

Seminar:

„XML und intelligente Systeme“

31. Oktober 2005

XML und Binärdaten

Kai Rabien

krabien@techfak.uni-bielefeld.de

Überblick

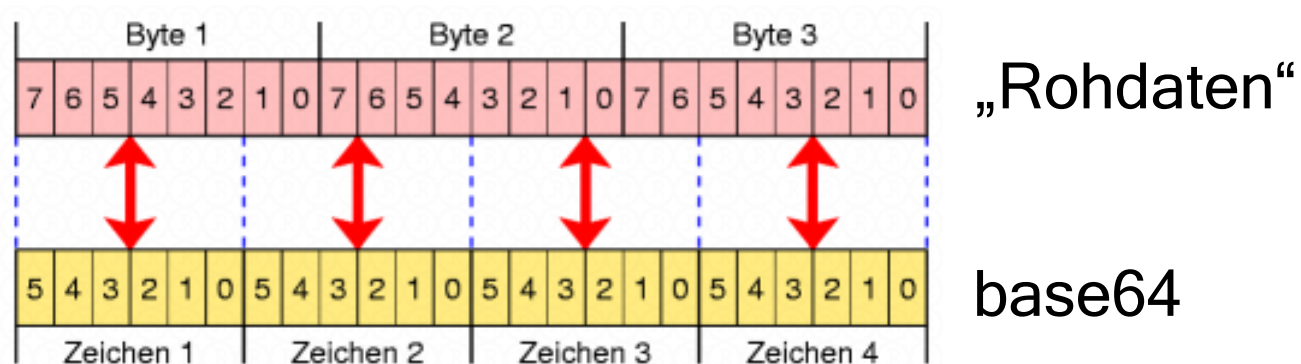
- Einführung
 - Binärdaten in XML – Die Problematik
 - Ansätze
- XOP
 - Vorstellung XOP, Beispiel
 - Was fehlt?
- Fast Infoset
 - Komponenten
 - Beispiel
- Fazit / Diskussion

Binärdaten in XML

- Daten sollen in XML eingebunden werden
 - z.B. Bilder in einer Adressen- Datenbank
- XML soll effizienter transportiert werden
 - kleinere Datenmengen -> schnellerer Transport
- Parsen / Serialisierung soll effizienter werden
 - Nicht-String-Daten sind in XML meist nicht in ihrer nativen Form gespeichert. Die Konvertierung kostet Rechenzeit / Energie

base64

- In XML-Dokumenten liegen Binärdaten i.d.R. base64 codiert vor.
Der base64- Algorithmus bildet je 3 Bytes auf 4 „6-bit-Bytes“ ab:



„Rohdaten“

base64

Quelle:
wikipedia

Das Problem:

- Demnach werden base64- codierte Daten 33% umfangreicher.
- Das kostet Speicherkapazität, und vor allem:
- Bandbreite!
Das ist in heutigen Web Services nicht akzeptabel.

Lösungsansätze

- Datenkomprimierung
 - Datenreduktion durch nachträgliche Komprimierung bzw. Dekomprimierung vor dem Parsen (kostet extra Rechenleistung)
- Infoset- basierte Modelle
 - Braucht kein Schema, bildet XML-Textdarstellung direkt in ein Binärformat ab (flexibel)
- Schema- basierte Modelle
 - Verwendet „Schema-Wissen“ zur Datenreduktion (effizient)
- Hybride

Schema-orientiert: XOP

- XOP ist eine w3c- Recommendation
- base64- Kodierung wird ersetzt durch Binärdaten-Blöcke, die im Datenstrom verlinkt werden

```
<m:data xmlns:m='http://example.org/data'>
  <m:photo>
    <xop:Include xmlns:xop='[...] /xop/include'
href='cid:http://example.org/me.png' />
  </m:photo>
</m:data>
```

```
--MIME_boundary
Content-Type: image/png
Content-Transfer-Encoding: binary
Content-ID: <http://example.org/me.png>
```

```
// binär codiertes png
```

```
...
...
...
...
```

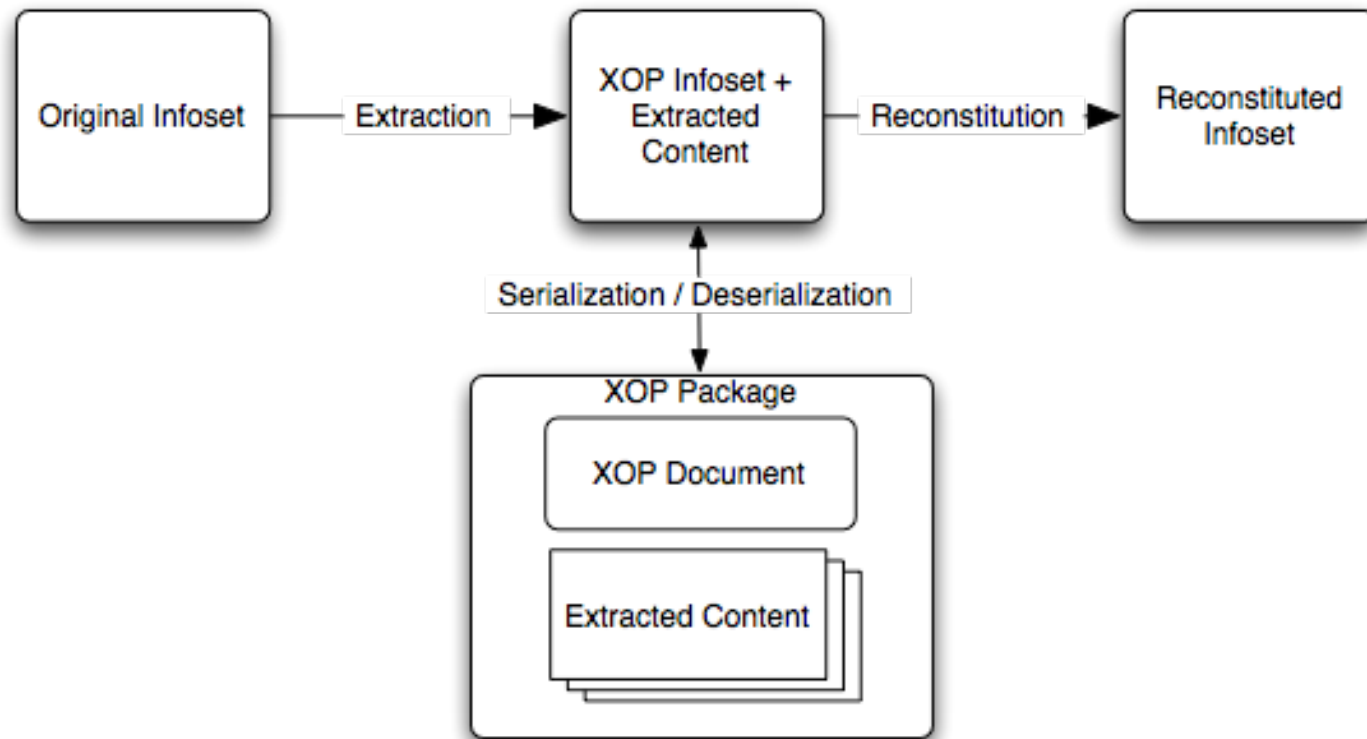
```
<m:data
  xmlns:m='http://example.org/data'>
  <m:photo>
    // base64 codiertes png
    GLWBIMSSJQSGGUqcuic1EmL
    RIYSp+5JDCOWLMEQhPGDIgA
    1Av5yy8En88Qp1q11q2xtDS
    eP7cIOKkkQioHp2it8cYG7V
    +BBmz0ZVALqLU/ekRAKzZ8O
    OBVJlgHxu3chNRXH1Z73ZTV
    51eden/0CKZPh8LCapJVpaN
    zhx49KiGvWtBYhh6+zbk59d
    .
    .
    .
  </m:photo>
</m:data>
```

Quelle: <http://www.w3.org/TR/xop10/>
(gekürzt/ergänzt)

Resultat:

- Deutlich kleinere Datenmengen
- Je nach Anwendung auch weniger Aufwand (base64 en- und decoding kann entfallen)
- Es ist ein Schema nötig (Schema- orientierter Ansatz)
- Wenn eine Konvertierung nach base64 nötig wird, entsteht zusätzlicher Aufwand

XML -> XOP -> XML



Was ist mit dem Text?

- XML- Daten enthalten immer noch „Overhead“:
- human-readable
- Viel whitespace, der keine verarbeitungsrelevanten Informationen enthält

```
<cac:Address>
```

```
██████████<cbc:StreetName>Marsh Lane</cbc:StreetName>
```

```
██████████<cbc:CityName>Nowhere</cbc:CityName>
```

```
██████████<cbc:PostalZone>NR18 4XX</cbc:PostalZone>
```

```
██████████<cbc:CountrySubentity>Norfolk</cbc:CountrySubentity>
```

```
</cac:Address>
```

Infoset-orientiert: Fast Infoset

- Eine Recommendation der ITU-T, die die Performance-Probleme mit XML-Daten lösen will:
- „To enable fast and efficient end-to-end, interoperable, Web Services for a whole spectrum of devices, from mobile phones to large back end servers“

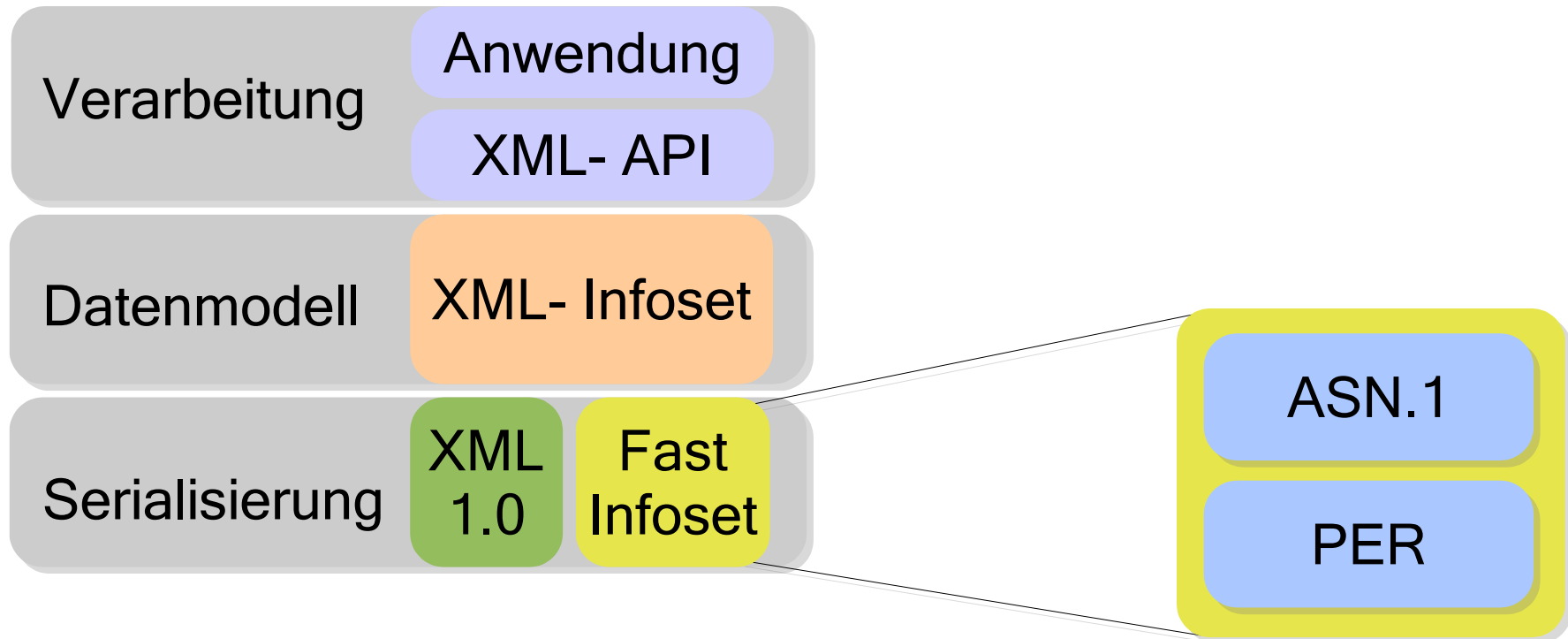
(Artikel „Fast Web Services“, developer.sun.com)

Einführung

XOP

Fast Infoset

Fast Infoset und XML- Infoset



ASN.1

- Abstract Syntax Notation One (ASN.1)
- Seit 20 Jahren in der Entwicklung
- Verwendet in Telefon-Netzwerken:
wired: Verarbeitung von Routing-Daten
wireless: Zell-Transfer Abwicklung
- ASN.1 ermöglicht Strukturdefinitionen ähnlich XML ->
es existiert ein direktes Mapping zu XSD

PER

- Packed Encoding Rules ist laut Sun:
- Die kompakteste Art der ASN.1 Kodierung
- Die schnellste ASN.1 Kodierung
- Die schnellste ASN.1 Dekodierung

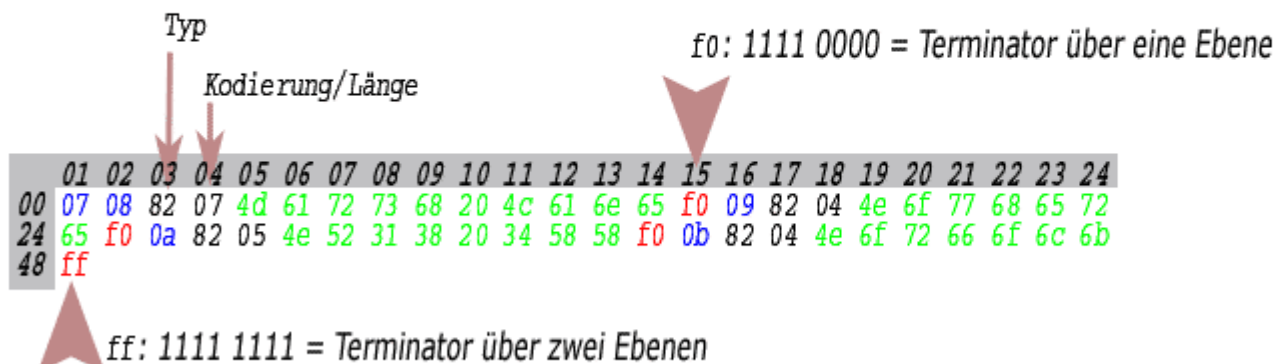
Beispiel:

```
<cac:Address>
  <cbc:StreetName>Marsh Lane</cbc:StreetName>
  <cbc:CityName>Nowhere</cbc:CityName>
  <cbc:PostalZone>NR18 4XX</cbc:PostalZone>
  <cbc:CountrySubentity>Norfolk</cbc:CountrySubentity>
</cac:Address>
```

```
<cac:Address>
  <cbc:StreetName>Marsh Lane</cbc:StreetName>
  <cbc:CityName>Nowhere</cbc:CityName>
  <cbc:PostalZone>NR18 4XX</cbc:PostalZone>
  <cbc:CountrySubentity>Norfolk</cbc:CountrySubentity>
</cac:Address>
```

Referenz- Tabelle

07	cac:Address
08	cbc:StreetName
09	cbc:CityName
0a	cbc:PostalZone
0b	cbc:CountrySubentity





Einführung

XOP

Fast Infoset

Resultat:

- Schnelleres Parsing
- Schnellere Serialisierung
- Interoperabilität hängt von FI-Schicht ab

Fazit / Diskussion

Xml-binary Optimized Packaging sorgt für einen effizienteren Transport von Binärdaten innerhalb von XML-Dokumenten. Das Infoset selbst bleibt, wie es ist.

Fast Infoset kodiert XML-Infosets in einem Binärformat, das nicht mehr wesentliche Eigenschaften von XML hat.

- Wie könnten Anwendungsszenarien aussehen?
- Wo liegen die Schwächen dieser Konzepte?
- Wie könnte man darauf reagieren?