

Datenspeicherung mit XML

Ulf Schaper

„Kulturerbe digital –
Methoden der Informatik zur Erschließung der kulturellen Überlieferung“
(Sommerakademie Greifswald 2008)

25. August 2008

Was euch erwartet:

Übersicht

Dateiformate im Wandel

anno 1998

anno 2008

Eine XML-Datei zum Anfassen

Kurzbesuch bei XSL

Was sind Namespaces?

Standards setzen: XML Schema

Dateiformate 1998

ein paar Negativbeispiele

- ▶ Fette Schrift in Word (▶ .doc)
- ▶ Spielstände im Hexeditor (▶ .sav)
- ▶ Messfahrt im BMW 335i (▶ .mdf)
- ▶ Fotos bestaunen (▶ .bmp)

Dateiformate 2008

weitere schlechte Beispiele

- ▶ Frames in Filmen (▶ .avi)
- ▶ Dokumente (▶ .pdf)

gute Beispiele

- ▶ Vektorgrafiken (▶ .svg)
- ▶ Office-Dateien (▶ .odt)
- ▶ Messfahrt im Mini Cooper S Clubman (▶ .xml)

XML: Textbasierte Datenspeicherung

- ▶ Extensible Markup Language
- ▶ viele abgeleitete Dateiformate
- ▶ breites Spektrum zwischen Maschinen- und Menschenlesbarkeit

Was gibt es darin?

- ▶ Kopfzeile/Prolog
- ▶ Elemente (in einer Baumstruktur)
- ▶ Attribute
- ▶ Kommentare

XML: Ein Beispiel zu unserer Arbeitsgruppe

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<sommerakademie ort="Greifswald" jahr="2008">
  <person typ="Akademieleiter">Thomas Ludwig</person>
  <person typ="Akademieleiter">Christine Schade</person>
  <person typ="Gastwirt" />
  <arbeitsgruppe nr="3">
    <titel>Wasserwellen: Phaenomene und Seemannsgarn</titel>
    <person typ="Gruppenleiter">Norbert Hoffmann</person>
    <person typ="Gruppenleiter">Nikolay Vitinov</person>
    <!-- to do: Teilnehmer nachtragen! -->
  </arbeitsgruppe>
  <arbeitsgruppe nr="4">
    <titel>Kulturerbe digital</titel>
    <person typ="Gruppenleiter">Guenther Goerz</person>
    <person typ="Gruppenleiter">Josef Schneeberger</person>
    <person>Sandro Bauer</person>
    <person>Saskia Becker</person>
    <person>Philip Heck</person>
    <person>Rick Hegewald</person>
    <person>Benjamin Holfeld</person>
    <person>Reha-Baris Incesu</person>
    <person>Christina Krieglstein</person>
    <person>Julia Kuehne</person>
    <person>Helen Maucher</person>
    <person>Katja Monika Poellmann</person>
    <person>Soeren Preibusch</person>
    <person>Julian Raschke</person>
    <person>Rebecca Rosenlund</person>
    <person>Ulf Schaper</person>
    <person>Karen Schewina</person>
    <person>Olaf Wittenstein</person>
  </arbeitsgruppe>
</sommerakademie>
```

XSL: Transformationen von XML zu *.*

- ▶ Extensible Stylesheet Language
- ▶ Ziele: HTML, XML, PDF, ...
- ▶ Datenbank-ähnliche Abfragen möglich

Beispiele

- ▶ Liste aller Arbeitsgruppen anzeigen
- ▶ Liste aller Personen mit Status „Gruppenleiter“
- ▶ ...

XSL: Arbeitsgruppenleiter herausfiltern

```
<?xml version="1.0" encoding="UTF-8" ?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:output method="html" indent="yes" encoding="UTF-8" />
  <xsl:template match="sommerakademie">
    <html<xsl:body<xsl:h1>Gruppenleiter in <xsl:value-of select="@ort" /></h1<xsl:ul>
      <xsl:for-each select="*/person">
        <xsl:if test="@typ='Gruppenleiter'">
          <li<xsl:value-of select="." /></li>
        </xsl:if>
      </xsl:for-each>
    </ul<xsl:/body<xsl/html>
  </xsl:template>
</xsl:stylesheet>
```

```
<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet type="text/xsl"
  href="sommerakademie_greifswald.xsl" ?>
<sommerakademie ort="Greifswald" jahr="2008">
  <arbeitsgruppe nr="1">
    <person typ="Gruppenleiter">
      Stephan Gielen</person>
    <person typ="Gruppenleiter">
      Volker Adams</person>
    ...
  </arbeitsgruppe nr="1">
  ...
</sommerakademie>
```



Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen

Gruppenleiter in Greifswald

- Stephan Gielen
- Volker Adams
- Roland Gunesch
- Marc Keßböhrmer
- Norbert Hoffmann
- Nikolay Vitanov
- Guenther Goerz
- Josef Schneeberger
- Rolf Sethe
- Christian Schröder
- Johann Kreuzer
- Tilo Wesche
- Ursula Kocher
- Shaswati Mazumdar
- Dieter Vieweger
- Jutta Häser

Namespaces: Verschiedene XML-Formate kombinieren

Beispiele

- ▶ SVG in XHTML einbinden
- ▶ MathML in SVG einbinden

Umsetzung

- ▶ entweder immer mit angeben:

```
<svg xmlns="http://www.w3.org/2000/svg"> ... </svg>
```

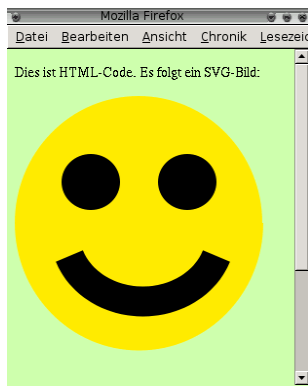
- ▶ oder Präfix definieren:

```
<html xmlns="http://www.w3.org/1999/xhtml"  
      xmlns:m="http://www.w3.org/1999/Math/MathML">  
<m:math> ... </m:math>  
</html>
```

Namespaces: Eine Vektorgrafik in einer HTML-Datei

```
<html xmlns=" http://www.w3.org/1999/xhtml">
<body>
<p>Dies ist HTML-Code.
  Es folgt ein SVG-Bild:</p>

<svg xmlns=" http://www.w3.org/2000/svg">
<g transform=" translate(-131,-141)">
<path d="M 386,268 A 127,127 0 1 1 131,268
  A 127,127 0 1 1 385,268 z"
  style=" fill:#ffeb00;" />
<path d="M 251,229 A 30,28 0 1 1 191,229
  A 30,28 0 1 1 251,229 z"
  transform=" translate(-12,-2)" />
<path d="M 251,229 A 30,28 0 1 1 191,229
  A 30,28 0 1 1 251,229 z"
  transform=" translate(87,-2)" />
<path d="M 270,572 A 80,68 0 0 1 119,572"
  transform=" translate(68,-271)"
  style=" fill:none ;
  stroke:#000000;stroke-width:30;" />
</g>/svg>
</body>/html>
```



XML Schema

- ▶ XML diktiert nur die Syntax
- ▶ SVG, XHTML, MathML, ... sind eigene Dateiformate!
 - ▶ Mit einem Schema werden diese Formate spezifiziert.

Beispiele

- ▶ „In einer SVG-Datei dürfen Pfade gruppiert werden, indem ihnen ein `<g>...</g>`-Element übergeordnet wird.“
- ▶ „Eine Sommerakademie besteht aus einer oder mehr Arbeitsgruppen.“

Schema für das Sommerakademie-Beispiel I

```
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">

  <xsd:annotation>
    <xsd:documentation xml:lang="de">
      Schema fuer Sommerakademie-Daten.
      Erstellt von Ulf Schaper, ulfschaper@gmx.de, August 2008
    </xsd:documentation>
  </xsd:annotation>

  <xsd:element name="sommerakademie" type="SommerakademieTyp" />

  <xsd:complexType name="SommerakademieTyp">
    <xsd:sequence>
      <xsd:element name="person" minOccurs="0"
        maxOccurs="unbounded" type="PersonTyp" />
      <xsd:element name="arbeitsgruppe" minOccurs="0"
        maxOccurs="unbounded" type="ArbeitsgruppeTyp" />
    </xsd:sequence>
    <xsd:attribute name="ort" type="xsd:string" />
    <xsd:attribute name="jahr" type="xsd:decimal" />
  </xsd:complexType>

  ...
```

Schema für das Sommerakademie-Beispiel II

...

```
<xsd:complexType name="ArbeitsgruppeTyp">
  <xsd:sequence>
    <xsd:element name="titel" minOccurs="0" type="xsd:string" />
    <xsd:element name="person"
      minOccurs="0" maxOccurs="unbounded" type="PersonTyp" />
  </xsd:sequence>
  <xsd:attribute name="nr" type="xsd:decimal" />
</xsd:complexType>

<xsd:complexType name="PersonTyp" mixed="true">
  <xsd:attribute name="typ"
    type="PersonenTypAttribut" default="Teilnehmer" />
</xsd:complexType>

<xsd:simpleType name="PersonenTypAttribut">
  <xsd:restriction base="xsd:string">
    <xsd:enumeration value="Akademieleiter" />
    <xsd:enumeration value="Gastwirt" />
    <xsd:enumeration value="Gruppenleiter" />
    <xsd:enumeration value="Teilnehmer" />
  </xsd:restriction>
</xsd:simpleType>
</xsd:schema>
```

Vielen Dank
für Eure Aufmerksamkeit!

Die Siedler im Hexeditor (vorher)

The screenshot shows a hex editor window titled "file://home/ufsch/Spiele/SIEDLER/SAVE8.DS - KHexEdit". The window has a menu bar with "Datei", "Bearbeiten", "Ansicht", "Dokumente", "Lesezeichen", "Extras", "Einstellungen", and "Hilfe". The main area displays a hex dump of the save file. The first few lines of the hex dump are as follows:

0001:0b20	00 00 c0 57 01 00 00 00 00 59 00 48 01 80 10	☐.....Y☐☐
0001:0b30	00 00 58 42 4b 00 00 00 2e 00 5c 00 c0 08 7c 44	..XK.....☐☐ID
0001:0b40	00 00 58 42 01 00 00 00 00 04 28 8f 00 80 10	..XK.....☐☐ID
0001:0b50	00 00 68 b4 01 00 00 01 00 08 28 8f 00 7c 44	..K☐.....☐☐ID
0001:0b60	00 00 70 c4 01 00 00 00 01 73 3f 00 84 10	..☐☐.....s?☐
0001:0b70	00 00 ff ab 42 05 5e 01 00 00 54 00 0f 27 7c 44	..TB *...T☐ID
0001:0b80	00 00 b9 72 01 00 00 00 00 5c 00 50 0c 7c 44	..☐.....A.P☐ID
0001:0b90	00 00 58 42 01 00 00 00 00 05 1e 92 00 7c 1a	..XK.....T☐ID
0001:0ba0	00 00 58 42 01 00 00 00 02 54 00 00 00 7c 44	..XK.....T☐ID
0001:0bb0	00 00 48 84 01 00 00 00 00 61 00 68 01 80 10	☐☐.....a.h...☐ID
0001:0bc0	00 00 58 42 01 00 00 00 01 54 00 00 00 7c 44	..XK.....T☐ID
0001:0bd0	00 00 a8 8d 01 00 00 00 00 01 70 3f 00 98 14	☐☐.....p?☐ID
0001:0be0	00 00 7c b9 42 01 8e 03 00 5c 00 f0 0f 7c 44	..TB.....\☐ID
0001:0bf0	00 00 58 42 01 00 00 00 01 72 9f 00 00 13	..XK.....p?☐ID

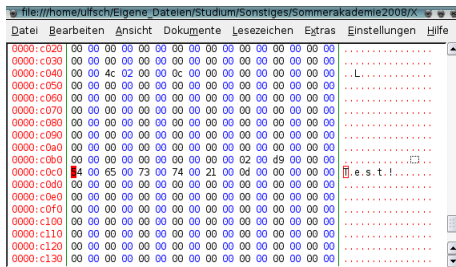
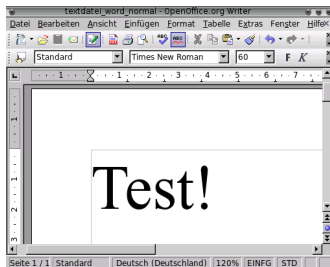


Die Siedler im Hexeditor (nachher)

```
file://home/ufsch/Spiele/SIEDLER/SAVE.DS - KHexEdit
Datei Bearbeiten Ansicht Dokumente Leesezeichen Extras Einstellungen Hilfe
0001:0b20 00 00 c0 57 01 00 00 00 00 59 00 48 01 80 10  :< . . . . Y . . . .
0001:0b30 00 00 58 42 4b 00 00 00 2e 00 5c 00 c0 08 7c 44  :.XBK.....\ . . . .ID
0001:0b40 00 00 58 42 01 00 00 00 00 04 28 8f 00 80 10  :.XB.....(.....)
0001:0b50 00 00 68 b4 01 00 00 01 00 08 28 8f 00 7c 44  :.kK.....(.....)ID
0001:0b60 00 00 70 c4 01 00 00 00 00 01 73 3f 00 84 10  :.K.....s?.....
0001:0b70 00 00 ff ab 42 45 5e 01 00 00 54 00 0f 27 0f 23  :.TB *.....]
0001:0b80 0f 27 0f 27 0f 27 0f 27 0f 27 0f 27 0f 27 0f 23  :
0001:0b90 0f 27 0f 27 0f 27 0f 27 0f 27 0f 27 0f 27 0f 23  :
0001:0ba0 0f 27 0f 27 0f 27 0f 27 0f 27 0f 27 0f 27 0f 23  :
0001:0bb0 00 00 48 84 01 00 00 00 00 00 61 00 68 01 80 10  :. . . . . a h . . . .
0001:0bc0 00 00 58 42 01 00 00 00 01 00 54 00 00 00 7c 44  :.XB.....T.....ID
0001:0bd0 00 00 a8 8d 01 00 00 00 00 01 70 3f 00 98 14  :.K.....p?.....
0001:0be0 00 00 fc b9 42 01 9e 03 00 00 5c 00 f0 0f 7c 44  :.TB.....\ . . . .ID
0001:0bf0 00 00 58 43 01 00 00 00 00 01 72 3f 00 90 13  :.vB.....p?.....
```

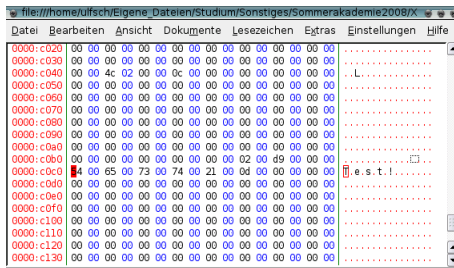


„test.doc“



- ▶ Dateigröße unverändert
52224 Bytes,
- ▶ davon ändern sich 404 Bytes
durch Fettschrift,
- ▶ ... nur nicht dort wo man sucht.

„test.doc“, jetzt mit Fettschrift



- ▶ Dateigröße unverändert
52224 Bytes,
- ▶ davon ändern sich 404 Bytes
durch Fettschrift,
- ▶ ... nur nicht dort wo man sucht.

Ein .avi-Video mit zwei Frames

```
0001fc0 0000 0000 0000 0000 0000 0000 0000 0000
0001fd0 0000 0000 0000 0000 0000 0000 0000 0000
0001fe0 0000 0000 0000 0000 0000 0000 0000 0000
0001ff0 0000 0000 0000 0000 0000 0000 0000 0000
0002000 494c 5453 0194 0000 6f6d 6976 3030 6264
0002010 00c0 0000 f7f7 00f7 0000 f7f7 00f7 0000
0002020 f7f7 00f7 0000 f7f7 00f7 0000 0000 f700
0002030 f7f7 0000 f700 f7f7 0000 f700 f7f7 0000
0002040 f700 f7f7 f7f7 00f7 0000 f7f7 00f7 0000
0002050 f7f7 00f7 0000 f7f7 00f7 0000 0000 f700
0002060 f7f7 0000 f700 f7f7 0000 f700 f7f7 0000
0002070 f700 f7f7 f7f7 00f7 0000 f7f7 00f7 0000
0002080 f7f7 00f7 0000 f7f7 00f7 0000 0000 f700
0002090 f7f7 0000 f700 f7f7 0000 f700 f7f7 0000
00020a0 f700 f7f7 f7f7 00f7 0000 f7f7 00f7 0000
00020b0 f7f7 00f7 0000 f7f7 00f7 0000 0000 f700
00020c0 f7f7 0000 f700 f7f7 0000 f700 f7f7 0000
00020d0 f700 f7f7 3030 6264 00c0 0000 f7f7 7ff7
00020e0 0000 f7f7 7ff7 0000 f7f7 7ff7 0000 f7f7
00020f0 7ff7 0000 007f f700 f7f7 007f f700 f7f7
0002100 007f f700 f7f7 007f f700 f7f7 f7f7 7ff7
0002110 0000 f7f7 7ff7 0000 f7f7 7ff7 0000 f7f7
0002120 7ff7 0000 007f f700 f7f7 007f f700 f7f7
0002130 007f f700 f7f7 007f f700 f7f7 f7f7 7ff7
0002140 0000 f7f7 7ff7 0000 f7f7 7ff7 0000 f7f7
0002150 7ff7 0000 007f f700 f7f7 007f f700 f7f7
0002160 007f f700 f7f7 007f f700 f7f7 f7f7 7ff7
0002170 0000 f7f7 7ff7 0000 f7f7 7ff7 0000 f7f7
0002180 7ff7 0000 007f f700 f7f7 007f f700 f7f7
0002190 007f f700 f7f7 007f f700 f7f7 6469 3178
00021a0 0020 0000 3030 6264 0010 0000 0004 0000
00021b0 00c0 0000 3030 6264 0010 0000 00cc 0000
00021c0 00c0 0000
```

- ▶ Der Übergang zwischen beiden Frames ist kaum zu erkennen.

Vektorgrafik (.svg) mit Smiley

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<!-- Created with Inkscape (http://www.inkscape.org/) -->

<svg
  xmlns="http://www.w3.org/2000/svg">
  <g transform="translate(-131,-141)">
    <path d="M 386,268 A 127,127 0 1 1 131,268 A 127,
      127 0 1 1 385,268 z" style="fill:#ffeb00;" />
    <path d="M 251,229 A 30,28 0 1 1 191,229 A 30,28
      0 1 1 251,229 z" transform="translate(-12,-2)" />
    <path d="M 251,229 A 30,28 0 1 1 191,229 A 30,28
      0 1 1 251,229 z" transform="translate(87,-2)" />
    <path d="M 270,572 A 80,68 0 0 1 119,572"
      transform="translate(68,-271)" style="fill:none;
      stroke:#000000;stroke-width:30;" />
  </g>
</svg>
```



▸ anno 2008

Textdatei (.odt) mal näher untersucht

```
Archive: ../textdatei_word_normal.odt
  extracting: mimetype
    creating: Configurations2/statusbar/
  inflating: Configurations2/accelerator/current.xml
    creating: Configurations2/floater/
    creating: Configurations2/popupmenu/
    creating: Configurations2/progressbar/
    creating: Configurations2/menubar/
    creating: Configurations2/toolbar/
    creating: Configurations2/images/Bitmaps/
  inflating: content.xml
  inflating: styles.xml
  extracting: meta.xml
  inflating: Thumbnails/thumbnail.png
  inflating: settings.xml
  inflating: META-INF/manifest.xml
```

Test!

.odt: Was steht in der content.xml?

```
<?xml version="1.0" encoding="UTF-8"?>
<office:document-content xmlns:office="urn:oasis:names:tc:opendocument:xmlns:office:1.0"
xmlns:style="urn:oasis:names:tc:opendocument:xmlns:style:1.0"
xmlns:ooo="http://openoffice.org/2004/office"
xmlns:ooow="http://openoffice.org/2004/writer"
...
<office:automatic-styles><style:style style:name="P1" style:family="paragraph"
style:parent-style-name="Standard" style:master-page-name="Standard">
<style:paragraph-properties style:page-number="auto"/><style:text-properties
fo:font-size="60pt" style:font-size-asian="60pt" style:font-size-complex="60pt"/>
</style:style></office:automatic-styles>

<office:body><office:text>
<text:sequence-decls><text:sequence-decl text:display-outline-level="0"
text:name="Illustration"/><text:sequence-decl text:display-outline-level="0"
text:name="Table"/><text:sequence-decl text:display-outline-level="0"
text:name="Text"/><text:sequence-decl text:display-outline-level="0"
text:name="Drawing"/></text:sequence-decls>
<text:p text:style-name="P1">Test!</text:p>
</office:text></office:body></office:document-content>
```

Messfahrt im Mini Cooper S Clubman

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml-styleSheet type="text/xsl" href="Auswertung_Verbrauch.xsl"?>
<resultList counter="104" complete="false" this="Auswertung_Verbrauch.xml">
  <result num="1">
    <data>
      <name>20080411_democar_ATtest_01.dat</name>
    </data>
    <date>02.07.2008</date>
    <time>10:45:03</time>
    <file num="0" fileFilter="MDF">
      <description>Errechnung des Verbrauchs in g/kWh</description>
      <event cred="1.0" id="1">
        <extraData>
          <Messdauer>0.09</Messdauer>
          <Motordrehzahl>2401.3026315789475</Motordrehzahl>
          <Moment>117.36842105263158</Moment>
          <Verbrauch_lph>10.68</Verbrauch_lph>
          <Verbrauch_gkWh>277.2395477586155</Verbrauch_gkWh>
          <Leistung_avg>29508.344195983384</Leistung_avg>
          <momentaner_Verbrauch_Start>28431.0</momentaner_Verbrauch_Start>
          <momentaner_Verbrauch_End>28698.0</momentaner_Verbrauch_End>
          <StartTime>21.984231490041438</StartTime>
          <n_Motor_diff>1.0</n_Motor_diff>
          <Moment_diff>0.0</Moment_diff>
        </extraData>
      </event>
    </file>
  </result>
  ...
</resultList>
```