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 direketen Analyse auf der Grafik
- Teil des Preprocessing bei 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 es 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 verschieden 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 exisitert 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

Page

Block 1

Block 2

Avant-bec *m.* *Voy.* Brise-glace.

Avant-chemin-couvert *m.* (chemin couvert fait au pied du glacis, et qui se trouve le plus avancé dans la campagne) (Fort.) *Der zweite gedeckte Weg.* Second *or* advanced covert-way.

Avant-coulant *m.*, **Première eau-de-vie** *f.* *Der Vorlauf.* First running of brandy, first short. *Voy.* Eau-de-vie.

Avant-creuset *m.* d'un haut-fourneau (espace creux qui se trouve avant le creuset) *Der Vorherd.* Breast-pan.

Avant-fossé *m.* (fossé fait au pied du glacis et qui précède immédiatement l'avant-chemin-couvert.) (Fort.) *Der Vorgraben, Aussengraben.* Advanced *or* second ditch, avant-fosse.

Avant-garde *f.* d'une armée navale (division qui, en ligne de bataille, se trouve en avant de deux autres) (Mar.) *Das Vordertreffen, die Avantgarde, die vordere Schlachtlinie.* Van of a fleet, van-guard, first-line, first-division.

Avant-glacis *m.* (glacis de l'avant-chemin-couvert) (Fort.) *Das Vor-Glacis.* Advanced glacis.

Avant-mur *m.*, **Barbacane** *f.* (Fort.) *Die Zwingermauer.* Barbacan, barbican. *Voy.* Barbacane.

Avant-port *m.* (entrée d'un grand-port au dehors de son enceinte) (Mar.) *Der Butenhafen, Aussenhafen, Vorhafen.* Outer-harbour.

Avant-titre *m.*, **Faux-titre** *m.* d'un livre (titre abrégé, à la fausse page) (Impr.) *Der Schmutztitel.* Bastard-title.

Avant-train *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.) *Der Vorderwagen.* Fore-carriage.

Avant-train *m.* d'un affût (train qui comprend les roues de devant et le timon d'un canon de campagne((Artill.) *Die Protze, der Vorderwagen, die Geschützprotze.* Limber, gun-limber.

Avant-train *m.* **à bras de limonière.** *Der Vorderwagen mit Gabeldeichsel.* Limber with shafts.

Avant-train *m.* **à timon.** *Der Vorderwagen*

Block 3

…rchandises dont il est chargé; dépenses nécessitées pour la conservation du navire (Mar.) *Die Haverie, der Seeschaden.* Average, damage by sea.

Avarie *f.* **grosse** ou **commune** (dépenses faites pour la préservation du navire et de la cargaison dont les frais doivent être partagés entre le propriétaire du vaisseau et celui des marchandises). *Die grosse, generale, gemeine Haverie.* General average, gross average.

Avarie *f.* **simple** ou **particulière** (dépenses extraordinaires faites, soit pour le bâtiment seul, soit pour la cargaison seule). *Die einfache, besondere Haverie.* Simple average, particular average.

Menue avarie *f.* (droit qu'on paie aux pilotes et aux mariniers qui font entrer le vaisseau dans le port, ou l'en font sortir). *Die kleine oder ordinaire Haverie.* Small *or* petty average.

Avec le soleil (sur la droite) (Mar.) *Mit der Sonne.* With the sun.

(Aventure) Grosse aventure *f.*, **Bomerie** *f.* (prêt d'argent à un fort intérêt, sur un navire ou sur sa cargaison) (Mar.) *Die Bodmerei, Bodmerie, das Geld auf Bodmerei.* Bottomry.

Aventurier *m.* (corsaire) (Mar.) *Der Freibeuter.* Free-booter. *Compar.* Flibustier.

Aventurier *m.*, **Interlope** *m.* (navire marchand qui trafique en fraude dans les pays de la concession d'une compagnie de commerce etc.) (Mar.) *Der Aventurier.* Adventurer, smuggler, free-booter.

Aventurine *f.*, **Quartz** *m.* **aventuriné** (variété de quartz présentant des parties brillantes disséminées dans un ciment plus obscur) (Minér.) *Der Avanturin, Aventurin.* Avanturine.

Aventurine *f.* **jaune à pluie d'or.** *Der Sonnenstein.* Sun-stone. *Voy.* Pierre de soleil.

Aviron *m.* (rame pour les nacelles) (Mar.) *Das*

Page

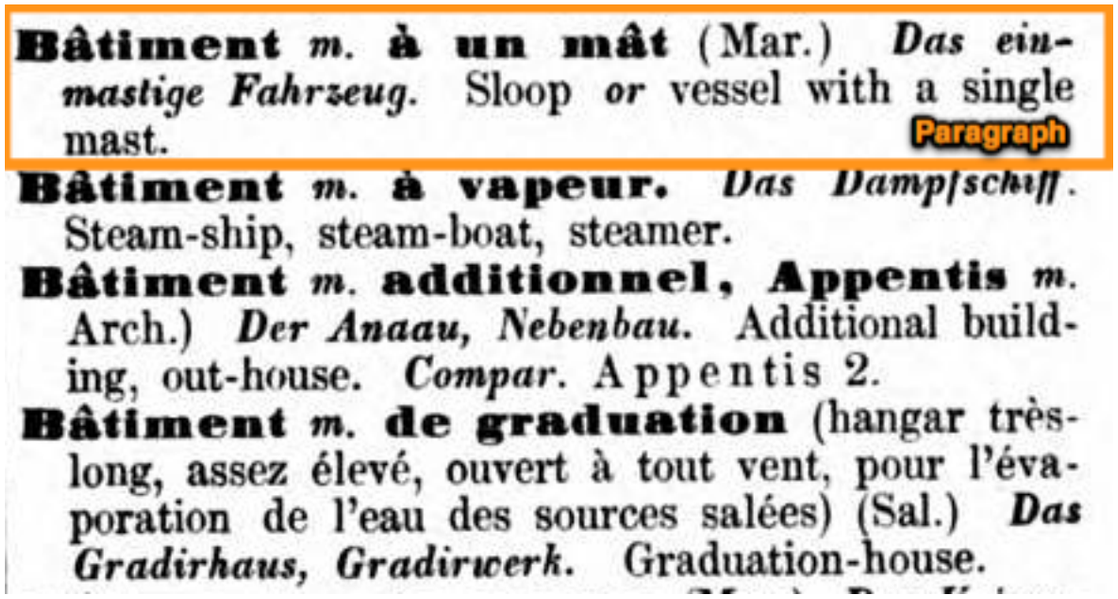
- Beschreibt die physische Buchseite
- Dienst als Bezugsrahmen für weiter Layoutelmente
- Seitenzahl optionales Element
- Umfasst Boxes, Paraggraphs, Lines und Words

Block

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

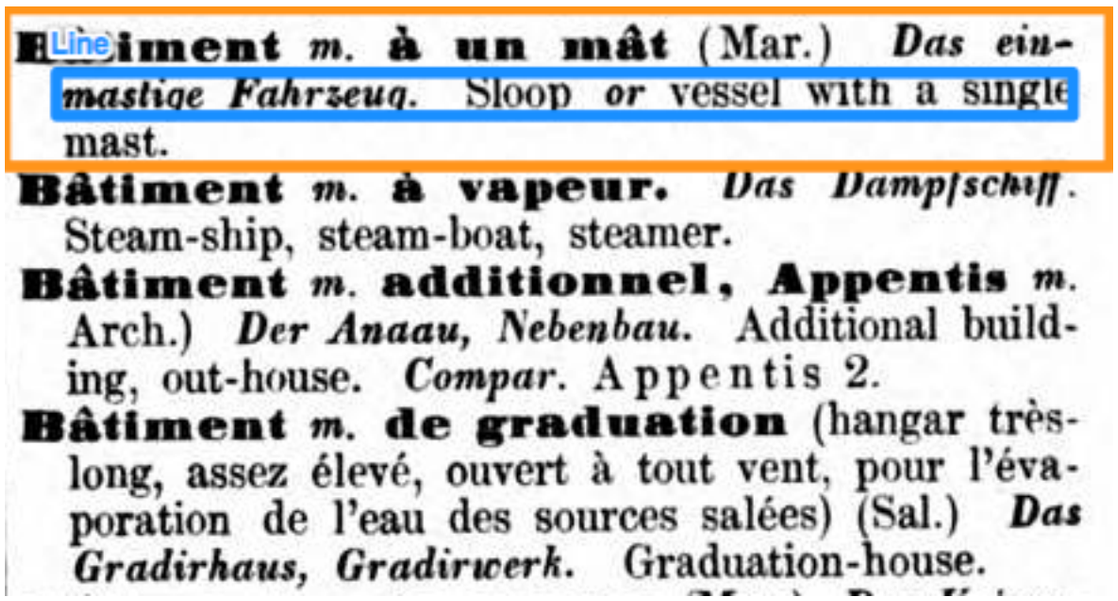
Paragraph

- Absatz
- Umfasst Lines und Words



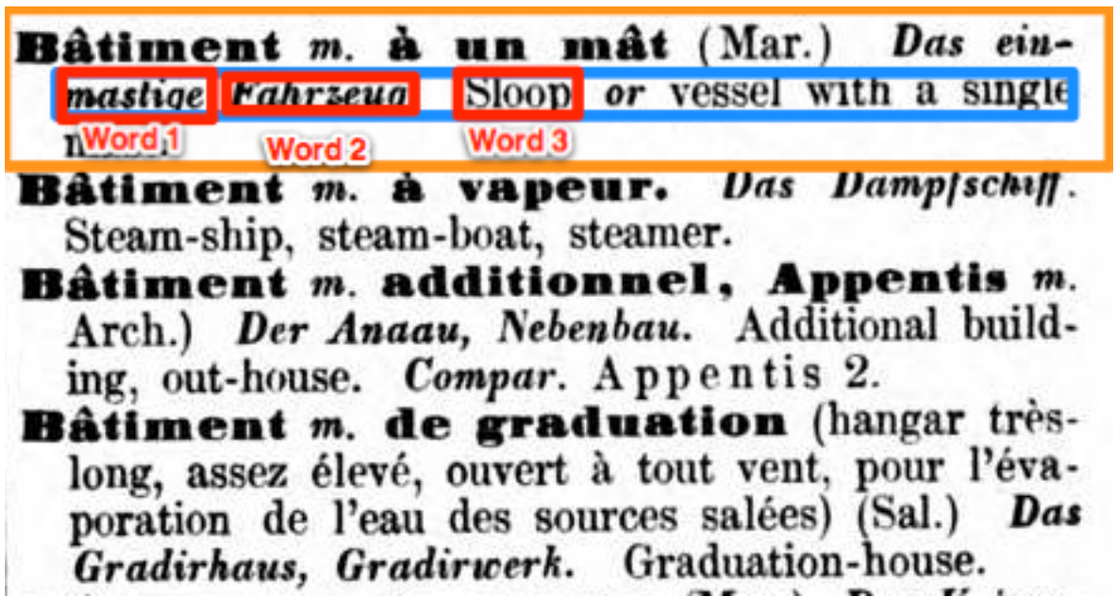
Line

- Zeile
- Umfasst Words



Word

- Wort
- Wortbegriff bei AbbyyXML nicht vorhanden



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

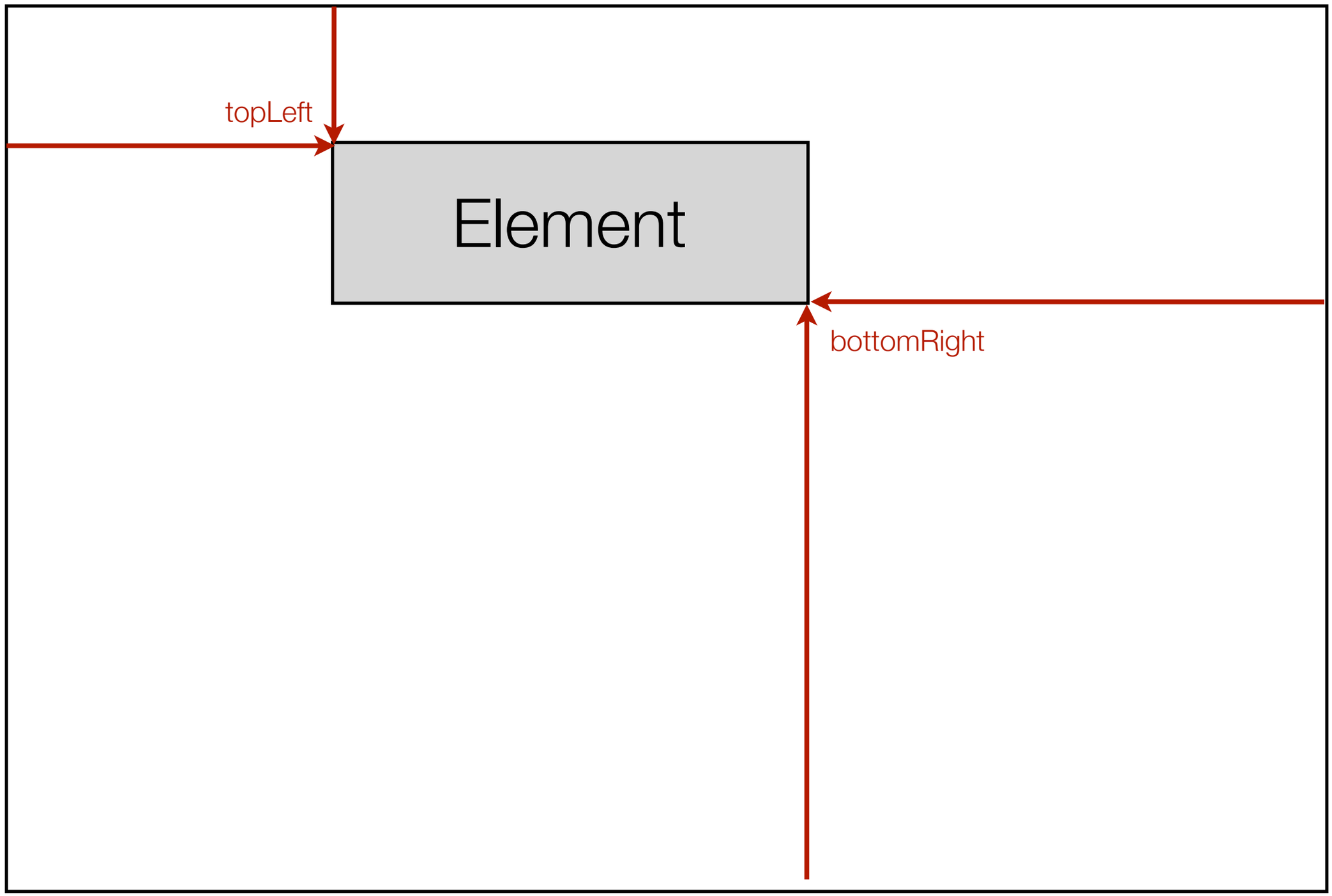
Character

- Zeichen
- Ein Word umfasst mehrere Chracters
- Nur bei AbbyyXML vorhanden

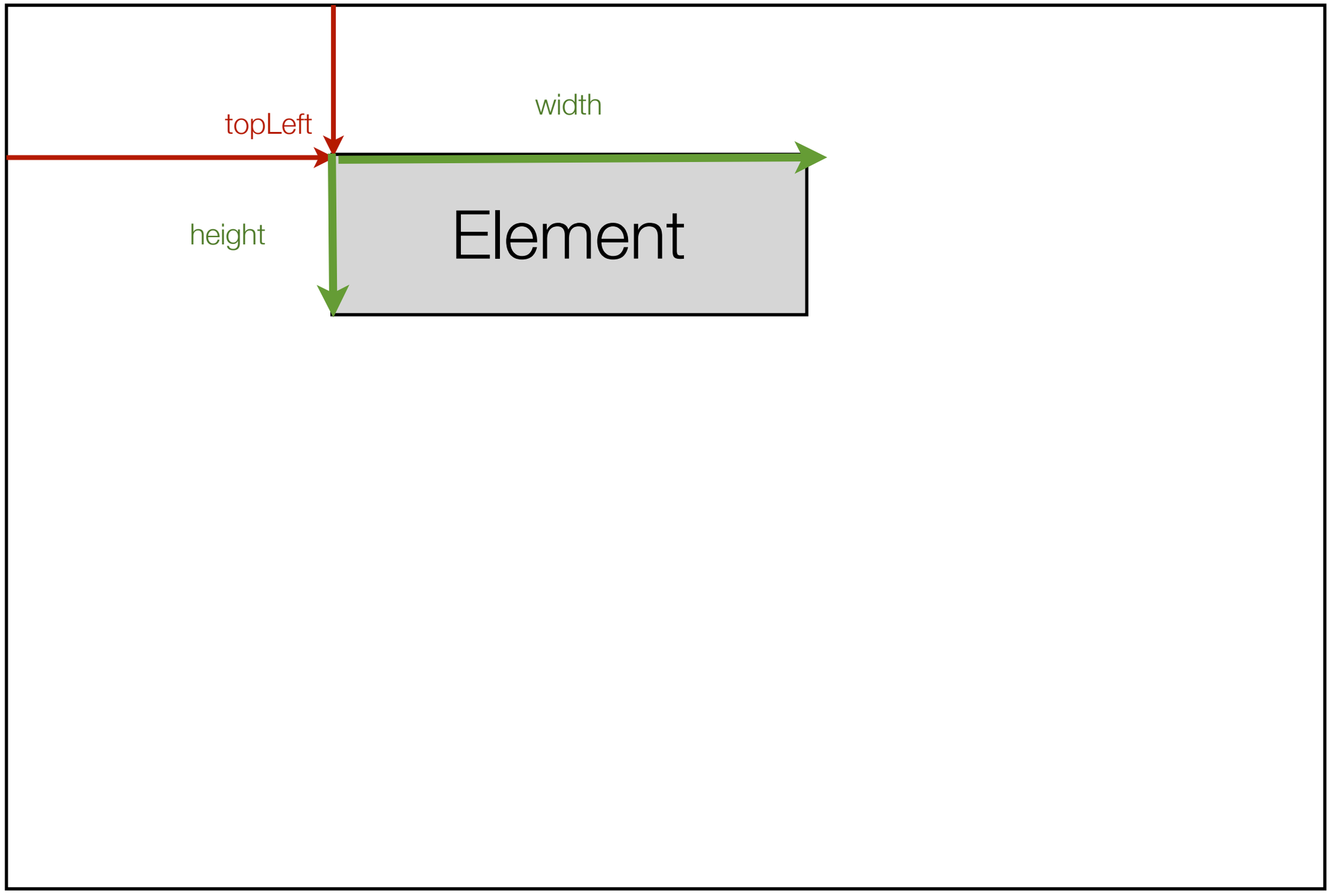
BoundingBox

- 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

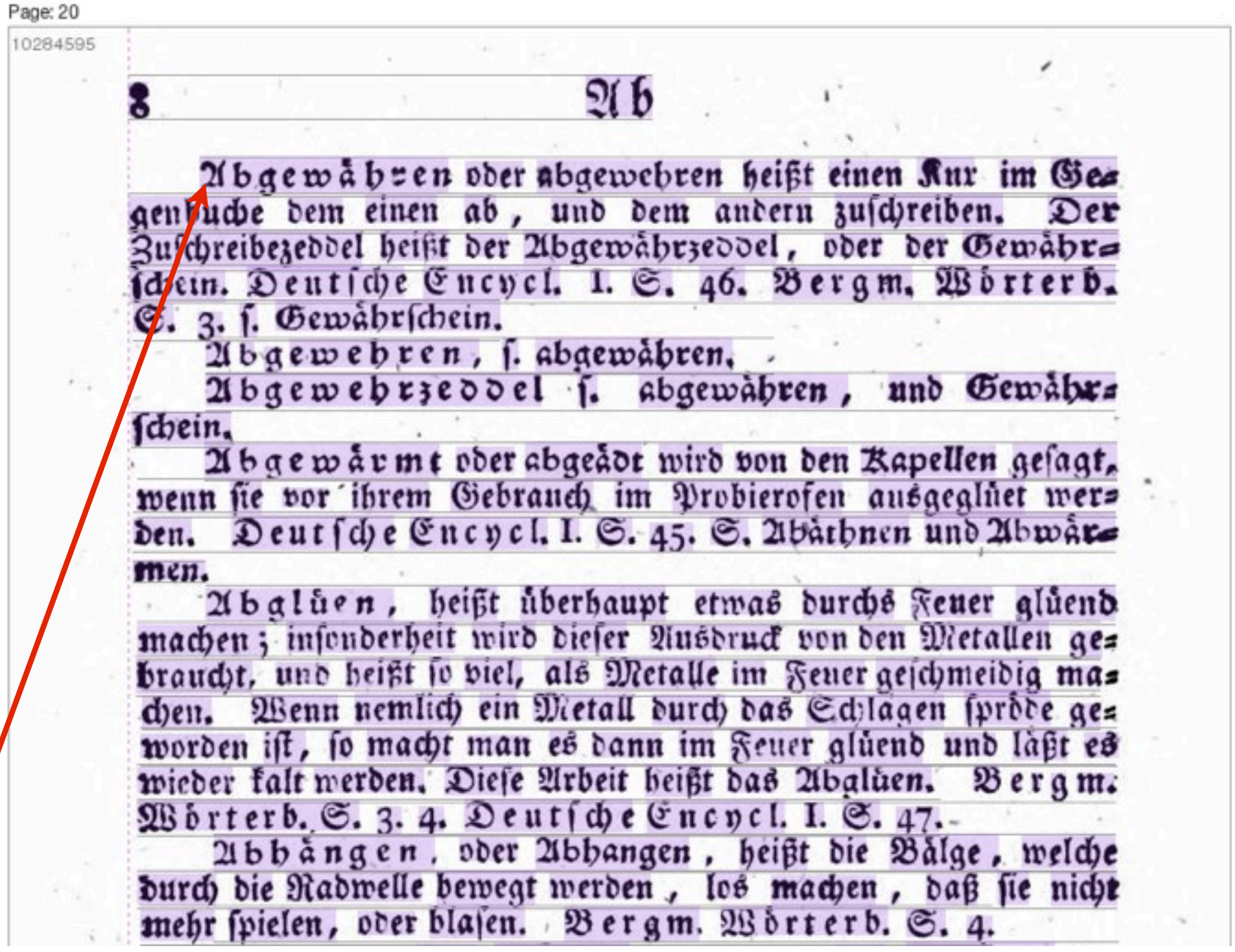


Bounding Box



Seite mit Bounding Boxes

Seite mit Bounding Boxes und Grafik




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

Mehr zu den Formaten: OCR

- Abbyy-XML: http://www.neumann.biz/data_munging/abbyy_xml
- Alto-XML: http://www.neumann.biz/data_munging/alto_xml
- hOCR: http://www.neumann.biz/data_munging/neumann.biz

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