# Beskrivelse af et webværktøj til redigering af ALTO-tekster med parallel visning af faksimile

## Behov

Det Kgl. Bibliotek (KB) har i 2018-2019 digitaliseret (scannet) og OCR-behandlet 28.650 tekstsider skrevet med fraktur (”gotiske” bogstaver) primært fra årene 1770-1773. Det er en samling af pamfletter hvor både papirkvaliteten og trykkekvaliteten er meget svingende. Resultaterne skal både anvendes til forskning og til visning på websider. OCR-resultaterne foreligger som XML-filer i standard ALTO-format.

På verdensplan er det typisk (pr. 2018-2019) at OCR-resultater af fraktur-tekster fra før år 1800 ikke er gode, men det er lykkedes KB at identificere størstedelen af ordene korrekt. Men til forskningsbrug er OCR-kvaliteten ikke tilstrækkelig.

Derfor skal ALTO-filerne efterredigeres manuelt. Arbejdet er estimeret til 9 personmåneder uden særlige hjælpeværktøjer ud over et XML-redigeringsprogram. Denne arbejdstid ønskes reduceret.

Det er vigtigt at redigeringsarbejdet kan foregå hurtigst muligt, og det er vigtigt at der ikke indskrives/slettes tegn ud over i tekstområderne. Desuden er det vigtigt ikke at miste informationer om ordenes positioner i forhold til originalen (faksimilen).

Derfor er det behov for en enkelt værktøj til visning og redigering af ALTO-teksterne med parallel visning af tilhørende faksimiler. Værktøjet skal fungere i en webbrowser.

## Forudsætninger

### ALTO-filer

En ALTO-fil i nærværende projekt indeholder OCR-tekst for kun én faksimile. Dvs. at der er én ALTO-fil for hver faksimile. ALTO-filerne anvender tegnsættet UTF-8.

Værktøjet skal kun kunne håndtere de XML-elementer og -attributter som findes i projektets ALTO-filer. Det er således ikke nødvendigt at supportere alle elementer og attributter som ALTO-standarden beskriver.

Alle ALTO-filer bliver placeret i GitHub grupperet i underkataloger i ét niveau. Således vil enhver ALTO-fil eksempelvis kunne tilgås vha. GitHub’s REST-API. Dokumentation:

<https://developer.github.com/v3/>

Under softwareudviklingen vil der blive givet adgang til et repræsentativt udvalg af ALTO-filer.

### Billeder = faksimiler

Der findes én faksimile (affotograferet tekstside) til hver ALTO-fil. Filnavnene er identiske på nær fil-extension.

Alle faksimiler findes i god kvalitet og med høj opløsning. Billederne vil blive gjort tilgængelige på KB’s billedserver der kan tilgås som URL’er med IIP 2.0-prokokollen. Hermed leveres et relevant uddrag af billedet ved at kopiere position, bredde og højde (bounding box) fra et tekststykke i ALTO-filen.

**Eksempel, tekstpositioner fra en linje i en ALTO-fil:**  
(filnavn ”001991704\_000\_0107.xml”)

<TextLine HPOS="**186**" VPOS="**1815**" WIDTH="**1452**" HEIGHT="**64**" …>

Værdierne med fed skrift kan overføres direkte til en URL. Den primære del af URL’en kan også udtrækkes af ALTO-filen. Klik for at se et eksempel på et uddrag (faksimilen ligner det aktuelle projekt):

[https://kb-images.kb.dk/public/dk\_breve/001991704\_000/001991704\_000\_0107/ **186,1815,1452,64**/!1000,300/0/default.jpg](https://kb-images.kb.dk/public/dk_breve/001991704_000/001991704_000_0107/%20186,1815,1452,64/!1000,300/0/default.jpg)

Dokumentation til IIP 2.0-protokollen:

<https://iiif.io/api/image/2.0/>

Se evt. billedet i fuld størrelse:

<https://kb-images.kb.dk/public/dk_breve/001991704_000/001991704_000_0107/>

Under softwareudviklingen vil der blive givet adgang til alle faksimiler der svarer til ALTO-filerne nævnt ovenfor.

## Funktionelle krav til værktøjet

Der skal laves en enkel brugergrænseflade baseret på HTML/CSS og JavaScript (helst jQuery).

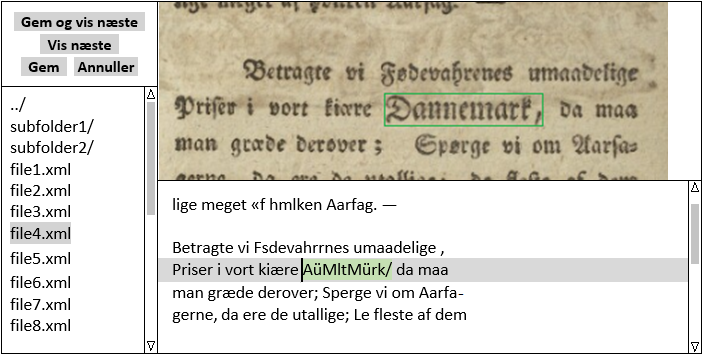
Værktøjet skal kun tilpasses det aktuelle redigeringsprojekt, og skal således ikke udstyres med egenskaber som der ikke er brug for i dette projekt.

Brugergrænsefladen behøver ikke at være responsiv. Den skal anvendes i desktop-browsere og ikke på smartphones eller tablets.

Der er ikke ønske om en grafisk veldesignet applikation. I stedet ønskes fokus på de vigtige funktionelle elementer, og at applikationen hurtigt er klar til anvendelse.

### Brugergrænseflade

Brugergrænsefladen skal i hovedtræk have nogenlunde dette udseende i en browser:



#### Forklaring til ovenstående billede

Knapper og filvælger til venstre, ALTO-tekst (OCR-tekst) nederst til højre, og et udsnit af faksimilen øverst til højre med grøn markering som matcher det med museklik markerede ord i ALTO-teksten.

Det er hensigtsmæssigt at feltet med ALTO-teksten øges i højden når browservinduet øges i højden.

Tekst-cursoren fremkommer når der klikkes på et ord. Herved skiftes fra visningsmodus til redigeringsmodus for det aktuelle ord. Det bemærkes at hele tekstlinjen og det aktuelle ord skal farvemarkeres.

Det aktuelle ord skal også markeres med en farvet ramme på udsnittet af faksimilen. KB’s billedserver har ikke faciliteter til dette så det skal programmeres i applikationen (med HTML5 <canvas>).

Det ønskes at det viste udsnit af faksimilen skal være noget større end det aktuelle ord. I højden bør der medtages hvad der svarer til minimum én tekstlinje før og én tekstlinje efter linjen som indeholder det aktuelle ord. Der bør også medtages noget af billedet til venstre for og til højre for det aktuelle ord – gerne hele den aktuelle linje hvis det er muligt. Dette muliggør at brugeren kan se det aktuelle ord i kontekst hvilket kan være nødvendigt for at forstå teksten.

Det er vigtigt at det som er grupperet i ét <TextLine> element i ALTO-filen, holdes sammen så brugeren kan forstå at det er én linje i faksimilen.

Det er vigtigt at det er hurtigt at navigere i teksten. Fx kunne det være en fordel at den aktuelle linje holdes i nogen afstand fra øverste og underste kant af tekstpanelet når der scrolles op/ned i ALTO-teksten.

Det er hensigtsmæssigt, men ikke et krav at udsnittet af faksimilen opdateres når der scrolles i ALTO-teksten. Men det er et krav at udsnittet af faksimilen opdateres når der klikkes på et ord (bortset fra gentagne klik på samme ord).

KB er åben over for andre forslag til udseende og funktionalitet så længe det er mindst lige så enkelt og hurtigt at anvende for brugeren som ovenstående og programmeringsopgaven ikke øges væsentligt.

### Oversigt over redigeringsfunktioner som skal implementeres

1. Rettelse af et ord
2. Sammenlægning af to på hinanden følgende ord (adskilte tekststrenge)
3. Sletning af et ord
4. Sletning af en linje

Det bemærkes at der kræves særlig håndtering når der er orddeling ved linjeskift.

### A. Rettelse af et ord

Brugen klikker på et ord i ALTO-feltet i brugergrænsefladen. KB tænker at der herved skiftes fra statisk visning af ordet til visning af et HTML INPUT-felt hvormed ordet kan redigeres (dette kræver DOM-manipulering med JavaScript).

Efter redigering af ordet klikkes på et andet ord etc. indtil redigeringen er færdig. Derefter klikkes på ’Gem’ eller ’Gem og vis næste’.

### Orddeling ved linjeskift

På det grafiske oplæg til brugergrænsefladen ovenfor mangler en visning af hvorledes orddeling ved linjeskift håndteres.

Det ønskes at ved klik på en af de to samhørende orddele - i slutningen eller i starten af en linje – skal kun den del som man klikker på være i redigeringsmodus. Derimod skal begge orddele være farvelagt i det nederste tekstpanel, og tilsvarende skal der på faksimilen være en farvet ramme om begge orddele. Dette gør det nemlig muligt for brugeren at se at orddelingen er registreret i ALTO-filen.

Når den første orddel i slutningen af en linje ikke er i redigeringsmodus, ønskes den vist med en orddelingsstreg. Se afsnit *Tip vedr. visning af orddelingsstreg* nedenfor.

Se følgende afsnit nedenfor:

* *Vedr. SUBS\_CONTENT (orddeling ved linjeskift)*
* *Vedr. OCR-fejl hvor to ord på faksimilen er placeret i én <String …>*
* *Vedr. OCR-fejl hvor to <String …> i ALTO-filen svarer til kun ét ord på faksimilen*

### B. Sammenlægning af to på hinanden følgende ord (adskilte tekststrenge)

Se nedenstående afsnit ”Vedr. OCR-fejl hvor to <String…> i ALTO-filen svarer til kun ét ord på faksimilen”.

### C. Sletning af et ord

Sletning af et ord foretages jf. afsnit ”A. Rettelse af et ord” ved at brugeren sletter alle tegn i tekstfeltet. Når ALTO-filen gemmes, skal det tilhørende <String> objekt slettes (se ALTO-detaljer nedenfor). Ved linjeskift skal eventuelle orddelingsinformationer opdateres.

### D. Sletning af en linje

Det er typisk at grafiske ornamenter i faksimilen kan give en ALTO-linje som udelukkende består af volapyk-ord. En sådan linje skal kunne slettes ved klik på en knap e.l. hvorved hele <TextLine> objektet fjernes (se ALTO-detaljer nedenfor). For at undgå fejl skal denne sletning evt. bekræftes med en Ja/Nej-dialog (afhængigt af den aktuelle løsning).

NB: KB har ikke taget stilling til hvordan og hvornår sletteknappen skal vises.

### Åbning/lagring af ALTO-filer

En mulighed er at ALTO-filerne læses fra/skrives til et lokalt katalog på PC’en/Mac’en. I tillæg til dette kan brugeren anvende et program (fx [GitHub Desktop](https://desktop.github.com/)) til at synkronisere det lokale katalog med indholdet på GitHub.

En anden mulighed er at ALTO-filerne læses og skrives direkte fra/til GitHub (med commit + push) vha. [GitHub’s REST-API](https://developer.github.com/v3/).

### ALTO-detaljer

Det er vigtigt at forstå at der til hvert enkelt ord i ALTO-filen (tekststreng med **grønt** herunder) er tilknyttet vigtige informationer som forsat skal være tilknyttet samme ord (tekststreng) når redigering er gennemført og der lagres.

Eksempel på to på hinanden følgende tekstlinjer i en ALTO-fil i projektet:

**<TextLine ID="LINE6" HEIGHT="100" WIDTH="1664" HPOS="185" VPOS="1019">**

<String ID="STRING30" STYLEREFS="STYLE2" HEIGHT="96" WIDTH="154" HPOS="185" VPOS="1023" CONTENT="**efter**" WC="0.63" LANG="dan"/>

<SP ID="SP25" WIDTH="30" HPOS="339" VPOS="1023"/>

<String ID="STRING31" STYLEREFS="STYLE2" HEIGHT="62" WIDTH="76" HPOS="369" VPOS="1041" CONTENT="**ar**" WC="0.31" LANG="dan"/>

<SP ID="SP26" WIDTH="44" HPOS="445" VPOS="1041"/>

<String ID="STRING32" STYLEREFS="STYLE2" HEIGHT="70" WIDTH="122" HPOS="489" VPOS="1029" CONTENT="**den**" WC="0.26" LANG="dan"/>

<SP ID="SP27" WIDTH="42" HPOS="611" VPOS="1029"/>

<String ID="STRING33" STYLEREFS="STYLE2" HEIGHT="94" WIDTH="204" HPOS="653" VPOS="1021" CONTENT="**haver**" WC="0.65" LANG="dan"/>

<SP ID="SP28" WIDTH="36" HPOS="857" VPOS="1021"/>

<String ID="STRING34" STYLEREFS="STYLE2" HEIGHT="92" WIDTH="202" HPOS="893" VPOS="1021" CONTENT="**kosted**" WC="0.66" LANG="dan"/>

<SP ID="SP29" WIDTH="38" HPOS="1095" VPOS="1021"/>

<String ID="STRING35" STYLEREFS="STYLE2" HEIGHT="72" WIDTH="238" HPOS="1133" VPOS="1043" CONTENT="**mange**" WC="0.36" LANG="dan"/>

<SP ID="SP30" WIDTH="52" HPOS="1371" VPOS="1043"/>

<String ID="STRING36" STYLEREFS="STYLE2" HEIGHT="94" WIDTH="248" HPOS="1423" VPOS="1025" CONTENT="**tusinde**" WC="0.56" LANG="dan"/>

<SP ID="SP31" WIDTH="38" HPOS="1671" VPOS="1025"/>

<String ID="STRING37" STYLEREFS="STYLE2" HEIGHT="98" WIDTH="140" HPOS="1709" VPOS="1019" CONTENT="**Rix**" SUBS\_TYPE="HypPart1" SUBS\_CONTENT="Rixdaler," WC="0.43" LANG="dan"/>

<HYP CONTENT="-"/>

**</TextLine>**

**<TextLine ID="LINE7" HEIGHT="102" WIDTH="1695" HPOS="185" VPOS="1117">**

<String ID="STRING38" STYLEREFS="STYLE2" HEIGHT="84" WIDTH="220" HPOS="185" VPOS="1131" CONTENT="**daler,**" SUBS\_TYPE="HypPart2" SUBS\_CONTENT="Rixdaler," WC="0.42" LANG="dan"/>

<SP ID="SP32" WIDTH="62" HPOS="405" VPOS="1131"/>

<String ID="STRING39" STYLEREFS="STYLE2" HEIGHT="74" WIDTH="82" HPOS="467" VPOS="1141" CONTENT="**og**" WC="0.35" LANG="dan"/>

<SP ID="SP33" WIDTH="48" HPOS="549" VPOS="1141"/>

<String ID="STRING40" STYLEREFS="STYLE2" HEIGHT="74" WIDTH="52" HPOS="597" VPOS="1125" CONTENT="**ei**" WC="0.41" LANG="dan"/>

<SP ID="SP34" WIDTH="46" HPOS="649" VPOS="1125"/>

<String ID="STRING41" STYLEREFS="STYLE2" HEIGHT="90" WIDTH="202" HPOS="695" VPOS="1123" CONTENT="**bragt**" WC="0.41" LANG="dan"/>

<SP ID="SP35" WIDTH="50" HPOS="897" VPOS="1123"/>

<String ID="STRING42" STYLEREFS="STYLE2" HEIGHT="76" WIDTH="190" HPOS="947" VPOS="1139" CONTENT="**noget**" WC="0.38" LANG="dan"/>

<SP ID="SP36" WIDTH="40" HPOS="1137" VPOS="1139"/>

<String ID="STRING43" STYLEREFS="STYLE2" HEIGHT="72" WIDTH="112" HPOS="1177" VPOS="1127" CONTENT="**ind**" WC="0.36" LANG="dan"/>

<SP ID="SP37" WIDTH="25" HPOS="1289" VPOS="1127"/>

<String ID="STRING44" STYLEREFS="STYLE1" HEIGHT="94" WIDTH="129" HPOS="1314" VPOS="1117" CONTENT="**(ft).**" WC="0.95" LANG="dan"/>

<SP ID="SP38" WIDTH="94" HPOS="1443" VPOS="1117"/>

<String ID="STRING45" STYLEREFS="STYLE2" HEIGHT="98" WIDTH="172" HPOS="1537" VPOS="1121" CONTENT="**Paa**" WC="0.5" LANG="dan"/>

<SP ID="SP39" WIDTH="34" HPOS="1709" VPOS="1121"/>

<String ID="STRING46" STYLEREFS="STYLE1" HEIGHT="93" WIDTH="137" HPOS="1743" VPOS="1125" CONTENT="**fags**" WC="0.98" LANG="dan"/>

**</TextLine>**

Hver tekstlinje er grupperet mellem <TextLine…> og </TextLine>. Alle TextLine-attributter skal bevares. *Bemærk, at to på hinanden følgende TextLine-elementer evt. er indeholdt i to forskellige TextBlock-elementer som også skal bevares.*

Identificerede ordmellemrum er indeholdt i elementerne <SP…/>. Alle attributter skal bevares.

Hvert identificeret ord er indeholdt i et <String…/> element. Alle attributter skal bevares. Dog skal nogle kunne redigeres.

Det findes særlige attributter når der er registreret orddeling ved linjeskift. De er markeret med grå baggrund ovenfor.

Flg. tabel beskriver alle attributter der vil forekomme i <String…/>:

|  |  |  |
| --- | --- | --- |
| Attributnavn | Skal kunne redigeres | Betydning |
| ID | Nej | Identifikation af XML-elementet |
| STYLEREFS | Nej | Reference til andet sted i ALTO-filen der beskriver fontstørrelse og -type |
| HEIGHT | Kun ved ordsammenlægning | Ordets maksimale højde på faksimilen (pixel) |
| WIDTH | Kun ved ordsammenlægning | Ordets maksimale bredde på faksimilen (pixel) |
| HPOS | Kun ved ordsammenlægning | Ordets vandrette position på faksimilen målt fra venstre kant (pixel) |
| VPOS | Kun ved ordsammenlægning | Ordets lodrette position på faksimilen målt fra overkant (pixel) |
| CONTENT | Ja | OCR-tekst. Dette er teksten som vises i brugergrænsefladen, og som brugeren evt. redigerer |
| WC | Nej | Word Confidence |
| LANG | Nej | Identificeret sprog |
| SUBS\_TYPE | Nej | Forekommer kun når der er registreret orddeling ved linjeskift.  Værdien er altid ” HypPart1” eller ”HypPart2” som angiver om der er tale om første eller anden del af det delte ord.  HypPart1 i slutningen af en linje hænger sammen med HypPart2 i begyndelsen af næste linje. |
| SUBS\_CONTENT | Ja | Forekommer kun når der er registreret orddeling ved linjeskift. Værdien (tekststrengen) er det ikke-delte ord. |

### Vedr. SUBS\_CONTENT (orddeling ved linjeskift)

Hvis brugeren redigerer CONTENT for et ord hvor attributten SUBS\_CONTENT er til stede, skal SUBS\_CONTENT opdateres for den aktuelle streng (HypPart1 eller HypPart2) **samt** for den tilhørende streng (hhv. HypPart2 eller HypPart1).

Et delt ord skal forblive delt efter redigeringen, og hver del skal forblive på hver deres linje.

SUBST\_CONTENT og informationerne om HypPart1 og HypPart2 skal også opdateres når et delt ord i begyndelsen/slutningen af en linje slettes (som enkeltord eller ved at slette hele linjen).

### Vedr. OCR-fejl hvor to ord på faksimilen er placeret i én <String…>

Det er okay. Dette behøver brugeren ikke ændre.

Værktøjet skal ikke håndtere dette.

### Vedr. OCR-fejl hvor to <String…> i ALTO-filen svarer til kun ét ord på faksimilen

Eksempel:

<String … CONTENT=”varesa” … />

<SP …/>

<String … CONTENT=”lg” … />

Her har OCR-softwaren delt ordet ”varesalg” i to strenge typisk fordi der er et lidt stort mellemrum mellem to bogstaver på faksimilen.

Det er et væsentligt problem og skal kunne rettes af brugeren. Ellers kan der ikke laves korrekte søgninger baseret på ALTO-filen.

NB: Fordi det er et væsentligt problem er det vigtigt at brugeren har let ved at finde/se sådanne tilfælde. Derfor skal softwareudvikleren sørge for ALTO-teksten i brugergrænsefladen vises med tydelige ordmellemrum, dvs. ekstra brede ordmellemrum (fx ved at vise <SP…/> elementer med en særlig bredde e.a.).

Det ønskes at brugergrænsefladen viser en knap e.l. hvormed **to på hinanden følgende** strenge i **samme** linje kan lægges sammen. KB har ikke taget stilling til hvordan og hvornår knappen skal vises.

Ved sammenlægning samles de to tidligere <String…/> elementer på følgende måde:

* Et <SP…/> element eller andre elementer mellem de to strenge fjernes
* I det nye <String…/> element sættes attributterne som vist i tabellen:

*I tabellen forstås at fx HEIGHT1 og HEIGHT2 er HEIGHT-værdierne fra hhv. den første og den sidste af de to sammenlagte <String…/> elementer*

|  |  |
| --- | --- |
| Ny attribut | Ny værdi |
| ID | = ID1 |
| STYLEREFS | = STYLEREFS1 |
| HPOS | = HPOS1 |
| VPOS | = minimum(VPOS1, VPOS2) |
| HEIGHT | = maximum(HEIGHT1, HEIGHT2) |
| WIDTH | = HPOS2 - HPOS1 + WIDTH2 |
| WC | = WC1 |
| LANG | = LANG1 |
| SUBS\_TYPE Oprettes kun hvis den findes på et af de to tidligere <String…/> | ”HypPart1” eller ”HypPart2” fra tidligere SUBS\_TYPE  (vil kun være til stede på ét af de to tidligere <String…/> elementer) |
| SUBS\_CONTENT Oprettes kun hvis den findes på et af de to tidligere <String…/> | **Hvis SUBS\_CONTENT2 eksisterer:**  = CONTENT1 + SUBS\_CONTENT2  og i begyndelsen af næste TextLine skal SUBS\_CONTENT udskiftes med samme værdi.  **Ellers hvis SUBS\_CONTENT1 eksisterer:**  = SUBS\_CONTENT1 + CONTENT2  og i slutningen af forrige TextLine skal SUBS\_CONTENT udskiftes med samme værdi. |

## Ikke-funktionelle krav

Der er ingen særlige krav til arkitektur, portabilitet og skalering.

Det er et krav at brugergrænsefladen skal fungere korrekt i nyere versioner af Google Chrome og FireFox i Windows 10 eller en nyere version af Safari i en Mac-computer. Internet Explorer skal ikke supporteres.

Værktøjet skal ikke anvendes på smartphones og tablets.

Værktøjet (programfilerne) skal placeres i et webkatalog på en af KB’s webservere.

Hvis det er tidsmæssigt fordelagtigt for softwareudviklingen, er det tilfredsstillende at værktøjet alene fungerer med programmering i front-end (JavaScript/jQuery) således at kun statiske filer leveres fra webserveren.

Det er også tilfredsstillende hvis værktøjet både har programmering i front-end og backend. Mht. programmering af backend er KB ikke særligt krævende og er åben for forslag. KB ønsker dog at der anvendes et **fortolket** sprog (fx Python og PHP), og ikke et sprog der nødvendiggør brug af compiler (fx Java, C# e.l.).

KB ville se det som en fordel hvis backend-koden er baseret på Python (med Django eller Flask) frem for fx PHP som KB ønsker at udfase.

Det er et krav at løsningen skal fungere med ”rimelig hurtighed” givet at der anvendes moderne computere, moderne browsere og hurtige internetforbindelser.

Det er et krav at den endelige løsning fungerer med meget basal adgangskontrol. Dette skal ikke programmeres i applikationen. I stedet vil KB’s IT-driftsafdeling etablere det således at applikationens webserverkatalog ikke kan tilgås medmindre brugeren er logget ind med sit KB-brugernavn i KB’s generiske loginløsning.

Al kildekode skal udleveres til Det Kgl. Bibliotek ved afslutning af softwareudviklingen.

Det foretrækkes, men det er ikke et krav, at softwareudvikleren som udvikler værktøjet, vil befinde sig fysisk i Den Sorte Diamant (etage 5, vest) i København under udviklingsarbejdet. I hvert fald en vis del af arbejdstiden. Det muliggør enklere adgang til filer og webserver, og vil desuden lette kontakten til KB’s projektleder (også etage 5, vest).

## Tip til softwareudvikleren

### Tip vedr. bevaring af ALTO-elementer og -attributter under redigering

Nedenstående fremgangsmåde er et tip til softwareudvikleren. Han/hun kan vælge en anden fremgangsmåde til bevaring og visning af ALTO-data.

En enkel måde hvorpå ALTO-elementer og -attributter kan bevares under redigering, er at transformere ALTO-elementer nogenlunde en-til-en til HTML-elementer hvor attributter forekommer med HTML5-præfikset ”data-”.

Eksempel på en linje i ALTO-filen:

<TextLine ID="LINE7" HEIGHT="102" WIDTH="1695" HPOS="185" VPOS="1117">

...

<String ID="STRING43" STYLEREFS="STYLE2" HEIGHT="72" WIDTH="112" HPOS="1177" VPOS="1127" CONTENT="ind" WC="0.36" LANG="dan"/>

<SP ID="SP37" WIDTH="25" HPOS="1289" VPOS="1127"/>

<String ID="STRING44" STYLEREFS="STYLE1" HEIGHT="94" WIDTH="129" HPOS="1314" VPOS="1117" CONTENT="(ft)." WC="0.95" LANG="dan"/>

...

</TextLine>

Kan fx repræsenteres således i HTML5:

<style>

span.SP { display: inline-block; width: 1px; }

div { white-space: nowrap; }

</style>

<div data-ID="LINE7" data-HEIGHT="102" data-WIDTH="1695" data-HPOS="185" data-VPOS="1117">

...

<span class=”String” data-ID="STRING43" data-STYLEREFS="STYLE2" data-HEIGHT="72" data-WIDTH="112" data-HPOS="1177" data-VPOS="1127" data-WC="0.36" data-LANG="dan">ind</span>

<span class=”SP” data-ID="SP37" data-WIDTH="25" data-HPOS="1289" data-VPOS="1127"></span>

<span class=”String” data-ID="STRING44" data-STYLEREFS="STYLE1" data-HEIGHT="94" data-WIDTH="129" data-HPOS="1314" data-VPOS="1117" data-WC="0.95" data-LANG="dan">(ft).</span>

...

</div>

Alle øvrige ikke-synlige dele af ALTO-filen kan transformeres til valide, skjulte dele af en HTML-side.

### Tip vedr. visning af orddelingsstreg

I forbindelse med orddeling ved linjeskift ønskes en orddelingsstreg vist når den første orddel ikke er i redigeringsmodus. Orddelingsstregen skal altså ikke vises når strengen redigeres.

Baseret på tippet ovenfor vedr. HTML5-attributter kan orddelingsstreg let opnås i visningsmodus med fx flg. CSS:

<style>

span[data-SUBS\_TYPE=HypPart1]::after { content: '-'; }

</style>