DDB-ALTO

> ⚠ **Work in Progress**
>
> Diese Seite befindet sich noch in der Bearbeitung. Ggf. wird es in den nächsten Wochen noch zu Änderungen kommen.

Das Akkronym ALTO steht für **A**nalyzed **L**ayout and **T**ext **O**bject. ALTO ist ein standardisiertes XML-Format zur Speicherung von Layout- und Inhaltsinformationen, das im Rahmen des EU-Projekts Metadata Engine entwickelt wurde. Seit 2009 wird der ALTO-Standard von der Library of Congress gepflegt.

ALTO wird meist in Kombination mit dem METS XML Schema (Metadata Encoding and Transmission Standard) als Erweiterung der administrative metadata section verwendet. Während METS Metadaten und strukturelle Informationen liefert, enthält ALTO inhaltliche und physische Informationen des Digitalisats. In der METS-Datei gibt es Dateizeiger auf die ALTO-Dateien. Sie befinden sich im METS Element <fileGrp> unter <fileSec>.

ALTO fußt auf einem seitenbasierten Verfahren, d.h. eine ALTO-Datei entspricht einer digitalisierten Seite. ALTO speichert Layout-Informationen und via OCR (Optical Character Recognition) erkannten Text von Seiten jeglicher Art von gedruckten Dokumenten wie Büchern, Zeitschriften und Zeitungen. Eine Seite wird in mehrere Bereiche unterteilt ("Print space", "left margin", "right margin", "top margin" und "bottom margin"). Für jeden Bereich werden alle Objekte aufgelistet, die darin erkannt wurden (z.B. Zeilen, Bilder, Textblöcke etc.). Alle erkannten Objekte werden mittels Koordinaten (Informationen zur Breite, Höhe, horizontaler und vertikaler Position und ggf. Angaben zur Rotation) erfasst. Der Viewer, den die DDB nutzt, kann Volltexte eines Digitalisats anzeigen, sofern die Volltexte im ALTO-Format mit Wortkoordinaten vorliegen.

Die offizielle Dokumentation des ALTO-Standards findet sich hier sowie hier auf Github. Version 4.4 (release 2023-04-07) ist die aktuelle Schema-Version.

Struktur und Aufbau von ALTO-Dateien

Für die Beschreibung eines Dokuments sind die folgenden ALTO-Elemente verpflichtend:

- <alto>: das Wurzelelement des ALTO-Datensatzes,
- <Layout>: verpflichtendes Top Level Element unter dem <alto> Wurzelelement, das die die eigentlichen Inhaltsinformationen – also was auf der digitalisierten Seite steht – enthält,
- <Page>: verpflichtendes Kindelement unter <Layout>. Alle <Page>-Elemente müssen einen gültigen ID Attributwert, einen PHYSICAL_IMG_NR Attributwert sowie die positionsbeschreibenden Attribute WIDTH (Breite) und HEIGHT (Höhe) beinhalten,
- <PrintSpace>: Kindelement unter <Page> mit den Attributen WIDTH (Breite), HPOS (horizontale Position), VPOS (vertikale Position), ID und HEIGHT (Höhe),
- <TextBlock>: Kindelement unter <PrintSpace> mit den Attributen WIDTH (Breite), HPOS (horizontale Position), VPOS (vertikale Position), ID und HEIGHT (Höhe). Textblöcke werden in Textzeilen unterteilt.
- <TextLine>: Kindelemente endlicher Anzahl unter <TextBlock>, die eine Zeile Text speichern und über die Attribute HPOS (horizontale Position), VPOS (vertikale Position), ID, WIDTH (Breite) und HEIGHT (Höhe) verfügen,
- <String>: Kindelemente endlicher Anzahl unter <TextLine>, die eine Zeichenkette/ein Wort speichern und über die Attribute HEIGHT (Höhe), CONTENT (Inhalt = das Wort), WIDTH (Breite), HPOS (horizontale Position), VPOS (vertikale Position) und ID verfügen. Da das ALTO-Schema <String> erfordert, ist bei der OCR eine Wortsegmentierung erforderlich.
- <SP>: Kindelemente endlicher Anzahl unter <TextLine>, die Leerzeichen zwischen <String> speichern. Im ALTO XML-Schema ist die Verwendung des SP-Tags als optional ausgewiesen. **In der DDB ist die Nutzung des SP-Tags zur Auszeichnung von Leerzeichen hingegen verpflichtend**.

Weitere mögliche Top Level Elemente (zusätzlich zum <Layout> Element) sind die Elemente <Description> und <Styles>. In der DDB sind Angaben in den Unterlementen <MeasurementUnit> (pixel) und <sourceImageInformation> erforderlich. <Styles> enthält Informationen zu Text- und Absatzstilen.

Einfaches Beispiel eines ALTO-Datensatz

Einfaches Beispiel eines ALTO-Datensatzes

```
<alto>
        <Description>
                <MeasurementUnit>pixel</MeasurementUnit>
                <sourceImageInformation>
                        <fileIdentifier>5177528</fileIdentifier>
                </sourceImageInformation>
        </Description>
        <Layout>
                <Page WIDTH="3174" PHYSICAL_IMG_NR="1" ID="p5177528" HEIGHT="4065">
                        <PrintSpace WIDTH="3174" HPOS="0" VPOS="0" ID="ps5177528" HEIGHT="4065">
                                <TextBlock WIDTH="3174" HPOS="0" VPOS="0" ID="tb5177528" HEIGHT="4065">
                                        <TextLine HPOS="1835" VPOS="325" ID="tl0" WIDTH="665" HEIGHT="57">
                                                <String HEIGHT="60" CONTENT="Bielefeld" WIDTH="184" HPOS="1835"
VPOS="325" ID="st0"/>
                                                <String HEIGHT="60" CONTENT="," WIDTH="21" HPOS="2019" VPOS="
325" ID="st1"/>
                                                <SP/>
                                                <String HEIGHT="60" CONTENT="den" WIDTH="58" HPOS="2062" VPOS="
322" ID="st2"/>
                                                <SP/>
                                                <String HEIGHT="60" CONTENT="4" WIDTH="21" HPOS="2157" VPOS="
320" ID="st3"/>
                                                <String HEIGHT="60" CONTENT="." WIDTH="21" HPOS="2178" VPOS="
320" ID="st4"/>
                                                <SP/>
                                                <String HEIGHT="60" CONTENT="Oktober" WIDTH="169" HPOS="2215"
VPOS="320" ID="st5"/>
                                                <SP/>
                                                <String HEIGHT="60" CONTENT="1924" WIDTH="90" HPOS="2410" VPOS="
322" ID="st6"/>
                                        </TextLine>
                                </TextBlock>
                        </PrintSpace>
                </Page>
        </Layout>
</alto>
```

Dieser beispielhafte ALTO-Datensatz speichert die Wortkoordinaten des Textblocks *Bielefeld, den 4. Oktober 1924*.