

Propädeutikum
Kurzeinführung XML, Zeichenkodierung, Digitale
Bilder, OCR

❖ XML I

- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

XML (eXtensible Markup Language) definiert Regeln zur Erstellung neuer Auszeichnungssprachen.

Elemente

- Elementnamen werden in spitzen Klammern angegeben
- beinhalten einfachen Text und / oder weitere Elemente
- haben ein Start- und ein Endtag
- können auch keinen Inhalt haben
- dürfen sich nicht überlappen

```
<a>Ein Verweis zum <b>W3C</b>?</a><br />
```

Attribute

- werden im Start-Tag angegeben
- bestehen aus einem Namen und einem Wert

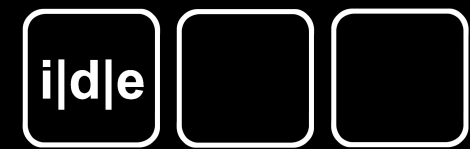
```
<a href="http://www.w3c.org">Ein Verweis zum  
<b>W3C</b>!</a><br />
```

- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

```
...<a href="http://www.w3c.org">Ein Verweis zum  
<b>W3C</b>!</a><br/>...
```

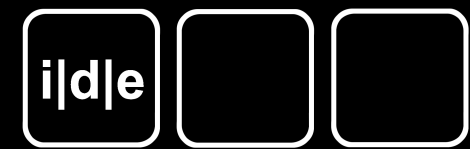
- Element- und Attributnamen werden (sinnvoll) vergeben
- Elemente und Attribute tragen Informationen
- Elemente, Attribute und deren Auftreten im Dokument können durch ein Schema festgeschrieben werden
- Mit einem Schema ist eine Überprüfung möglich (wohlgeformt vs. valide)
- Es gibt mehrere Schema-Sprachen (XML Schema, RelaxNG, DTD)
- Es muss immer ein Wurzelement geben

XML: Aussehen?



- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

- Es gibt kein Programm, das eine XML-Ressource auf Anhieb im Sinne des Herausgebers richtig anzeigen kann
- Trennung von Inhalt (XML) und Aussehen (z.B. HTML)
- Inhalte veralten nicht, Layouts und Anzeigemedien dagegen schon
- Es erfolgt eine Wandlung von XML in ein bestimmtes Zielformat



- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ **Kodierung I**
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

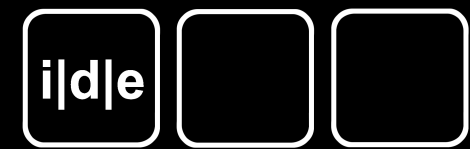
```
<?xml version="1.0" encoding="utf-8"?>
```

Ein Schritt zurück: Technisches

- Zeichenkodierung
- Zeichenspeicherung
- Zeichendarstellung
- Zeicheneingabe

Noch ein Schritt zurück: Historisches

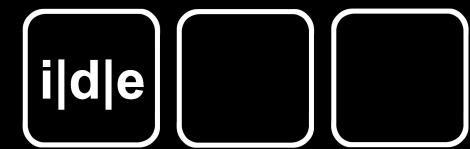
- ASCII, 128 Zeichen: elementare Zeichen
- ISO-8859-X, maximal 256 Zeichen: länderspezifische Erweiterungen
- ISO-8859-1, westeuropäische Umlaute und Sonderzeichen



- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ **Kodierung II**
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

Unicode

- Ziel: alle sinntragenden Schriftzeichen aus allen Kulturen und Sprachen zu kodieren
- Die aktuelle Version (5.1) enthält 100713 Zeichen
- Es sind 1114112 Zeichen möglich
- Wird am häufigsten in UTF-8 gespeichert
- Moderne Schriftfonts können „nur“ 65536 Zeichen aufnehmen
- Aktuell kann kein Unicode-Font alle möglichen Zeichen darstellen



- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ **Kodierung III**
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

„Mediävistische“ Zeichen

- Viele sind schon in Unicode integriert
- Es existieren mehrere Fonts (Junicon, Titus)
- Zentrale Sammelstelle: Medieval Unicode Font Initiative (<http://www.mufi.info/>)
- Ziel der MUFI ist es, alle mediävistischen Zeichen in Unicode zu integrieren
- Bis dies der Fall ist, wird die Private Use Area von Unicode genutzt
- Die unkomplizierte Eingabe der Zeichen ist noch nicht gelöst

Hilfe! Mein Zeichen ist nicht dabei!

- Kodierung als Element, z.B. `<ax />`
- Kodierung als Zeichen, z.B. `#z+ax#s-`
- Zeichen bei der MUFI melden

- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

Einsatzzweck: Bilddigitalisierung

- Bei Druckwerken vor 1750 oder bei unikatlicher Überlieferung gehört eine Digitalisierung zum guten Ton
- Auch wenn nur ein Teil benötigt wird, immer die komplette Vorlage digitalisieren
- Graustufen-Aufnahmen nur dann, wenn keine Informationen über Farbe transportiert werden
- Auflösung der Kamera beachten: mindestens mit 300dpi digitalisieren (bezogen auf die Originalvorlage)
- Es wird zwischen Master-Version und Derivaten unterschieden
- Die Frage der Archivierung der Master-Version klären
- Auf eine sinnvolle und eindeutige Benennung achten, z.B. `kn28_83II_0059_30r.tif`

- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

Die Erzeugung der Digitalisate hängt von der besitzenden Einrichtung ab.

Einrichtung digitalisiert selbst

- Die Schonung des Originals steht im Vordergrund
- Eine professionelle Ausrüstung ist meist vor Ort vorhanden, was sich positiv auf die Qualität auswirkt.
 - ❖ Wolfenbütteler Buchspiegel
 - ❖ Grazer Buchwippe
- Es muss mit Gebühren, einem zeitlichen Vorlauf und Einschränkungen beim späteren Gebrauch gerechnet werden

- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

Die Digitalisierung wird in die eigene Hand genommen

- Ist mit einem erhöhten Aufwand verbunden
- Hochwertige Kamera (> 12 MegaPixel)
- Stativ
- Gute, gleichbleibende Beleuchtung
- Eine entsprechende Rechnerausstattung ist nötig
- Lineal? Farbskala? Farbprofil?
- Alle Aufnahmen unter den gleichen Bedingungen aufnehmen
- Eine sofortige und umfassende Qualitätskontrolle ist notwendig

- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

- Es wird ein Mikrofilmscanner benötigt
- Die Qualität des Films bestimmt die Qualität der Scans
- Die Filme wurden meist nicht in Farbe aufgenommen, Aufnahmen in Graustufen sind die Regel

- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ **Bildformate I**
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

● TIFF

- ❖ Eignet sich für die Master-Version
- ❖ Ein sehr komplexes Format, daher wird eine Beschränkung auf „Baseline-TIFF“ angeraten
- ❖ Keine zusätzliche Kompression verwenden
- ❖ Hat sich als Archivformat durchgesetzt
- ❖ Als Arbeitsversion nicht sonderlich geeignet, da die Bilder einen zu großen Dateiumfang haben

● JPEG

- ❖ Eignet sich für Bildderivate
- ❖ Eigentlich nur ein Kompressionsalgorithmus
- ❖ Komprimiert verlustbehaftet (Grad einstellbar)
- ❖ gut geeignet für Farb-Digitalisate
- ❖ wegen der Artefaktbildung schlecht für reine Textseiten

- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ **Bildformate II**
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

- PNG
 - ❖ Eignet sich für Bildderivate
 - ❖ Komprimiert verlustfrei
 - ❖ Gut geeignet für alle Digitalisate
 - ❖ Dateiumfang nicht immer der kleinste
- GIF
 - ❖ Eignet sich für Bildderivate
 - ❖ Komprimiert verlustfrei
 - ❖ Gut geeignet für Graustufen-Digitalisate
 - ❖ Unterstützt nur 256 Farben

- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ **Volltext I**
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

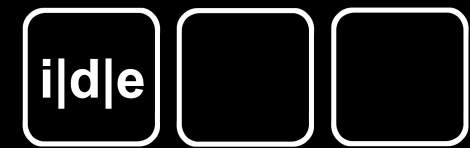
Einsatzzweck: Volltext

- Eine Rechtefreiheit ist erst 80 Jahre nach dem Tod des Autors gegeben
- Eine Digitalisierung ist einfacher, wenn das Buch entleihbar ist
- Es ist keine hochwertige Ausrüstung nötig: ein herkömmlicher Flachbettscanner reicht in vielen Fällen aus
- Mit einem richtigen Buchscanner werden aber i.d.R. in kürzerer Zeit bessere Aufnahmen gemacht
- Eine gute Vorlage ermöglicht ein gutes Ergebnis
- Abtippen (lassen) oder OCR?
 - ❖ Double-Keying: 99,997% Genauigkeit
 - ❖ OCR: Genauigkeit schwankt je nach Art der Vorlage, der eingesetzten Software und dem betriebenen Aufwand beim Korrigieren
 - ❖ OCR: „Schmutziger“ Volltext als einfache Alternative

- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ **Volltext II**
- ❖ Volltext III
- ❖ XML erzeugen

Unter den OCR-Programmen ist aktuell der FineReader von Abby das Produkt der Wahl

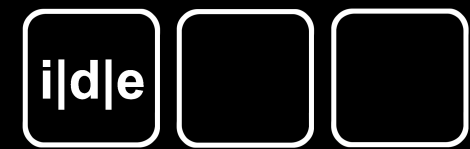
- Mehrsprachigkeit innerhalb eines Textes soll kein Problem mehr sein
- Erkennung von Fraktur-Schriften ist teuer, falls sie überhaupt funktioniert (vorher testen)
- Das gleichzeitige Vorkommen von Fraktur- und Antiqua-Schrift funktioniert nicht



- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ **Volltext III**
- ❖ XML erzeugen

OCR: Arbeitsschritte

- Bilddateien umwandeln (erledigt die Software)
- Bereiche festlegen. Ein manueller Eingriff könnte lohnen
- Zeichen erkennen (erledigt die Software)
- Ggf. ist ein Trainieren der OCR-Engine nötig
- Erkannten Text korrigieren
- Exportieren des erkannten Textes
- Optional kann sich eine Weiterverarbeitung des exportierten Textes anschließen (Wandlung nach XML)



- ❖ XML I
- ❖ XML II
- ❖ XML III
- ❖ Kodierung I
- ❖ Kodierung II
- ❖ Kodierung III
- ❖ Digitalisierung I
- ❖ Digitalisierung II
- ❖ Digitalisierung III
- ❖ Mikrofilm
- ❖ Bildformate I
- ❖ Bildformate II
- ❖ Volltext I
- ❖ Volltext II
- ❖ Volltext III
- ❖ XML erzeugen

- Im Layout der Vorlage stecken viele Informationen
- Implizites explizit machen
- Anwenden von erweiterten Suche- und Ersetze-Operationen
- Umwandlung auf Grundlage des XML-Formates der Textverarbeitung (Programmierung notwendig)
- Es ist eine manuelle Nachbearbeitung nötig
- Je nach Einsatzzweck kann es aufwändig werden