Bilder und TEI

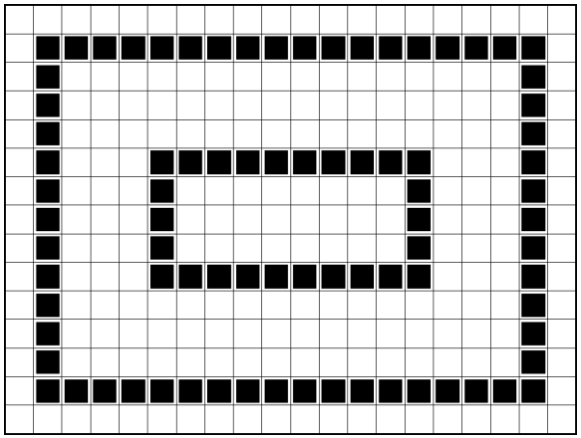
O. Duntze

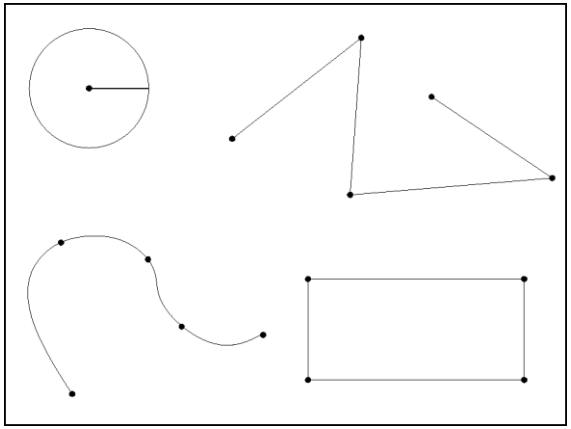
Gliederung

- **Grundlagen der Bilddigitalisierung**
- Digitale Faksimiles in TEI definieren
- Text und Bild verbinden
- Hilfsmittel
- Darstellung von Text-Bild-Verbindungen

Grundlegendes zur Bilddigitalisierung

- Digitale Reproduktion analoger „Objekte“
- Häufig Textträger (Manuskripte, Druckseiten, Inschriften, Keilschrifttafeln usw.) -> Für den Computer sind das Fotos, keine Texte!!!
- Bilddigitalisierung erzeugt Rastergrafiken, keiner Vektorgrafiken!
  - Rastergrafik: x mal y Bildpunkte (**Pixel**), für jeden Bildpunkt ein Farbe
  - Vektorgrafik: Beschreibung eines Bildes durch geometrische Figuren bzw. Pfade



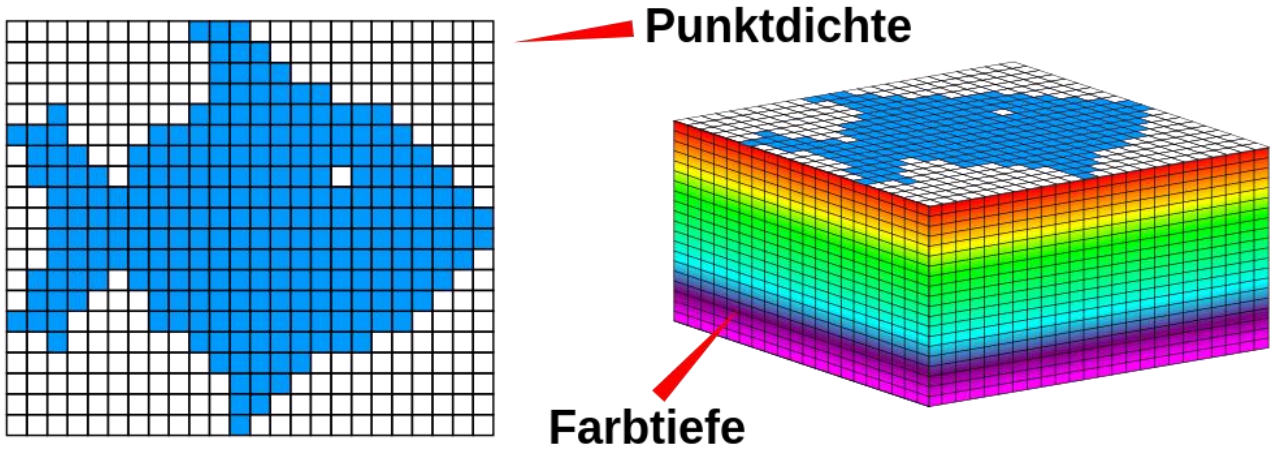


Quelle: Kunze, Flash Weather Ein Generator fuer Macromedia Flash zur interaktiven Visualisierung XML basierter Daten. Universität Osnabrück, Dipl.Arbeit 2001.
http://www-lehre.inf.uos.de/~rkunze/flashweather/Diplomarbeit/node8.html

Bildqualität bei Rastergrafiken

- **Auflösung** (Abtastauflösung, Punktdichte)
  - Rasterung einer Längeneinheit im Original in x Bildpunkte
  - Angaben in dpi (dots per inch)
  - 300 dpi = 1 Inch (~2,54 cm) des Originals ergibt 300 Bildpunkte in der digitalen Reproduktion (entspricht ca. 118px/cm, d.h. 1px ~ 0,08mm)
  - Wichtig: die in einer Bilddatei angegebene Auflösung (sofern überhaupt angegeben) entspricht nicht unbedingt der Größe des gescannten Originals, sondern gibt an, wie groß das Bild beim Druck ausgegeben wird!
- **Farbtiefe**
  - Farbinformationen für jeden Bildpunkt
  - Angegeben in bpp (bit per pixel), z.B.
    - bitonal (schwarz/weiß, 1 bpp)
    - Graustufen (256 Grauwerte, 8 bpp)
    - 256 Farben (ausreichend für grafische Illustrationen / Zeichnungen, 8 bpp)
    - 16.777.216 Farben (True Color, 24 bpp, RGB-Modell, je 8 bit pro Farbkanal)
    - 281.474.976.710.656 (48 bpp, je 16 bit pro RGB-Kanal)

Punktdichte und Farbtiefe



Quelle: http://de.wikipedia.org/w/index.php?title=Datei:Punktdichte%2BFarbtiefe.svg&filetimestamp=20111016165836 (F. Graf)

- Bitonale Scans (z.B. für reine Druckseiten) sollten mit mindestens 600dpi gescannt sein
- Farbscans (meist 24bpp) und Graustufen (8bpp) mindestens 300 dpi, bei schwierigem Material (z.B. Karten, kleine Handschriften etc.) auch mehr
- Wichtig: Farben werden immer verfälscht
  - Normierte Farbkarte mitscannen (inkl. Maßstab)
  - Ggf. Farbkalibrierung
  - Gleichbleibende Lichtverhältnisse

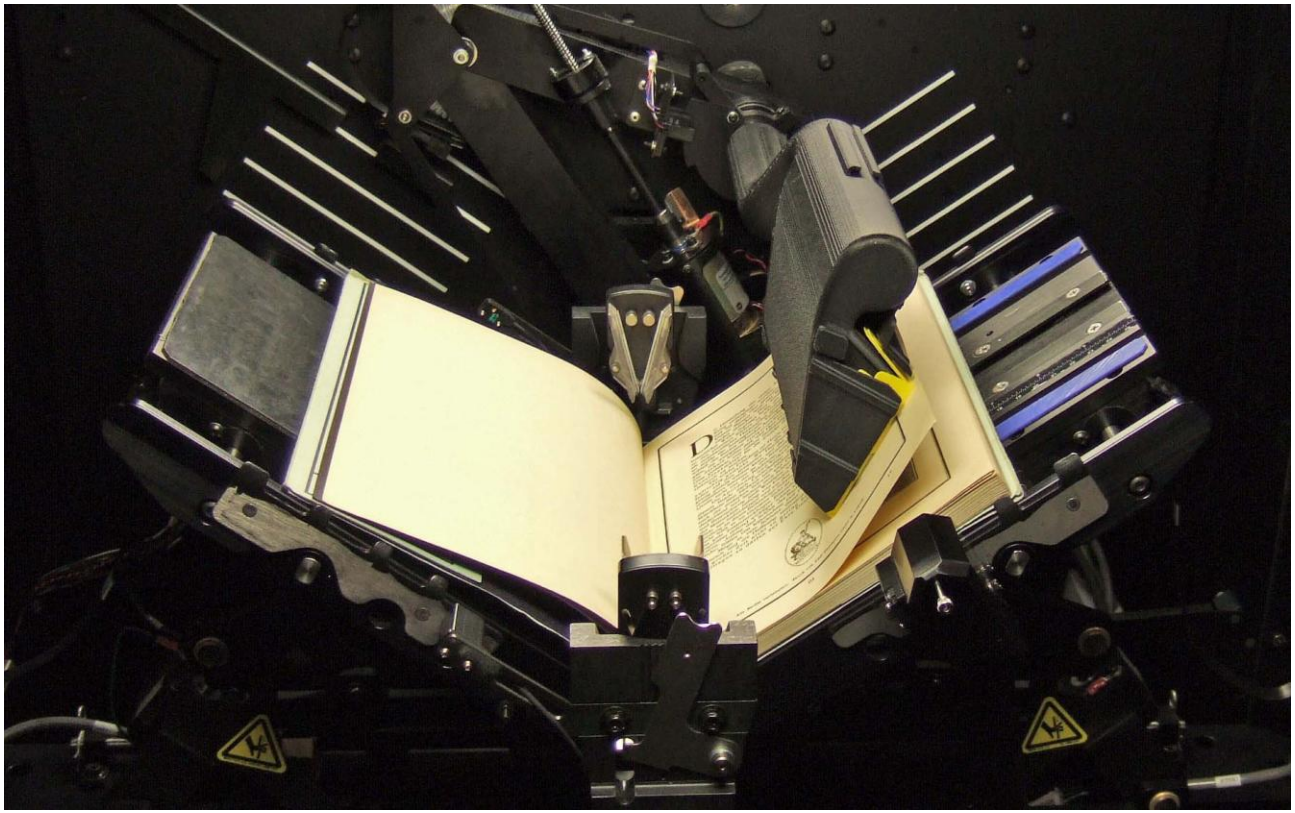
Wege zum digitalen Bild

- Digitale Fotografie
  - Schneller als Scannen
  - Erlaubt Digitalisierung problematischer Vorlagen wie Inschriften
  - Auflösung meist geringer als bei Scannern und „Rauschen“
  - Maßstab muss mitgegeben bzw. mit fotografiert werden
- Scanner
  - hohe Auflösung
  - Maßstab implizit
- Durchlichtfotografie, Thermographie, Röntgenaufnahmen, Multispektralfotografie, Reflectance Transformation Imaging (RTI)
  - Wasserzeichen
  - Palimpseste
  - schwer leserliche Papyrii
  - Inschriften

Hardware (Flachbettscanner)

Hardware (Aufsichtscanner)

Hardware (Scanroboter)



Hardware
(Wolfenbütteler Buchspiegel, Grazer Buchtisch)

Komprimierung

- Problem: Bilddaten werden schnell sehr umfangreich
  - Z.B. Größe 1000 x 1000px, 24bit (=3 Byte) Farbtiefe = 3MB Bilddaten (+ggf. Headerinformationen etc.)
  - Normales Handy: 8 Megapixel -> 24MB Bilddaten
- Lösung: Komprimierung
  - Z.B.: Eine Zeile (1000px) ist komplett schwarz -> ca. 3 KB Bilddaten
  - Komprimiertes Dateiformat speichert „Jetzt kommen 1000 Pixel, die alle schwarz sind“ –> ca. 300 Byte
  - Problem: Dateien in der Regel fehleranfälliger, deshalb werden für die Langzeit-Archivierung (LZA) oft unkomprimierte Formate empfohlen

Verlustfreie vs. Verlustbehaftete Komprimierung

- Verlustfreie Komprimierung
  - Bilddaten können exakt rekonstruiert werden, z.B.
    - RLE (Run-Length Encoding) -> sucht aufeinanderfolgende gleiche Werte
    - LZW (Lempel-Ziv-Welch-Algorithmus) -> sucht wiederkehrende Muster und ersetzt sie durch ‚Abkürzungen‘
- Verlustbehaftete Komprimierung
  - Bilddaten können so rekonstruiert werden, dass sie wie das Original wirken, aber bei genauem Hingucken abweichen.
    - JPEG (Standard 1992 von der Joint Photographic Experts Group entwickelt)
    - Audio: MP3

TIFF, LZW-komprimiert, 350KB

JPEG mit hoher Kompressionsrate, 2KB

Wichtige Dateiformate

- TIFF (Tagged Image File Format) *.tif
  - Containerformat, kann unkomprimierte, verlustfrei (LZW, RLE) komprimierte oder verlustbehaftet (JPEG) komprimierte Bilddaten enthalten
  - Metadaten in eigenem Header
  - Häufig für Masterbilder
- JPG/JPEG/JFIF (JPEG File Interchange Format) *.jpg
  - Format für JPEG-codierte Bilddaten (meist verlustbehaftet)
  - Für Webpräsentation
- GIF (Graphics Interchange Format) *.gif
  - Verlustfrei (LZW) komprimierte Daten mit geringer Farbtiefe (8 bpp / 256 Farben)
- PNG (Portable Network Graphics) *.png
  - Als Alternative zu GIF entwickelt
  - Verlustfreie Komprimierung (nicht LZW wg. Patentforderungen)
  - variable Farbtiefe

Gliederung

- Grundlagen der Bilddigitalisierung
- **Digitale Faksimiles in TEI definieren**
- Text und Bild verbinden
- Hilfsmittel
- Darstellung von Text-Bild-Verbindungen

TEI und Bilder

- Guidelines Kap. 11
- Modul **`transcr`** stellt die für Digitale Faksimiles und Transkriptionen wichtigen Elemente und Attribute bereit. U.a.
  - <facsimile/>
  - <surface/>
  - <zone/>
  - att.global.facs (stellt @facs bereit)
  - att.global.change (stellt @change bereit)

Die einfache Methode

```
<text>
   <body>
       <div>
       <head>Mein Text</head>
       <pb n="1"/>
       <p>Hier kommt der Text von Seite 1</p>
       <pb n="2"/>
       <p>Hier kommt der Text von Seite 2</p>
       </div>
   </body>
</text>
```

Die einfache Methode: @facs mit Verweis auf Datei

```
<text>
   <body>
       <div>
       <head>Mein Text</head>
       <pb n="1" facs="seite1.tif"/>
       <p>Hier kommt der Text von Seite 1</p>
       <pb n="2" facs="seite2.tif"/>
       <p>Hier kommt der Text von Seite 2</p>
       </div>
   </body>
</text>
```

➔ Bei fast jedem Element möglich

Beispiel: 001 facs-Attribut.xml

Die bessere Methode: `<facsimile/>`

- **`<facsimile>`** … **`</facsimile>`** steht normalerweise nach dem **`<teiHeader>`**
- Enthält im einfachsten Fall ein oder mehrere <graphic url="xyz">-Elemente:

```
<facsimile>
   <graphic url="seite1.tif"/>
   <graphic url="seite2.tif"/>
   <graphic url="seite3.tif"/>
</facsimile>
```

- Tipp: Falls möglich mit @width und @length die Größe der Bilddatei in Pixel angeben, das erleichtert später die Verarbeitung

<surface/> und <zone/>

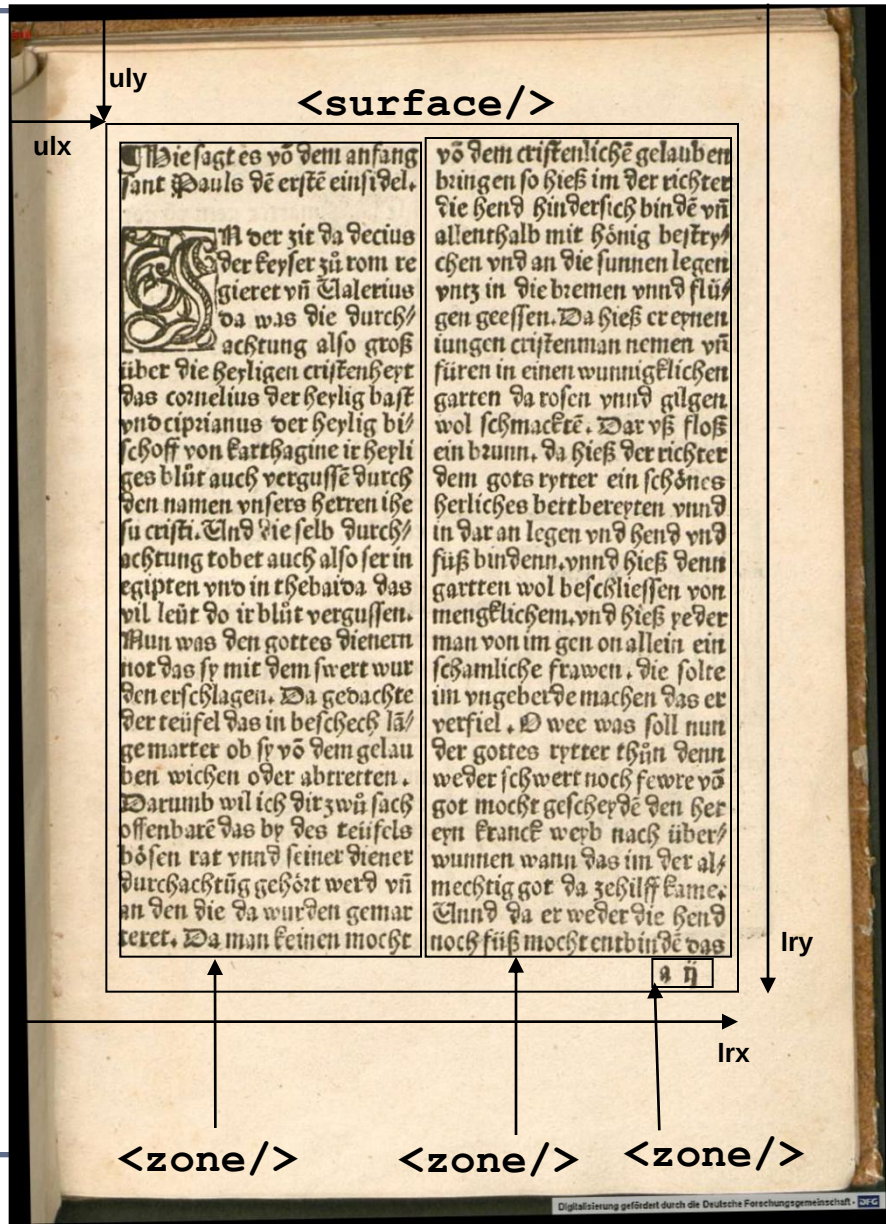
- <surface> definiert als Kindelement von <facsimile> eine beschriebene (bzw. bedruckte, beschriftete usw.) Oberfläche
- <surface> kann ein <graphic>-Element enthalten
- <zone> dient der Feingliederung der <surface>
- Beispiel: Manuskriptseite (<graphic>) enthält einen beschriebenen Bereich (<surface>) in zwei Spalten (2mal <zone>) und Marginalien (<zone>)
- Koordinaten werden folgenden Attributen angegeben
  - @ulx -> Upper left x
  - @uly -> Upper left y
  - @lrx -> Lower right x
  - @lxy -> Lower right y
- Angaben der Koordinaten können in Pixeln (bezogen auf das in <graphic> codierte Bild) gemacht werden, aber auch in cm, mm usw.

Beispiel <surface>

```
<facsimile>
    <surface ulx="120" uly="150" lrx="860"
    lry="1140">
    <!-- Satzspiegel der Seite -->

          <graphic url="M29857-002r.jpg"/>
          <!-- Verweist auf die Bilddatei -->

          <zone ulx="120" uly="150" lrx="480"
    lry="1125"/> <!-- Spalte 1 -->

          <zone ulx="500" uly="150" lrx="860"
    lry="1125"/> <!-- Spalte 2 -->

          <zone ulx="775" uly="1130" lrx="860"
    lry="1140"/> <!-- Bogensignatur aij -->

    </surface>
</facsimile>
```

➔ Mächtiges Instrument, um das Layout zu beschreiben

Beispiel: 002 facsimile.xml



<surface/> und <zone/>

- <surface> und <zone> verwenden standardmäßig das gleiche Koordinatensystem, d.h. <zone> ist nicht relativ zu <surface>!
  - D.h. auch, dass <zone> theoretisch einen größeren Bereich als <surface> beschreiben kann
- <surface> kann mehrere <graphic>-Elemente enthalten, z.B. Bilder in verschiedenen Auflösungen
- Mehrere <surface>-Elemente können mit <surfaceGrp> gruppiert werden
  - Z.B. Blätter, Lagen oder zusammengebundene Hefte
  - beliebig tiefe Schachtelung
- Weitere Attribute zur Spezifizierung
  - @type -> z.B. "printedInitial", "column", "figure", "word" usw.
  - @attachment -> z.B. attachment="glued" für aufgeklebte Zettel
  - @flipping -> gefaltet und mit zweiseitig beschrieben

<surface/> und <zone/>

- <zone> kann ein <graphic>-Element enthalten, z.B. eine Detailaufnahme
  - Damit können z.B. auch Doppelseiten als eine <surface> definiert werden
- <zone> kann auch vieleckige Zonen definieren (@points)
  - Orientiert an SVG-Standard
  - Details in den Guidelines
  - Vorsicht: mit XSLT schwierig auszuwerten!

- **<zone> kann bzw. sollte mit dem Attribut @xml:id versehen werden!**
- Ggf. auch <surface>

Gliederung

- Grundlagen der Bilddigitalisierung
- Digitale Faksimiles in TEI definieren
- **Text und Bild verbinden**
- Hilfsmittel
- Darstellung von Text-Bild-Verbindungen

Parallele Transkription

- <zone>-Elemente werden mit @xml:id eindeutig referenzierbar gemacht
- Gegliederter TEI-Text verweist mit @facs in den entsprechenden Elementen auf die Zonen
- Kann beliebig fein gegliedert werden, theoretisch bis zur Buchstabenebene
- Vorteile
  - Flexibel
  - Text kann mit allen Kunstgriffen der TEI ausgezeichnet werden
  - Text und Bild bleiben getrennt und können theoretisch in zwei Dateien aufgeteilt werden
- Nachteile
  - Durch Flexibilität der Textauszeichnung z.T. problematisch in der Darstellung bzw. bei der Bearbeitung mit XSLT
  - Flexibilität des <graphic>-Elements (bezieht es sich auf eine <zone> oder eine <surface>?)

Beispiel <facsimile> und <text>

```
<facsimile>
    <surface ulx="120" uly="150" lrx="860"
    lry="1140" xml:id="s002r">
          <!-- Satzspiegel der Seite -->

          <graphic url="M29857-002r.jpg"/>
          <!-- Verweist auf die Bilddatei-->

          <zone xml:id="z002r001" ulx="120"
    uly="150" lrx="480" lry="1125"/>
          <!-- Spalte 1 -->

          <zone xml:id="z002r002" ulx="500"
    uly="150" lrx="860" lry="1125"/>
          <!-- Spalte 2 -->

          <zone xml:id="z002r003" ulx="775"
    uly="1130" lrx="860" lry="1140"/>
          <!-- Bogensignatur aij -->

    </surface>
</facsimile>
```

```
<text>
  <body>
    <div>
      <pb n="2r" facs="#s002r"/>
      <cb n="1"  facs="#z002r001"/>
      <head>Hie sagt es von dem anfang
      sant Pauls des ersten einsidel.</head>
      <p>In der zit da decius der keyser zuo
         rom regieret vnd Valerius da
         was die durchachtung also groß über
         die heiligen christenheyt das
         cornelius der heilig bast und
         ciprianus der heylig bischoff von
         carthagine ....
         <cb n="2"  facs="#z002r002"/>
         von dem christenlichen gelauben
         bringen so hieß im der richter die
         hend hinderlich binden ...
         noch füß mocht entbinden das
         <fw type="signatureMark"
             facs="#z002r003">a ij</fw>
         <pb n="2v"/>
         ...
      </p>
    </div>
  </body>
</text>
```

Beispiel: 003 parallelTranscription.xml

Eingebettete Transkription

- Element <sourceDoc> (Kindelement v. <TEI>) speziell für Transkriptionen eines Quelldokumentes (z.B. einer Edition oder einer Quellensammlung)
- Innerhalb von <sourceDoc> stark eingeschränktes Inventar von Elementen
- Innerhalb von <sourceDoc> können <surface> und <zone> definiert werden
- Transkription erfolgt direkt in den <zone>-Elementen, ggf. untergliedert durch <line> oder <seg>-Elemente (Kinder v. <zone>)
- Vorteile:
  - Direkter Ansatz
  - Reduktion von möglichen Elementen schließt Interpretationen aus
  - Keine Milestone-Elemente zur Verknüpfung mit den Bildern, einfacher in der Verarbeitung mit XSLT
- Nachteile:
  - Unübersichtlich, Vermischung von Bild und Textinformationen

Beispiel <sourceDoc>

```
<sourceDoc>
   <surface ulx="120" uly="150" lrx="860" lry="1140">
      <graphic url="M29857-002r.jpg"/>
      <zone ulx="120" uly="150" lrx="480" lry="1125"
            type="column"> <!-- Spalte 1 -->
            <line>Hie sagt es von dem anfang</line>
            <line>sant Pauls des ersten einsidel.</line>
            <line><seg>I<seg>n der zit da decius</line>
            <line>der keyser zuo rom re</line>
            <line>gieret vnd Valerius</line>
      </zone>
   </surface>
</sourceDoc>
```

Beispiel: 004 sourceDoc.xml

Gliederung

- Grundlagen der Bilddigitalisierung
- Digitale Faksimiles in TEI definieren
- Text und Bild verbinden
- **Hilfsmittel**
- Darstellung von Text-Bild-Verbindungen

Bildbetrachter

- Z.B. IrfanView (http://www.irfanview.de/)
  - Öffnet viele Bildformate
  - Kann viele Formate speichern
  - Batchkonvertierung/-umbenenung (z.B. um Thumbnails oder Präsentationsbilder zu erzeugen)
  - Informationen über Auflösung, Farbtiefe usw.
  - Zur Not auch, um Bildkoordinaten zu bestimmen (ulx, uly, lrx, lry)
  - Verschiedene Filter
  - Keine avancierten Bildbearbeitungsmöglichkeiten

Image Markup Tool

- Tool zum Erzeugen von <facsimile>-Gruppen (http://www.tapor.uvic.ca/~mholmes/image_markup/index.php)
  - Graphische Oberfläche
  - Relativ leicht zu bedienen
  - Erzeugt valides TEI
  - Ggf. als Grundlage für eigene Weiterbearbeitung zu verwenden
  - Beschränkt auf je ein Bild

Beispiel: 006 IMT.xml

Beispieltransformation: 006 IMT2html.xsl

Gliederung

- Grundlagen der Bilddigitalisierung
- Digitale Faksimiles in TEI definieren
- Text und Bild verbinden
- Hilfsmittel
- **Darstellung von Text-Bild-Verbindungen**

Darstellung von Text-Bild-Verbindungen

- Die Darstellung von Text und Bild ist vom Ziel des Projekts abhängig!
  - Bei einer sorgfältigen Textedition können eingestreute Bilder ablenken und sollten eher nur der Dokumentation der Arbeit dienen
  - Manche Rezipienten wollen einfach nur einen Text lesen!
  - Für linguistische Textcorpora sind Bilder nicht unbedingt notwendig
  - Text-Bild-Synopsen sind z.B. bei Transkriptionen oder für die Lehre sinnvoll
  - Oberfläche bei Webpräsentationen nicht überfrachten
- Denken sie an rechtliche Probleme! Dürfen sie 'ihre' Bilder im Web anzeigen?
- Für die Darstellungsebene zwingend erforderlich:
  - **Konsistente (und dokumentierte) Auszeichnung der TEI-Quelle**
  - XSLT, HTML, CSS
- Sinnvoll:
  - Javascript, PHP, Python o.ä. für interaktive Funktionen

Beispiele

- Bilder im Fließtext integriert
  - Beispiel: http://teiviewer.org/examples/manuscript/thjostolfs.xml
- Text und Bild parallel
  - Beispiel: 003 parallelTranscription.html
  - www.deutschestextarchiv.de
- Interaktiv
  - Beispiel: 005 wortkoordinaten.xml
  - Beispieltransformation: 005 facsimile2html.xsl
  - Problem: Das ist (fast) nur möglich, wenn die @facs-Attribute in nicht-leeren Elementen angegeben werden (also nicht in <pb/> o.ä.)

Übungen

- 1)
  - Definieren sie für ein Bild ihrer Wahl manuell ein surface-Element mit mehreren Zonen, transkribieren sie etwas Text und verknüpfen sie beide (parallele oder eingebettete Transkription)

- 2)
  - Laden sie das Image Markup Tool herunter, installieren sie es und bearbeiten Sie ein Bild ihrer Wahl (oder das im Beispielverzeichnis liegende)
  - Sehen sie den vom IMT erzeugten TEI-Code an und vollziehen sie ihn nach