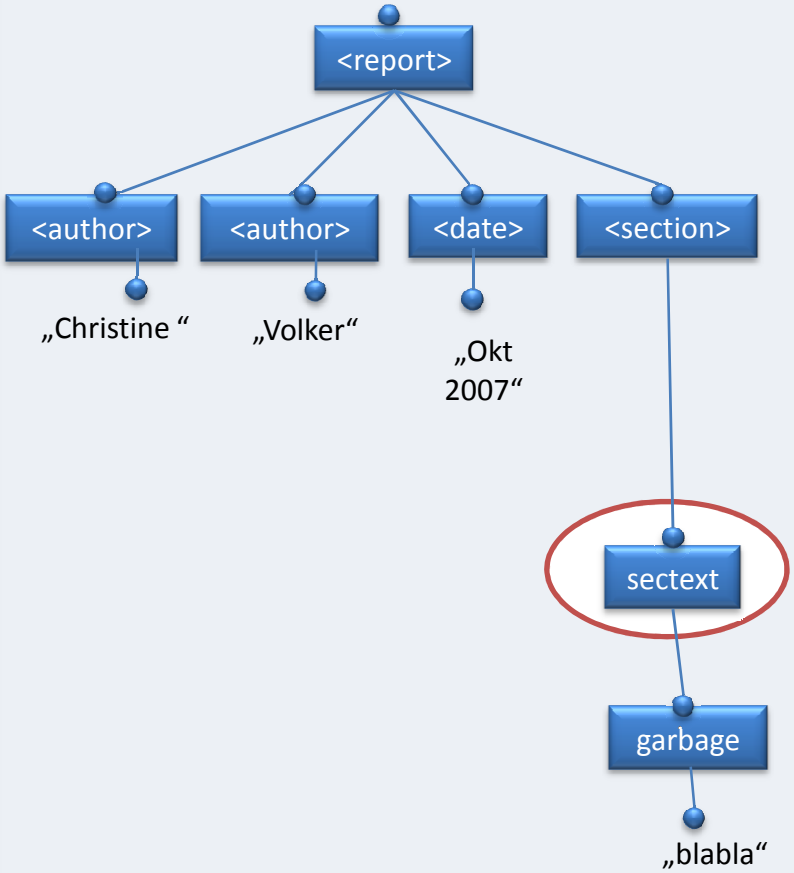


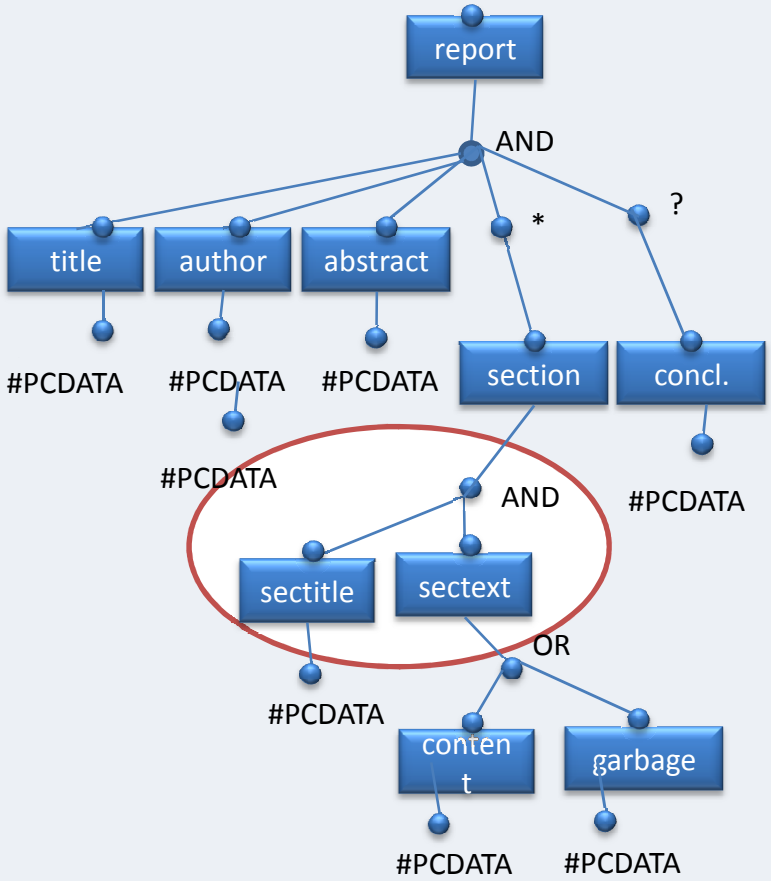
Matching Algorithm (M)



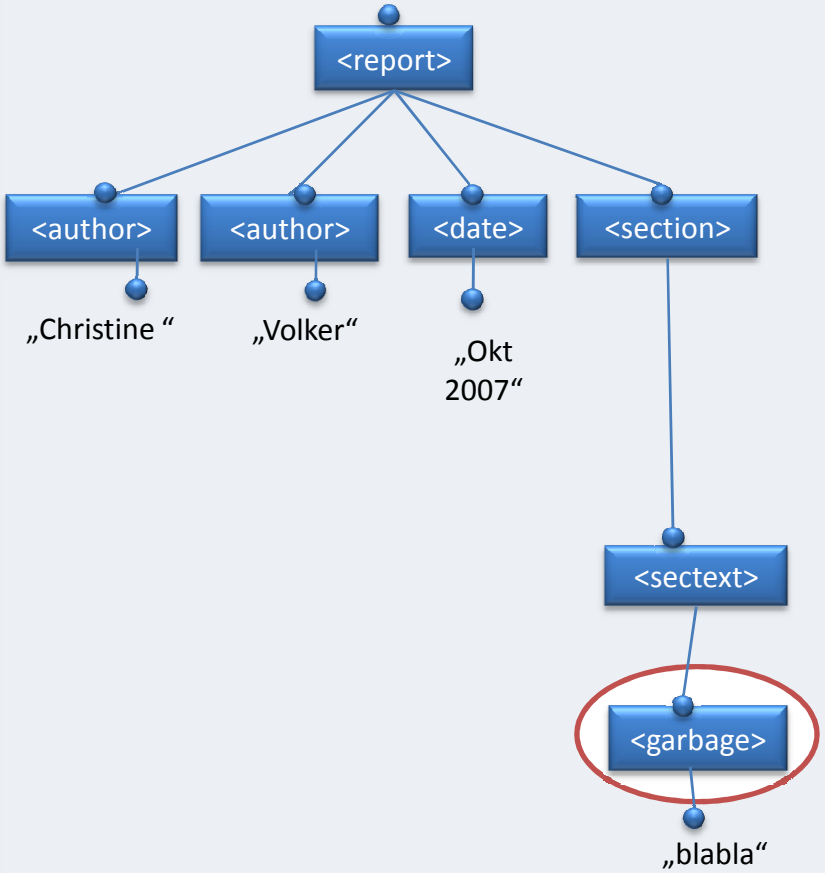


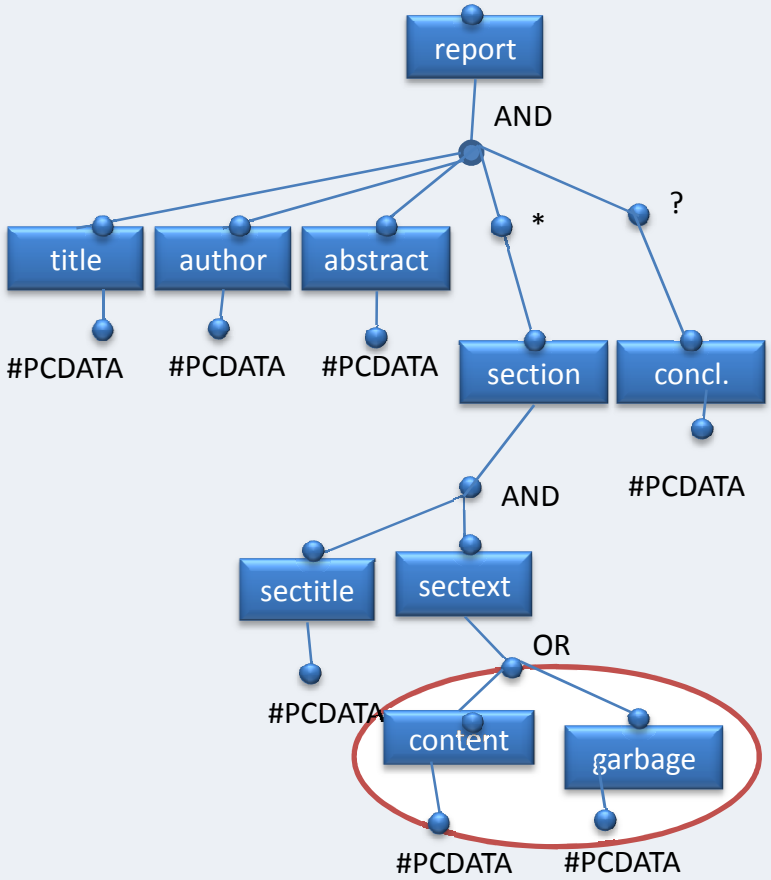
Matching Algorithm (M)





Matching Algorithm (M)





Übersicht

1. Einführung
2. Grundbegriffe
3. Algorithmen
4. Schwachstelle

Schwachstelle