

# Logische Dokumentenanalyse - OCR-Datenformate

---

Andreas Neumann M.A.  
Centrum für Informations- und Sprachverarbeitung

# Ausgangsmaterial

# Ausgangsmaterial kann vorliegen als:

---

- Image
- Plain Text
- Text mit Strukturinformationen
- Text mit Positionsangaben

# Image

---

- Verfahren zur direkten Analyse auf der Grafik
- Diese Verfahren sind auch Teil des Preprocessing-Prozesses bei der OCR

# Plain Text

---

- Textdatei
- Physisches Layout nachgebildet:
  - meist entspricht ein File einer Seite
  - Zeilen durch \n getrennt
  - Wörter durch „ “ getrennt
- Exakter Seitenaufbau verloren

# Strukturierter Text

---

- XML artig
- Physische Strukturen durch Tags angegeben:
  - Table
  - Line
  - ...
- Qualität stark von Erkennung abhängig
- Beispiel: ePub, HTML
- Ähnlich zum Original, aber Informationsverlust der genauen Positionierung

# Strukturierter Text mit Positionsangaben

---

- meist Text mit Strukturinformation erweitert um Koordinaten bei den Strukturen
- Koordinaten bilden eine sogenannte Bounding Box, die es erlaubt Position auf der Seite zu bestimmen
- Erlaubt eine eigene Interpretation des physischen Aufbaus

# Text mit typographische Angaben

---

- Schriftgröße
- Type
- Kann mit strukturiertem Text und Text mit Positionsangaben auftreten
- Z.B. Bei HTML mit CSS

# Strukturierter Text mit Positionsangaben

---

- Es existieren verschiedene Standards um Text strukturiert mit Positionsangaben und typographische Information zu kodieren
- Verbreitet sind:
  - hOCR (Format der Google OCR Daten)
  - Abbyy XML (Formate der Firma Abbyy)
  - Alto XML (Im Bibliotheksumfeld entstanden)
- Ein allgemeingültiger Standard existiert nicht

# Layoutelemente aus der OCR-Analyse

---

- Moderne OCR-Software benutzt eine Analyse bestimmter Layout Elemente um das OCR-Ergebnis zu verbessern
- Diese Elemente können auch für weitere Analyse nützlich sein
- Die Analyse ist nicht immer korrekt (Nachkorrektur)

# Typische Layoutelemente aus der OCR-Analyse

---

- Page
- Block
- Paragraph
- Line
- Word
- (Character)

# Page and Block

Block1	Page
38	<b>Avant-bec — Aviver.</b>
<b>Avant-bec</b> m. <i>Voy.</i> Brise-glace. <b>Avant-chemin-couvert</b> m. (chemin couvert fait au pied du glacis, et qui se trouve le plus avancé dans la campagne) (Fort.) <i>Der zweite gedeckte Weg.</i> Second or advanced covert-way.	<b>Block2</b>
<b>Avant-coulant</b> m., <b>Première eau-de-vie</b> f. <i>Der Vorlauf.</i> First running of brandy, first short. <i>Voy.</i> Eau-de-vie.	<b>Block3</b>
<b>Avant-creuset</b> m. d'un haut-fourneau (espace creux qui se trouve avant le creuset) <i>Der Vorherd.</i> Breast-pan.	<b>Block3</b>
<b>Avant-fossé</b> m. (fossé fait au pied du glacis et qui précède immédiatement l'avant-chemin-couvert.) (Fort.) <i>Der Vorgraben, Aussengraben.</i> Advanced or second ditch, avant-fosse.	
<b>Avant-garde</b> f. d'une armée navale (division qui, en ligne de bataille, se trouve en avant de deux autres) (Mar.) <i>Das Vordertreffen, die Avantgarde, die vordere Schlachlinie.</i> Van of a fleet, van-guard, first-line, first-division.	
<b>Avant-glacis</b> m. (glacis de l'avant-chemin-couvert) (Fort.) <i>Das Vor-Glacis.</i> Advanced glacis.	
<b>Avant-mur</b> m., <b>Barbacane</b> f. (Fort.) <i>Die Zwingermauer.</i> Barbican, barbican. <i>Voy.</i> Barbacane.	
<b>Avant-port</b> m. (entrée d'un grand-port au dehors de son enceinte) (Mar.) <i>Der Butenhafen, Aussenhafen, Vorhafen.</i> Outer-harbour.	
<b>Avant-titre</b> m., <b>Faux-titre</b> m. d'un livre (titre abrégé, à la fausse page) (Impr.) <i>Der Schmutztitel.</i> Bastard-title.	
<b>Avant-train</b> m. d'une voiture à quatre roues (cette partie de la voiture où se trouvent l'essieu de devant, les roues d'avant-train et le timon) (Charr.) <i>Der Vorderwagen.</i> Fore-carriage.	
<b>Avant-train</b> m. d'un affût (train qui comprend les roues de devant et le timon d'un canon de campagne) (Artill.) <i>Die Protze, der Vorderwagen, die Geschützprotze.</i> Limber, gun-limber.	
<b>Avant-train à bras de limonière.</b> <i>Der Vorderwagen mit Gabeldeichsel.</i> Limber with shafts.	
<b>Avant-train à timon.</b> <i>Der Vorderwagen</i>	<b>Aviron</b> m. (rame pour les nacelles) (Mar.) <i>Das</i>

# Page

---

- Beschreibt die physische Buchseite
- Dienst als Bezugsrahmen für weiter Layoutelmente
- Seitenzahl optionales Element
- Umfasst Boxes, Paragraphs, Lines und Words

# Block

---

- Textblock
- Beschreibt zusammengehörige Layoutelemente
- Umfasst Paragraphs, Lines und Words

# Paragraph

---

- Absatz
- Umfasst Lines und Words

**Bâtiment m. à un mât** (Mar.) *Das einmastige Fahrzeug.* Sloop or vessel with a single mast.

Paragraph

**Bâtiment m. à vapeur.** *Das Dampfschiff.* Steam-ship, steam-boat, steamer.

**Bâtiment m. additionnel, Appentis m.** (Arch.) *Der Anbau, Nebenbau.* Additional building, out-house. *Compar.* Appentis 2.

**Bâtiment m. de graduation** (hangar très-long, assez élevé, ouvert à tout vent, pour l'évaporation de l'eau des sources salées) (Sal.) *Das Gradirhaus, Gradirwerk.* Graduation-house.

# Line

---

- Zeile
- Umfasst Words

**Lineiment m. à un mât** (Mar.) *Das einmastige Fahrzeug.* Sloop or vessel with a single mast.

**Bâtimen<sup>t</sup> t m. à vapeur.** *Das Dampfschiff.*  
Steam-ship, steam-boat, steamer.

**Bâtimen<sup>t</sup> t m. additionnel, Appentis m.**  
Arch.) *Der Anbau, Nebenbau.* Additional building, out-house. *Compar.* Appentis 2.

**Bâtimen<sup>t</sup> t m. de graduation** (hangar très-long, assez élevé, ouvert à tout vent, pour l'évaporation de l'eau des sources salées) (Sal.) *Das Gradirhaus, Gradirwerk.* Graduation-house.

# Word

---

- Wort
- Wortbegriff bei AbbyyXML explizit nicht vorhanden

**Bâtiment m. à un mât (Mar.) Das ein-mastige Fahrzeug Sloop or vessel with a single**  
Word 1 Word 2 Word 3

**Bâtiment m. à vapeur. Das Dampfschiff.**  
Steam-ship, steam-boat, steamer.

**Bâtiment m. additionnel, Appentis m.**  
Arch.) Der Anbau, Nebenbau. Additional building, out-house. Compar. Appentis 2.

**Bâtiment m. de graduation** (hangar très-long, assez élevé, ouvert à tout vent, pour l'évaporation de l'eau des sources salées) (Sal.) **Das Gradirhaus, Gradirwerk.** Graduation-house.

# Character

---

- Zeichen
- Ein Word umfasst mehrere Characters
- Beispielsweise bei Abbyy-XML vorhanden

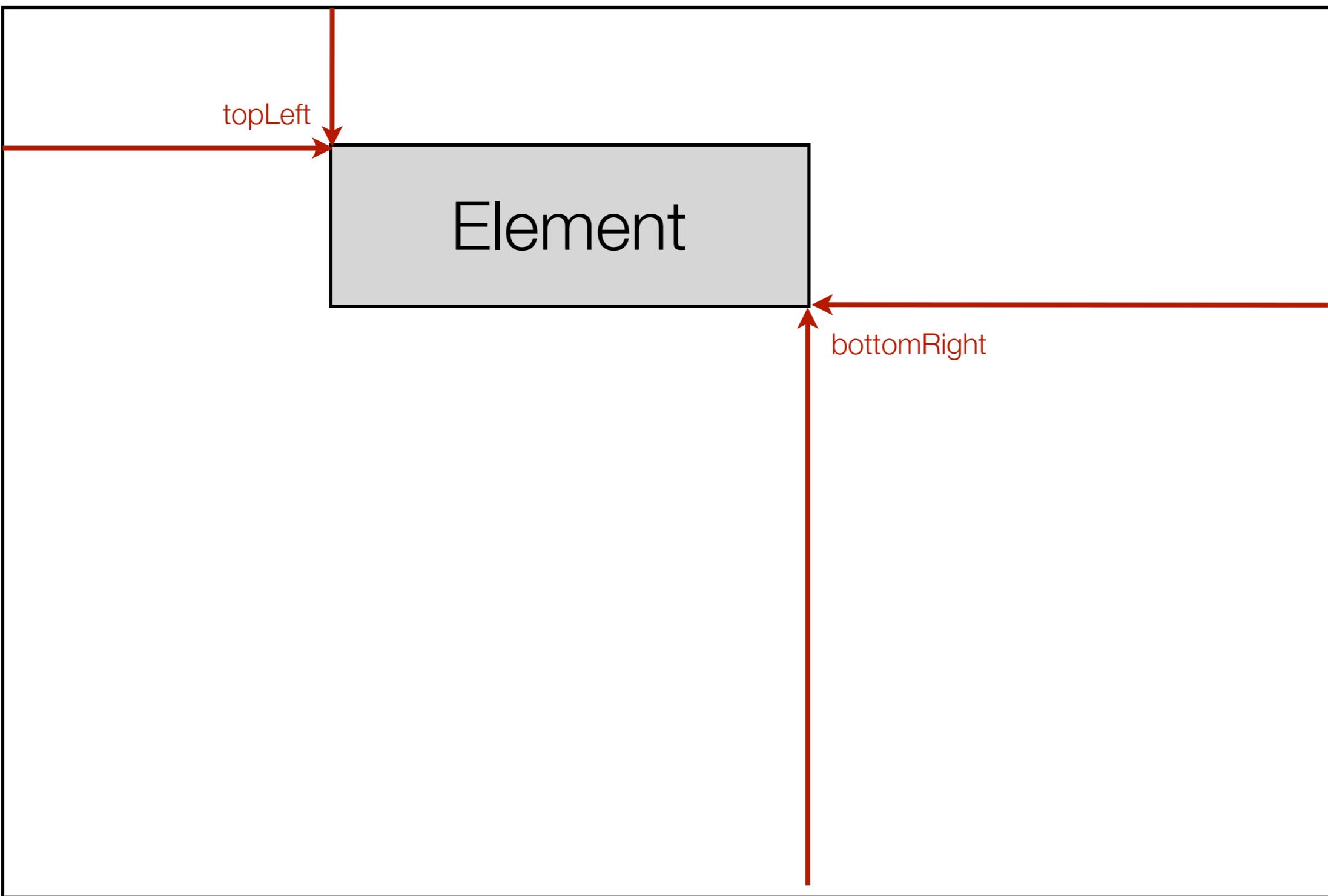
# BoundingBox

---

- Möglichkeit Position relativ zu kodieren
- Meist über zwei Punkte aufgespannt, oder ein Punkt + width und height
- zB. topLeft, bottomRight
- erlaubt das errechnen von Position in Dokument
- erlaubt die umfasste Fläche zu bestimmen
- es existieren verschiedene Varianten

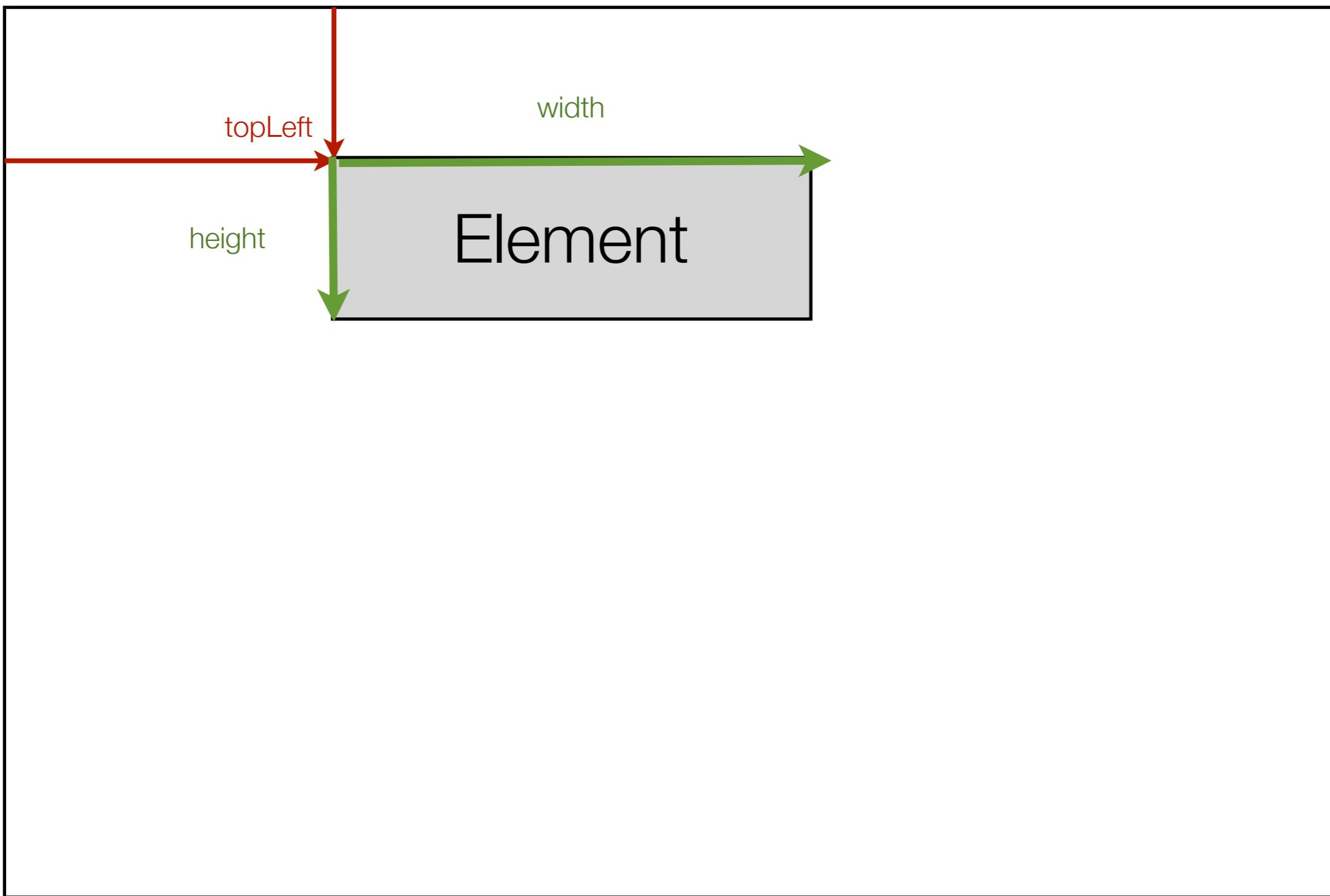
# Bounding Box

---



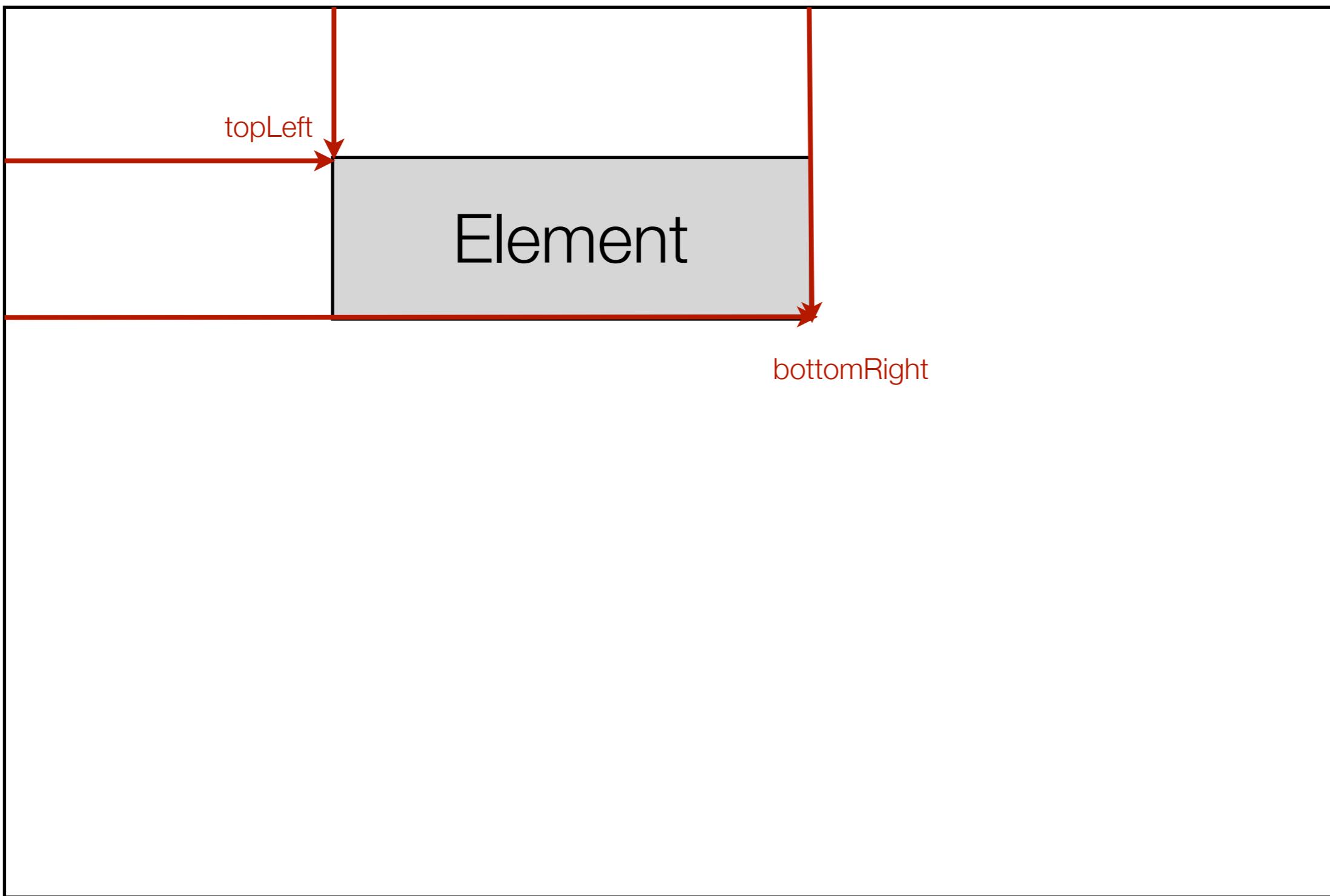
# Bounding Box

---



# Bounding Box

---



# Seite mit Bounding Boxes

---

Page: 20



```
<span data-text="Abgewöhren" class="OCRWord" data-features style="position:absolute; top:85px; left:125px; height: 25px; width:164px;"></span>
```

# Seite mit Bounding Boxes und Grafik

Page: 20

10284595

Ab

Ab

Abgewähren oder abgewehren heißt einen Anr im Gespenk Buche dem einen ab, und dem andern zuschreiben. Der Zuschriftebezeddel heißt der Abgewährzeddel, oder der Gewährschein. Deutsche Encycl. I. S. 46. Bergm. Wörterb. S. 3. s. Gewährschein.

Abgewehren, s. abgewähren,

Abgewehrzeddel s. abgewähren, und Gewährschein.

Abgewärmt oder abgeärdt wird von den Kapellen gesagt, wenn sie vor ihrem Gebrauch im Probierofen ausgeglüht werden. Deutsche Encycl. I. S. 45. S. Abhähnen und Abwärmen.

Abglühen, heißt überhaupt etwas durchs Feuer glüend machen; insonderheit wird dieser Ausdruck von den Metallen gebraucht, und heißt so viel, als Metalle im Feuer geschmeidig machen. Wenn nemlich ein Metall durch das Schlagen spröde geworden ist, so macht man es dann im Feuer glüend und läßt es wieder kalt werden. Diese Arbeit heißt das Abglühen. Bergm. Wörterb. S. 3. Deutsche Encycl. I. S. 47.

Abhängen, oder Abhangen, heißt die Bälge, welche durch die Radwelle bewegt werden, los machen, daß sie nicht mehr spielen, oder blasen. Bergm. Wörterb. S. 4.

<span data-text="Abgewähren" class="OCRWord" data-features="position:absolute; top:85px; left:125px; height: 25px; width:164px;"></span>

# Verbreitete Formate

---

- Abbyy-XML
- Alto-XML
- hOCR

# AbbyyXML

---

- Basiert auf XML
- Kennt keinen Wortbegriff
- bis auf Zeichenebene,

# AbbyyXML - Basisimage

---

90

## Entstehung der Materie.

---

und Empfindung seiner Wirkungen als Eigentum gegeben hat.

Aus dieser allmächtigen Quelle sind auch die Kräfte genommen worden, die der Vater aller Geister, der Macht des freien Willens verliehen hatte, und welche die Freiheit des Willens durch die Wissenschaft des Bösen missbrauchte; Sie haben der in sie wirkenden Urkraft widerstanden, und sind von ihr geschieden worden.

# AbbyyXML - Beispieldaten

```
<document xmlns="http://www.abbyy.com/FineReader_xml/FineReader8-schema-v2.xml" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="1.0" producer="FineReader 8.0" xsi:schemaLocation="http://www.abbyy.com/FineReader_xml/FineReader8-schema-v2.xml http://www.abbyy.com/FineReader_xml/FineReader8-schema-v2.xml">  
  <mainLanguage>OldGerman</mainLanguage>  
  <languages>OldGerman</languages>  
  <page width="6112" height="4721" resolution="600" originalCoords="true">  
    <block blockType="Text" blockName="" isHidden="true" l="1143" t="416" r="2547" b="541">...</block>  
    <block blockType="Text" blockName="" isHidden="true" l="1054" t="687" r="2884" b="3748">...</block>  
    <block blockType="Text" blockName="" isHidden="true" l="3673" t="406" r="4793" b="525">  
      <region>  
        <rect l="3673" t="406" r="3841" b="407"/>  
        <rect l="3673" t="407" r="4236" b="408"/>  
        <rect l="3673" t="408" r="4631" b="409"/>  
        <rect l="3673" t="409" r="4793" b="522"/>  
        <rect l="3838" t="522" r="4793" b="523"/>  
        <rect l="4233" t="523" r="4793" b="524"/>  
        <rect l="4628" t="524" r="4793" b="525"/>  
      </region>  
      <text>  
        <par align="Justified">  
          <line baseline="491" l="3686" t="418" r="4781" b="504">  
            <formatting lang="OldGerman">  
              <charParams l="3686" t="425" r="3734" b="484" suspicious="true" characterHeight="46" hasUncertainHeight="false" baseline="0" wordStart="true" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" wordPenalty="7" meanStrokeWidth="77" charConfidence="63" serifProbability="255"/><span style="color:red">Beispielwort</charParams>  
              <charParams l="3754" t="422" r="3807" b="489" characterHeight="46" hasUncertainHeight="false" baseline="0" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" wordPenalty="7" meanStrokeWidth="77" charConfidence="96" serifProbability="255"/>E</charParams>  
              <charParams l="3818" t="437" r="3859" b="486" characterHeight="46" hasUncertainHeight="false" baseline="0" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" wordPenalty="7" meanStrokeWidth="77" charConfidence="80" serifProbability="255"/>n</charParams>  
              <charParams l="3876" t="433" r="3897" b="487" characterHeight="46" hasUncertainHeight="false" baseline="0" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" wordPenalty="7" meanStrokeWidth="77" charConfidence="90" serifProbability="255"/>t</charParams>  
              <charParams l="3915" t="425" r="3955" b="502" characterHeight="46" hasUncertainHeight="false" baseline="0" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" wordPenalty="7" meanStrokeWidth="77" charConfidence="94" serifProbability="255"/>s</charParams>  
              <charParams l="3915" t="425" r="3955" b="502" characterHeight="46" hasUncertainHeight="false" baseline="0" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" wordPenalty="7" meanStrokeWidth="77" charConfidence="94" serifProbability="255"/>t</charParams>  
              <charParams l="3972" t="443" r="3995" b="486" characterHeight="46" hasUncertainHeight="false" baseline="0" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" wordPenalty="7" meanStrokeWidth="77" charConfidence="96" serifProbability="255"/>e</charParams>  
              <charParams l="4011" t="419" r="4050" b="499" characterHeight="46" hasUncertainHeight="false" baseline="0" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" wordPenalty="7" meanStrokeWidth="77" charConfidence="96" serifProbability="255"/>h</charParams>  
              <charParams l="4070" t="439" r="4104" b="486" characterHeight="46" hasUncertainHeight="false" baseline="0" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" wordPenalty="7" meanStrokeWidth="77" charConfidence="94" serifProbability="255"/>u</charParams>  
              <charParams l="4125" t="440" r="4161" b="485" characterHeight="46" hasUncertainHeight="false" baseline="0" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" wordPenalty="7" meanStrokeWidth="77" charConfidence="94" serifProbability="255"/>n</charParams>  
              <charParams l="4176" t="439" r="4212" b="499" characterHeight="46" hasUncertainHeight="false" baseline="0" wordStart="false" wordFromDictionary="true" wordNormal="true" wordNumeric="false" wordIdentifier="false" wordPenalty="7" meanStrokeWidth="77" charConfidence="90" serifProbability="255"/>g</charParams>
```

# AltoXML

---

- Basiert auf XML
- Häufig mit METS-Daten (Metadata Encoding and Transmission Standard) zusammen verwendet
- findet sich im bibliothekarischen Umfeld
- Standard wird von der Library of Congress verwaltet

# AltoXML - Abschnitte und sonstige Besonderheiten

---

## <description> - Abschnitt

Hier finden sich Metadaten zum beschriebenen Dokument.

## <Styles> - Abschnitt

Im Style Abschnitt werden alle Informationen zu Typen und Layout kodiert.

- <TextStyle> : Informationen zu der verwendeten Schriftart
- <ParagraphStyle> : Layoutinformationen, wie Ausrichtung

## Komplettüberblick über Alto-XML-Styles

## <Layout> - Abschnitt

Im Layoutabschnitt findest sich wiedersprüchlicherweise der Inhalt des Dokuments. Ein Dokument besteht aus einer oder mehreren <Page>-Elementen.

## Maßeinheiten

Positionasangaben werden in 1/10mm oder 1/1200inch gemacht.

## Umrechnung in Pixel

- Bei Angabe in 1/1200inch: Wert \* Auflösung des Dokuments / 1200
- Bei Angabe in 1/100mm: Wert \* Auflösung des Dokuments / 254

# AltoXML - Grafik

Beispielwort

## Planet Earth

Each planet of the Solar system is unique in its own right, yet Earth has a whole set of really unique features. First, it is the only active planet – the earthquakes and volcano eruptions constantly change its appearance. Second, it is the only planet that boasts vast resources of liquid water: it's too hot for that on Venus and too cold on Mars. The Earth's atmosphere is also very unlike any other planet's gaseous shells. Earth neighbors' atmospheres consist mainly of carbon dioxide, while the Earth's contains great amounts of oxygen and nitrogen, which form shield from the most dangerous components of solar radiation. The Earth's atmosphere also protects the planet from meteorites. More so, it is this unique combination of constantly changing land surface, oceans and aerial shield that made it possible for another unique phenomenon – Life – to exist on Earth.



*Our planet photographed from "Apollo - 17". You can notice the big ice polar cap in Antarctica.*



*Grand Canyon: the stream cut through the layers of soft sandstone and limestone in Arizona (USA) into a*

### Seasons

Each 24 hours the Earth completes a full revolution around its axis, which is inclined by 23.5° to the vertical. This inclination is the reason why the seasons change on the Earth as it rotates around the Sun.

### Structure

The central part of Earth is a metal core; it's very hot – some 4000°C, and it's surrounded by a shell of liquid iron that creates the magnetic field of Earth. Outer layers form the mantle made up of rocky substances, over which are lighter substances that form the crust. The atmosphere is made of nitrogen (77%), oxygen (21%), and a mixture of water vapor and other gases.

### Magnetic bubble

The rotation of Earth around its axis generates forceful electrical currents in the iron core of the planet and this creates the magnetic field. This field forms a giant "bubble" in the near-Earth space called the "magnetosphere". Magnetosphere protects Earth from the solar wind – a flow of charged particles emitted by the Sun. These particles are trapped by the magnetic field in two huge rings – Van Allen's belts. When spacecrafts travel through the Van Allen's belts, the electrical equipment of the former may suffer malfunction caused by these particles.

### Clashing Continents

The Earth crust is made from parts called plates, which float on its surface driven by the flows in the liquid mantle. The continents lie on these plates, and so their location is subject to constant change. Some 200 million years ago, all the dry land on Earth was a single continent called Pangea by the scientists, which further split into the continents we know now. The lava rises by millimeters around the mountain ridges located on the ocean floor, and moves the continents apart. When the continents clash, as they do around the shores of the Pacific, the

# AltoXML - auf echten Daten

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<alto xmlns="http://www.loc.gov/standards/alto/alto-v2_0.xsd">
  <Description>
    <MeasurementUnit>pixel</MeasurementUnit>
    <OcrProcessing>
      <ocrProcessingStep>
        <processingDateTime>129556338579510000</processingDateTime>
        <processingSoftware softwareCreator="ABBYY"/>
      </ocrProcessingStep>
    </OcrProcessing>
  </Description>

  <Styles>
    <ParagraphStyle ID="{FFFFFFF-FFFF-FFFF-FFFF-FFFFFFF}" ALIGN="Left" LEFT="0" RIGHT="0" FIRSTLINE="0"/>
    <ParagraphStyle ID="{28C382FD-5FBB-41B6-96AD-C573415A8A04}" ALIGN="Block" LEFT="0" RIGHT="0" FIRSTLINE="0" LINESPACE="0"/>
    <ParagraphStyle ID="{43E10D91-0C1B-4382-9781-16B0D68F5E73}" ALIGN="Block" LEFT="0" RIGHT="0" FIRSTLINE="0" LINESPACE="1176"/>
    <ParagraphStyle ID="{2C449421-5C92-44CB-9B45-0ACB05AFBF5B}" ALIGN="Left" LEFT="0" RIGHT="0" FIRSTLINE="0" LINESPACE="0"/>
    <ParagraphStyle ID="{2871F514-20E8-483C-81A4-8E257FD06E0B}" ALIGN="Block" LEFT="0" RIGHT="0" FIRSTLINE="0" LINESPACE="1056"/>
    <ParagraphStyle ID="{9C6696A1-BBD2-4D79-83D2-86B0CC1E0383}" ALIGN="Block" LEFT="0" RIGHT="0" FIRSTLINE="0" LINESPACE="1224"/>
    <ParagraphStyle ID="{29FCBACF-E1F9-4BA3-A1AD-D3DDAFDE4511}" ALIGN="Left" LEFT="0" RIGHT="0" FIRSTLINE="0" LINESPACE="0"/>
  </Styles>

  <Layout>
    <Page ID="0" PHYSICAL_IMG_NR="0">
      <PrintSpace>
        <ComposedBlock ID="{358274E0-5271-4FA1-9881-9F7C12478045}" HEIGHT="3508" WIDTH="2479" VPOS="0" HPOS="0" TYPE="container">
          <textBlock ID="{5925CDD5-FABD-4624-A5A8-61F094ABCDC2}" HEIGHT="602" WIDTH="2086" VPOS="187" HPOS="183" STYLEREFS="{9C6696A1-BBD2-4D79-83D2-86B0CC1E0383}">
            <TextLine HEIGHT="84" WIDTH="692" VPOS="192" HPOS="196" STYLEREFS="{28C382FD-5FBB-41B6-96AD-C573415A8A04}">
              <String STYLE="bold" CONTENT="Planet" WIDTH="354" VPOS="192" HPOS="196"/>
              <SP WIDTH="44" VPOS="192" HPOS="551"/>
              <String STYLE="bold" CONTENT="Earth" WIDTH="292" VPOS="192" HPOS="596"/>
            </TextLine>
            <TextLine HEIGHT="42" WIDTH="2072" VPOS="341" HPOS="190">
              <String STYLE="bold" CONTENT="Each" WIDTH="94" VPOS="342" HPOS="190"/>
              <SP WIDTH="13" VPOS="343" HPOS="285"/>
              <String STYLE="bold" CONTENT="planet" WIDTH="116" VPOS="343" HPOS="299"/>
              <SP WIDTH="13" VPOS="345" HPOS="416"/>
              <String STYLE="bold" CONTENT="of" WIDTH="37" VPOS="342" HPOS="430"/>
              <SP WIDTH="9" VPOS="342" HPOS="468"/>
              <String STYLE="bold" CONTENT="the" WIDTH="57" VPOS="343" HPOS="478"/>
              <SP WIDTH="13" VPOS="341" HPOS="536"/>
              <String STYLE="bold" CONTENT="Solar" WIDTH="98" VPOS="341" HPOS="550"/>
            </TextLine>
          </textBlock>
        </ComposedBlock>
      </PrintSpace>
    </Page>
  </Layout>
</alto>
```

Description

Style

Layout

# AltoXML - Der Layout-Abschnitt

```
<Layout>  
  <Page ID="0" PHYSICAL_IMG_NR="0">  
    <PrintSpace>  
      <ComposedBlock ID="{358274E0-5271-4FA1-9881-9F7C12478045}" HEIGHT="3508" WIDTH="2479" VPOS="0" HPOS="0" TYPE="container">  
        <textBlock ID="{5925CDD5-FABD-4624-A5A8-61F094ABCDC2}" HEIGHT="602" WIDTH="2086" VPOS="187" HPOS="183" STYLEREFS="{9C6696A1-BBD2-4D79-83D2-86B0CC1E0383}">  
          <TextLine HEIGHT="84" WIDTH="692" VPOS="192" HPOS="196" STYLEREFS="{28C382FD-5FBB-41B6-96AD-C573415A8A04}">  
            <String STYLE="bold" CONTENT="Planet" WIDTH="354" VPOS="192" HPOS="196"/>  
            <SP WIDTH="44" VPOS="192" HPOS="551"/>  
            <String STYLE="bold" CONTENT="Earth" WIDTH="292" VPOS="192" HPOS="596"/>  
          </TextLine>  
        <TextLine HEIGHT="42" WIDTH="2072" VPOS="341" HPOS="190">  
          <String STYLE="bold" CONTENT="Each" WIDTH="94" VPOS="342" HPOS="190"/>  
          <SP WIDTH="13" VPOS="343" HPOS="285"/>  
          <String STYLE="bold" CONTENT="planet" WIDTH="116" VPOS="343" HPOS="299"/>  
          <SP WIDTH="13" VPOS="345" HPOS="416"/>  
          <String STYLE="bold" CONTENT="of" WIDTH="37" VPOS="342" HPOS="430"/>  
          <SP WIDTH="9" VPOS="342" HPOS="468"/>  
          <String STYLE="bold" CONTENT="the" WIDTH="57" VPOS="343" HPOS="478"/>  
          <SP WIDTH="13" VPOS="341" HPOS="536"/>  
          <String STYLE="bold" CONTENT="Solar" WIDTH="98" VPOS="341" HPOS="550"/>  
          <SP WIDTH="13" VPOS="352" HPOS="649"/>  
          <String STYLE="bold" CONTENT="system" WIDTH="124" VPOS="345" HPOS="663"/>  
          <SP WIDTH="14" VPOS="342" HPOS="788"/>  
          <String STYLE="bold" CONTENT="is" WIDTH="27" VPOS="342" HPOS="803"/>  
          <SP WIDTH="12" VPOS="352" HPOS="831"/>  
          <String STYLE="bold" CONTENT="unique" WIDTH="129" VPOS="342" HPOS="844"/>  
          <SP WIDTH="12" VPOS="342" HPOS="974"/>  
          <String STYLE="bold" CONTENT="in" WIDTH="34" VPOS="342" HPOS="987"/>  
          <SP WIDTH="13" VPOS="342" HPOS="1022"/>  
          <String STYLE="bold" CONTENT="its" WIDTH="42" VPOS="342" HPOS="1036"/>  
          <SP WIDTH="13" VPOS="352" HPOS="1079"/>  
          <String STYLE="bold" CONTENT="own" WIDTH="75" VPOS="352" HPOS="1093"/>  
          <SP WIDTH="14" VPOS="352" HPOS="1169"/>  
          <String STYLE="bold" CONTENT="right," WIDTH="102" VPOS="342" HPOS="1184"/>  
          <SP WIDTH="11" VPOS="353" HPOS="1287"/>  
          <String STYLE="bold" CONTENT="yet" WIDTH="56" VPOS="345" HPOS="1299"/>  
          <SP WIDTH="11" VPOS="342" HPOS="1356"/>  
          <String STYLE="bold" CONTENT="Earth" WIDTH="109" VPOS="342" HPOS="1368"/>  
        </TextLine>  
      </ComposedBlock>  
    </PrintSpace>  
  </Page>
```

# hOCR

---

- hybrides Format auf Basis von HTML4
- versucht als HTML in der Darstellung dem Originaldokument nahe zu kommen
- intern werden weitere, für das Dokumentenverstehen wichtige Daten gespeichert
- findet Einsatz im Google Digitalisierungsprojekt
- Spezifikation unter: [https://docs.google.com/document/preview?id=1QQnIQtvdAC\\_8n92-LhwPcjAUFWBlzE8EWnKAxlgVf0&pli=1](https://docs.google.com/document/preview?id=1QQnIQtvdAC_8n92-LhwPcjAUFWBlzE8EWnKAxlgVf0&pli=1)

# hOCR - Grafik als Basis des OCR

**W**Un höret zu und schweiget still / habt  
acht was ich euch singen will/von einem  
guten Lande/ so blieb mancher daheim  
nicht/ wann ihm das wäre bekandte. Beispiel

2. Der Weg der ist auch ziemlich weit/ junge  
Kinder und alte Leut/ machen sich dadurch  
viel Kommer/ im Winter ist es ihnen zu kalt/  
und auch zu heiß im Sommer.

3. Die Gegend heißt Schlauraffenland/  
ist faulen Leuten wol bekandt/ red ich ohn allen  
Schaden/ darinnen sind die Häuser gedect/  
mit lauter Eyer-Fladen.

4. Welche Mägde oder Gesellen/ des Lan-  
des Art erfahren wollen/mögen sich dahin ver-  
fügen/ wann man die Dächer bricht ab/ hat  
er Fladen voll Genügen.

5. Thür und Wände das ganze Haus/  
mit Pfefferkuchen gemauert aus/die Träm mit

# hOCR - Darstellung im Browser

---

Un höret zu und schweiget still/ hab?  
icht was ich euch singen will/von eitafi  
guten Lande/ so blieb mancher daheime  
wann ihm das wäre blkandte.

**Beispiel**

i\* Der Weg der ist auch ziemlich weit/ sätti-  
ge Kinder und alte Lcut/ machen sich dadurch  
viel im Winter ist es ihnen zu kalt/  
«Vdauchzuhrißim Sommer.

z. Die Gegend heist Schlauraffenland/  
ist faulen Leuten wol bekandt/ red ich ohn all« n  
Schaden/ darinnen sind die Häuser gedeckt/  
Mit lauter Eyer-Flaben.

4. Welche Mägde oder deß Lan»  
des Att erfahren wollen/ mögen sich dahin ver-  
fügen / wann man die Dächer brichet ab / hat  
krFladen voll Genügen»

5. Thür und Wände das gantze Hauß/  
mit Pfefferkuchen gemauert aus/die Tram mit  
Schweinen-Braten/kaufft einer dort für ein  
Pfenning hier gilt einen Ducaten.

# hOCR: Daten

---

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
<title>OCR Output</title>
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
<meta http-equiv="content-style-type" content="text/css" />
<meta name="ocr-capabilities" content="ocr_page ocr_par ocrx_word ocr_line" />
<meta name="ocr-system" content="ABBYY fre-8.0.1.1024" />
<meta name="ocr-number-of-pages" content="1" />
</head>
<body bgcolor="#ffffff">
<div class="ocr_page" title="bbox 0 0 1709 1709;ppageno 20"> Beginn Page: ocr_page
<div class="ocrx_block" title="bboxnull 110 1186 1942" style="font-size:9pt;font-family:'Arial';font-style:normal"><br>
<p class="ocr_par" align="justify" left-indent="200" linespacing="1509" style="font-size:10pt;font-family: times new roman";font-style: normal"><br><span
class="ocr_line" baseline="169" title="bbox 342 117 1162 181"><span class="ocrx_word" title="bbox 342 117 415 170">Un</span> <span class="ocrx_word"
title="bbox 424 119 536 180">höret</span> <span class="ocrx_word" title="bbox 559 129 603 177">z</span> <span class="ocrx_word" title="bbox 621 120 702
168">und</span> <span class="ocrx_word" title="bbox 722 118 917 179">schweiget</span> <span class="ocrx_word" title="bbox 938 120 1043
181">still</span> <span class="ocrx_word" title="bbox 1063 123 1162 181">hab?</span></span> <br><span class="ocr_line" baseline="231" title="bbox 355
176 1163 242"><span class="ocrx_word" title="bbox 355 186 432 242">icht</span> <span class="ocrx_word" title="bbox 446 189 534 230">was</span> <span
class="ocrx_word" title="bbox 544 185 596 240">ich</span> <span class="ocrx_word" title="bbox 609 185 693 240">euch</span> <span class="ocrx_word"
title="bbox 707 182 841 242">singen</span> <span class="ocrx_word" title="bbox 854 182 1035 238">will/von</span> <span class="ocrx_word" title="bbox
1048 176 1163 234">eitafi</span></span> <br><span class="ocr_line" baseline="295" title="bbox 350 245 1164 308"><span class="ocrx_word" title="bbox 350
255 466 306">guten</span> <span class="ocrx_word" title="bbox 477 245 624 299">Lande/</span> <span class="ocrx_word" title="bbox 636 246 892
302">so</span> <span class="ocrx_word" title="bbox 691 246 786 294">blieb</span> <span class="ocrx_word" title="bbox 797 249 925 304">mancher</span>
<span class="ocrx_word" title="bbox 988 248 1164 308">daheim</span></span> <br><span class="ocr_line" baseline="356" title="bbox 170 308 958 368"><span
class="ocrx_word" title="bbox 307 322 428 357">wann</span> <span class="ocrx_word" title="bbox 443 311 525 365">ihm</span> <span class="ocrx_word"
title="bbox 540 311 615 355">das</span> <span class="ocrx_word" title="bbox 634 308 735 356">wäre</span> <span class="ocrx_word" title="bbox 757 309 958
357">blkandte.</span></span> <br></p>
```

# hOCR: Daten

---

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
<title>OCR Output</title>
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
<meta http-equiv="content-style-type" content="text/css" />
<meta name="ocr-capabilities" content="ocr_page ocr_par ocrx_word ocr_line" />
<meta name="ocr-system" content="ABBYY fre-8.0.1.1024" />
<meta name="ocr-number-of-pages" content="1" />
</head>
<body bgcolor="#ffffff">
<div class="ocr_page" title="bbox 0 0 1709 1709;ppageno 20">

<div class="ocrx_block" title="bboxnull 110 1186 1942" style="font-size:9pt;font-family:'Arial';font-style:normal"><br>
<p class="ocr_par" align=Justified leftIndent=200 lineSpacing=1309 style="font-size:10pt;font-family:'Times New Roman';font-style:n
class="ocr_line" baseline= 169 title="bbox 342 117 1162 181"><span class="ocrx_word" title="bbox 342 117 415 170">Un</span> <span c
title="bbox 424 119 536 180">höret</span> <span class="ocrx_word" title="bbox 536 119 603 177">z</span> <span class="ocrx_word" titl
168">und</span> <span class="ocrx_word" title="bbox 722 118 917 179">schweiget</span> <span class="ocrx_word" title="bbox 938 120 104
181">still</span> <span class="ocrx_word" title="bbox 1063 123 1162 181">nab?</span></span> <br><span class="ocr_line" baseline=
176 1163 242"><span class="ocrx_word" title="bbox 355 186 432 242">icht</span> <span class="ocrx_word" title="bbox 446 189 534 230">
class="ocrx_word" title="bbox 544 185 596 240">ich</span> <span class="ocrx_word" title="bbox 609 185 693 240">euch</span> <span clas
title="bbox 707 182 841 242">singen</span> <span class="ocrx_word" title="bbox 854 182 1035 238">will/von</span> <span class="ocrx_w
1048 176 1163 234">eitafi</span></span> <br><span class="ocr_line" baseline= 295 title="bbox 350 245 1164 308"><span class="ocrx_wor
255 466 306">guten</span> <span class="ocrx_word" title="bbox 477 245 624 299">Lande/</span> <span class="ocrx_word" title="bbox 636
302">so</span> <span class="ocrx_word" title="bbox 691 246 786 294">blieb</span> <span class="ocrx_word" title="bbox 797 249 975 304"
<span class="ocrx_word" title="bbox 988 248 1164 308">daheim</span></span> <br><span class="ocr_line" baseline= 356 title="bbox 170
class="ocrx_word" title="bbox 307 322 428 357">wann</span> <span class="ocrx_word" title="bbox 443 311 525 365">ihm</span> <span clas
title="bbox 540 311 615 355">das</span> <span class="ocrx_word" title="bbox 634 308 735 356">wäre</span> <span class="ocrx_word" titl
357">blkandte.</span></span> <br></p>
```

Bounding Box

Text

# PDF - Extraktion

---

- Als reines Darstellungsformat kennen einige PDFs nicht einmal einen Wortbegriff
- Es lassen sich / müssen Techniken wie beim OCR angewandt werden um ein solches Dokument korrekt Analysieren zu können
- Im Unterschied zur OCR sind die Positionen exakt
- [http://www.neumann.biz/data\\_munging/pdf](http://www.neumann.biz/data_munging/pdf)