DE BASIS voor de vervaardiging van tekst

De richtlijnen in DE BASIS voor de vervaardiging van digitale tekst zijn primair bedoeld voor het creëren van digitale tekst, die zowel voor mens als machine leesbaar en doorzoekbaar is. Het gaat daarbij zowel om gedigitaliseerde als digitaal geboren tekst.

Afhankelijk van het eindgebruik worden inhoud, structuur en opmaak opgenomen in de digitale tekst en de bijbehorende metadata. Een handig overzicht van verschillende mogelijkheden, keuzes en workflows is te vinden in het kennisdossier voor digitale tekst. Bij DE BASIS staat voorop dat digitale tekst in een duurzaam en flexibel formaat wordt opgeslagen. DE BASIS voor de vervaardiging van digitale tekst bestaat uit een minimum richtlijn en een reeks van aanbevelingen.

Drie verschijningsvormen

Digitale tekst heeft drie mogelijke verschijningsvormen:

1. Machine-leesbare tekst (letterlijk tekst die door een machine kan worden doorzocht en geïnterpreteerd). Om de tekst goed te kunnen lezen is opmaak en tekststructuur nodig.
   Voorbeelden: teksten in DBNL, tekst als product van een tekstverwerker, e-books en teksten op een website.
2. Machine-leesbare tekst in combinatie met digitaal beeld. De machine leesbare tekst wordt dan vaak exclusief gebruikt om in te zoeken.
   Voorbeelden: gedigitaliseerde kranten van de KB, een PDF met ingesloten beeld en OCR output, of Google books.
3. Tekst (exclusief) als digitaal beeld via bijvoorbeeld fotografie of scanning. De tekst is dan niet machine doorzoekbaar en voldoet strikt gezien niet aan de minimale eisen van DE BASIS. Wanneer de tekst wel is voorzien van voldoende inhoudelijke kenmerken/metadata, zoals vaak het geval bij archiefmateriaal, is het natuurlijk wel zo dat de tekstbron goed vindbaar is.

De machine-leesbare vorm kan een papieren bron als origineel hebben die is getranscribeerd of automatisch gegenereerd (en eventueel gecorrigeerd) via OCR, of een digitale bron. In DE BASIS wordt in principe geen onderscheid gemaakt tussen deze twee vormen. *Het uiteindelijke eindproduct moet hetzelfde zijn*: een goed leesbare en doorzoekbare digitale tekst.

Fotografie of scanning vanaf een papieren origineel

Omdat de fotografie of scanning van tekst is in principe niet anders is als andere digitaliseringstrajecten vanaf een analoge bron kun je daarvoor DE BASIS en het kennisdossier voor vervaardigen van beeld gebruiken.

Minimum richtlijnen vervaardiging tekst

Om digitale tekst voor mens en machine-leesbaar en doorzoekbaar te maken schrijft DE BASIS minimaal voor:

- Het gebruik van een open bestandsformaat voor de opslag.
- Het gebruik van een tekenset die op Unicode gebaseerd is, bij voorkeur UTF-8, voor de codering van de tekst.

(Bron: Stadsarchief Amsterdam)

Aanbevolen richtlijnen vervaardiging tekst

Naast een minimum richtlijn geeft DE BASIS ook een aantal aanbevelingen. Om de digitale tekst naast lees- en doorzoekbaar ook goed navigeerbaar (voor mens en machine) te maken wordt het aanbrengen van een (XML) structuur sterk aanbevolen.

XML

- Gebruik XML om structuur toe te voegen aan je tekst en publiceer het bijhorende XML schema.
- Een verdergaande aanbeveling geldt het gebruik van een XML schema dat zich conformeert aan de TEI richtlijnen. Dit betekent dat het XML schema gebruik maakt van TEI conforme tags om de tekst semantisch te coderen. Door TEI tags te gebruiken kun je de tekst inhoudelijk door een machine laten analyseren. Omdat het toekennen van TEI tags een ingewikkelde klus is, wordt er over het algemeen gewerkt met behulp van speciale editors waarmee de tekst op eenvoudiger wijze kan worden gecodeerd.
- Wanneer je OCR inzet, gebruik dan het ALTO XML Schema als standaard datastructuur om de lay-out van de tekst te coderen. ALTO maakt gebruik

van coördinaten om de tekst te matchen met het digitale beeld, waardoor je tekst kan terugvinden in het digitale beeld.

PDF/A

Een alternatief voor het gebruik van XML is het opnemen van de tekststructuur in een PDF/A bestand. Dit wordt met name toegepast bij digitaal geboren tekst. Een voordeel hiervan ten opzichte van XML is dat de originele opmaak in het PDF/A bestand is gefixeerd. Dit kan voor projecten waarbij de authentieke opmaak van belang is, bijvoorbeeld bij officiële stukken, een belangrijk pré zijn. Een nadeel van die aanpak is dat de conversie van een PDF bestand naar een ander, bijvoorbeeld e-book formaat erg moeizaam is. Een gestructureerd XML formaat is veel eenvoudiger te converteren naar allerlei outputformaten. PDF/A wordt in DE BASIS daarom vooral aanbevolen wanneer ook de originele opmaak van de tekst in de digitale tekst behouden moet blijven.

Je kunt PDF/A ook gebruiken om gescande documenten en OCR output in één bestand op te nemen. DE BASIS beveelt een dergelijke samengesteld bestand echter alleen aan als raadplegingsbestand. Voor duurzame archivering en vanuit oogpunt van flexibiliteit is het beter om de scans en de OCR bestanden afzonderlijk te bewaren. De elektronische tekst en afbeeldingen kunnen dan in de toekomst los van elkaar naar elk gewenst formaat worden omgezet.

Opslagformaten voor archivering

DE BASIS schrijft het gebruik van een open bestandsformaat voor. Dat wil zeggen dat de specificaties van het formaat vrijelijk beschikbaar zijn. Het gebruik van een open standaard is een belangrijke voorwaarde voor interoperabiliteit en duurzame bewaring. In het kader van DE BASIS wordt het volgende aanbevolen:

- Machine-leesbare en gestructureerde tekstbestanden kun je het beste opslaan in een XML bestand voorzien van een UTF-8 tekenset.
- Voor opslag van gescande documenten wordt verwezen naar DE BASIS voor vervaardiging van beeld.
- OCR bestanden kun je het beste opslaan in platte tekst of XML.
- Wanneer gebruik wordt gemaakt van PDF/A worden twee varianten aanbevolen:
    - Bij digitaal geboren tekst: gebruik de PDF/A -1a of- 2a variant die een rijke, op XML gebaseerde, logische tekststructuur toestaat.
    - Bij gedigitaliseerde tekst: met PDF/A-1b of 2b kunnen digitale afbeelding en de ge-ocr'de tekst in een samengesteld en doorzoekbaar bestand worden opgenomen. Dit wel met bovengenoemd voorbehoud.

Koppeling van afbeeldingen aan OCR output

Om de afbeeldingsbestanden aan de OCR tekst en eventuele andere bestanden aan elkaar te koppelen wordt het gebruik van METS of MPEG-21 DIDL

aanbevolen. Het wordt afgeraden om de structuur exclusief vast te leggen in bestandsnamen en/of mappenstructuren. Het nadeel daarvan is dat je de structuur daarna moeilijk kunt verbeteren of veranderen.

Verantwoording

Dit is een herziene, tweede versie van DE BASIS voor vervaardiging van tekst. De eerste versie stamt uit 2008. Deze herziene tekst is opgesteld door experts die werkzaam op het gebied van tekstvervaardiging. De tekst is ook voorgelegd aan de adviesraad van de kennisbank. Tenslotte is de tekst open gesteld voor commentaar vanuit de hele erfgoedsector.

Denk mee!

Deze tekst staat altijd open voor commentaar. Wil je reageren of meedenken? Maak dan gebruik van het onderstaande reactieformulier of lever je commentaar rechtsreeks in de tekst via dit bestand. Stuur deze naar den@den.nl (onder vermelding van DE BASIS voor vervaardiging van tekst).