**Input-Datenformat**

Aufbau der Page XML-Dateien:

Die XML-Dateien sind in zwei grundlegende Bereiche aufgeteilt. Zum einen *Metadata* und zum anderen *Page*.

1. Metadaten
   1. Ersteller (Tag: *Creator*)
   2. Erstellungsdatum (Tag: *Created*)
   3. Letzte Änderungen (Tag: *LastChanged*)
2. Page
   1. Seitenzahl (Tag: *Page*, Attribute: *imageFilename*)
   2. Seitenlänge (Tag: *Page*, Attribute: *imageHeight*)
   3. Seitenbreite (Tag: *Page*, Attribute: *imageWidth*)
   4. Textbereich (Tag: *TextRegion*)

Im Tag *TextRegion* steht der Textinhalt. Dieser ist folgend aufgeteilt:

1. Textregion
   1. Koordinaten (Tag: *Coords*)
   2. Textzeilen (Tag: *TextLine*)
   3. Text (als Ganzes) (Tag: *TextEquiv*)

Die Elemente *TextLine* enthalten die einzelnen Zeilen des Textes und sind folgendermaßen aufgebaut:

1. Textline
   1. Koordinaten (Tag: *Coords*)
   2. Text (Tag: *TextEquiv*, Attribut: *Unicode*)

**Koordinaten**

Die Koordinaten *Coords*, welche für *TextRegion* und *TextLine* angegeben sind, bestimmen die Position des Textes sowie der Textzeilen auf der Seite. Dabei sind die Koordinaten in folgendem Format angegeben: [x1,y1 x2,y2 x3,y3 x4,y4]. Der Text bzw. Textbereich wird als ein Rechteck modelliert und die vier (x,y) - Koordinaten definieren jeweils eine Ecke dieses Bereiches. Dasselbe gilt für jede Textzeile. In der Abbildung 1 wird das verdeutlicht.

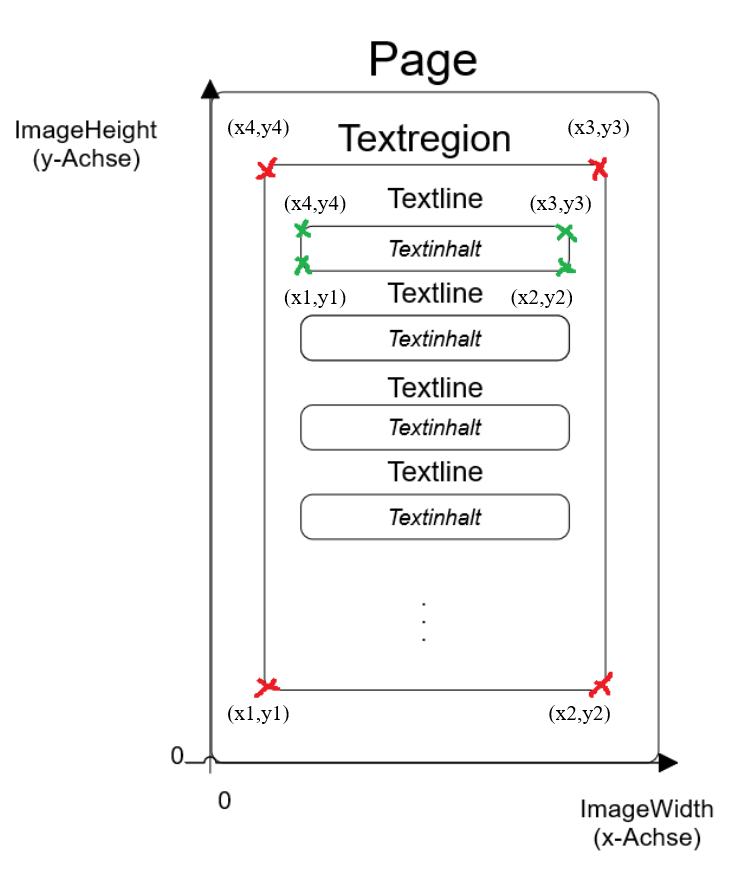
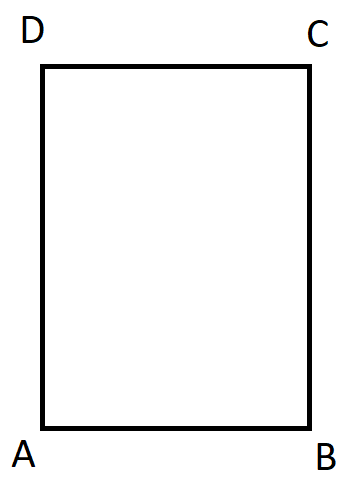


Abbildung 1 Modellierung einer Seite

Die farbigen, eingezeichneten Punkte zeigen die Koordinatenpositionen. Dabei kann die Page als Koordinatensystem betrachtet werden, wobei die ImageWidth (Seitenbreite) die x-Achse modelliert und ImageHeight (Seitenhöhe) die y-Achse, *TextRegion* und *TextLine* befinden sich somit innerhalb dieser Page bzw. innerhalb des Koordinatensystems. Die Koordinatenliste : [x1,y1 x2,y2 x3,y3 x4,y4] ist wie folgt aufgeteilt:

1. x1,y1 - Koordinate für untere, linke Ecke (A)
2. x2,y2 - Koordinate für untere, rechte Ecke (B)
3. x3,y3 - Koordinate für obere, rechte Ecke (C)
4. x4,y4 - Koordinate für obere, linke Ecke (D)

**Datenmodell**

Zur Modellierung der Daten, welche in den XML-Dateien enthalten sind, sollen *pandas Dataframes* verwendet werden. Dabei soll es ein *pages\_df* und eine *textlines\_df* geben.

*pages\_df*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| pageNum | height | width | textRegID | A | B | C | D | text | creator | created | lastChange |

*textlines\_df*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| id | textline | A | B | C | D | pageNum |