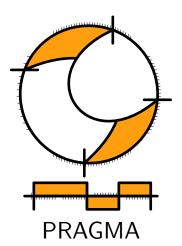
Typografisch Programmeren

Hans Hagen

Voorjaar 1996

In dit artikel wordt de ontwikkeling van een interactief document beschreven. De getoonde voorbeelden demonstreren de kracht van $T_{\rm E}$ X als typografische programmeertaal. Omdat dergelijke documenten het moeten opnemen tegen (vaak dedicated) programma's, kan het ontwikkelen ervan worden omschreven als typografisch programmeren.



Advanced Document Engineering | Ridderstraat 27 | 8061GH Hasselt NL tel: +31 (0)38 477 53 69 | e-mail: pragma@wxs.nl | ConTEXt: www.pragma-ade.nl

Dit artikel is eerder in kleur verschenen in de *Minutes and Appendices* van de NTG (Nederlandstalige T_EX Gebruikersgroep), Nummer 96.1.

Inleiding

Met de komst van (1) snelle computers, (2) opslagmedia met grote capaciteit en (3) viewers als Acrobat Reader dient zich voor TeX een nieuw werkterrein aan: de produktie van complexe, interactieve teksten. In dit artikel gaan we in op enkele aspecten van het produceren van zulke teksten.

De afdeling Control van de Koninklijke Landmacht is verantwoordelijk voor de distributie van documenten binnen dit krijgsmachtdeel. We praten in dat verband al snel over duizenden documenten, variërend van een partituur voor het Wilhelmus (2 varianten) tot beschrijvingen van een Leopard-tank. Omdat het begrip publicatie in de ruimste betekenis van het woord wordt opgevat, komen we ook stickers, formulieren en slablonen tegen.

Elk jaar worden er door de betreffende afdeling catalogi uitgegeven van publicaties. Bovendien worden in de loop van het jaar updates rondgestuurd. Aangezien het aantal publicaties in de duizenden loopt en het aantal abonnee's op de catalogi rond de tweeduizend ligt, zal duidelijk zijn dat de jaarlijkse verschijning ervan aardig wat papier kost.

Tot voor kort werden de catalogi samengesteld uit prints van de database waarin de gegevens zijn opgeslagen. Zowel de omvang in termen van papier als de wens volledige updates te verstrekken, zijn aanleiding geweest elektronische uitgave te overwegen. De beschikbaarheid van ACROBAT READER op verschillende platforms en de zich aandienende standaard formaat PDF voor elektronische documenten, opende de weg daartoe. Overigens werken op dit moment de meeste gebruikers met MSDOS.

In dit artikel gaan we in op de wijze waarop twee catalogi met behulp van het in huis ontwikkelde macropakket Context zijn geproduceerd. We doen dit aan de hand van voorbeelden van de schermen zoals de gebruiker die ook te zien krijgt.

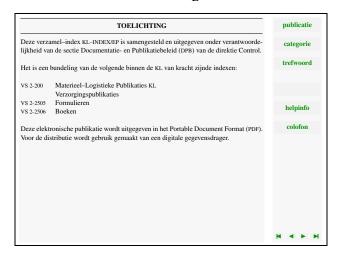
Het eerste document

We beginnen met de eerste catalogus: die van afzonderlijke publicaties.

Bij wijze van introductie krijgt de gebruiker een titelblad te zien. Aan de rechterkant van het scherm



zien we het navigatiegedeelte, links staat de tekst. De aan te klikken delen hebben een wat zwaardere achtergrond en de woorden cq. karakters worden weergegeven in kleur. Zowel achtergronden als aanvullende ruimte voor navigatie-instrumenten worden standaard ondersteund door CONTEXT.



In de toelichting vertellen we uit welke databases het document is opgebouwd. Het is de ruwe uitvoer van deze bestanden die aan de basis staat van het document. Alle gegevens van een publicatie stonden daarbij op één regel.

In eerste instantie lieten we T_EX de verschillende velden uit de regel filteren. Dit gebeurde op basis van een serie database-macro's, waarbij de velden op naam konden worden opgeroepen. Bovendien was een mechanisme ingebouwd om gecodeerde velden om te zetten. Tegenover het voordeel van onafhankelijkheid stond het nadeel van lange rekentijden. Dit soort handelingen zijn nu eenmaal niet T_EX's sterkste kant. Bovendien gaven bijzondere en actieve karakters problemen. (Hier zouden enkele e-T_EX opties wellicht uit-

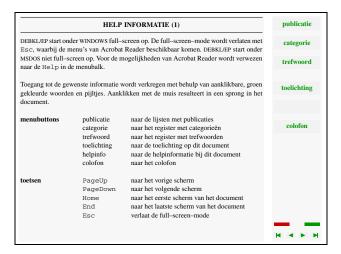
komst hebben geboden.)

Uiteindelijk is besloten de aangeleverde bestanden om te zetten in een wat sneller te behappen formaat. Het hiertoe geschreven programma zal te zijner tijd worden uitgebreid, maar kon in dit verband worden beperkt tot het (met behulp van een definitiefile) genereren van een file met commando's als:

```
\beginofrecord
\field{name 1}{data of field 1}
\field{name 2}{data of field 2}
\endofrecord
```

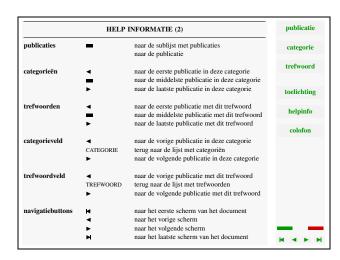
Bij het verwerken van dergelijke records en velden wordt veelvuldig gebruik gemaakt van \csname, maar dat terzijde.

Het verwerken van het eerste bestand, dat uit bijna 3000 records bestaat, kost als we afzien van interactiviteit ongeveer een half uur. Het verwerken van de definitieve interactieve versie duurt drie maal zo lang. Deze verwerkingstijden zijn mede een gevolg van de vrij uitgebreide functionaliteit van ConTEXT. Hier betalen we de prijs van flexibiliteit.

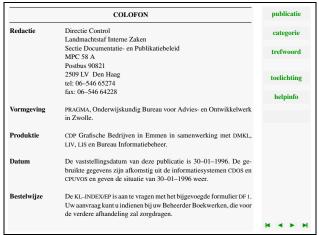


Bij de helpinformatie zien we naast de buttons voor heen- en weerbladeren een nieuw navigatie-instrument opduiken. De twee blokjes staan voor de twee pagina's waaruit de helpinformatie bestaat. Deze blokjes worden automatisch gegenereerd en maken gebruik van het in ConTEXT aanwezige mechanisme om subpagina's te nummeren.

Het subpaginanummer is zichtbaar gemaakt in de titel. Deze titel wordt herhaald omdat, in tegenstelling tot een papieren document, in een interactief document de tactiele en visuele signalen ontbreken met be-



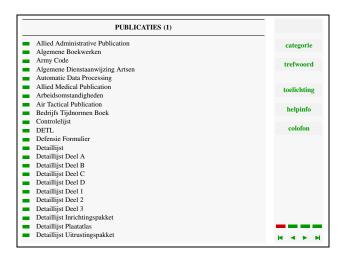
trekking tot de plaats in het document. Bovendien kunnen we minder informatie kwijt op een bladzijde. Hoewel een soort \everypage denkbaar is, is hier gebruik gemaakt van kopregels (headers). Omdat een kopregel meerdere regels tekst kan bevatten, wordt per hoofdstuk de kopregel opnieuw ingesteld.

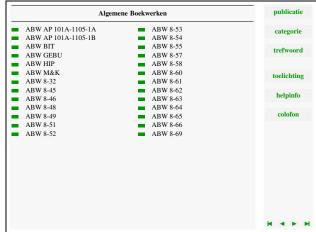


Naast deze introducerende bladzijden bevat het document eigenlijk nauwelijks typografisch aantrekkelijke informatie. Dit is mede het gevolg van de beperkingen van databases. Bovendien wordt inconsistent gebruik van spaties, leestekens en haakjes, door TeX genadeloos afgestraft in de vorm van overfull boxes. Vandaar dat in dit document het afbreekmechanisme soepel is ingesteld. Bovendien worden enkele velden (door TeX) op orde gebracht. Zo zetten we afgebroken woorden als zie- zo om in ziezo. Dergelijke afbrekingen komen voor omdat in de database de informatie over meerdere regels kan worden ingevoerd.

De publicaties zijn ondergebracht in ongeveer vijftig categorieën. Daarnaast is aan iedere publicatie een trefwoord gekoppeld en heeft elke publicatie een uniek

PRAGMA ADE





nummer.

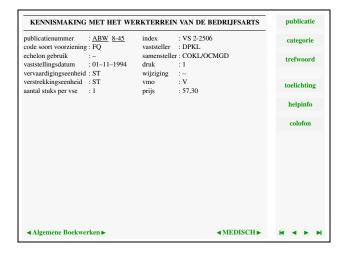
De eerste ingang in het document loopt via het publicatienummer. Via een categorie komt men bij de lijsten met publicaties per categorie. Deze lijsten worden automatisch gegenereerd.

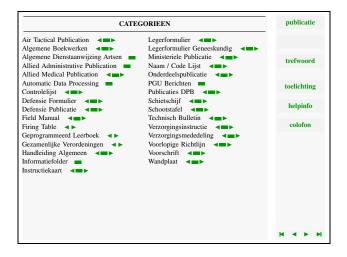
Binnen ConT_EXT kan men in principe een onbeperkt aantal lijsten definiëren. Standaard zijn lijsten beschikbaar met hoofdstukken, paragrafen, subparagrafen enz. Dergelijke lijsten kunnen worden gegroepeerd tot samengestelde lijsten, zoals inhoudsopgaven. Lijsten en samengestelde lijsten kunnen op ieder niveau worden opgeroepen. Binnen een hoofdstuk krijgt men dus alleen de paragrafen, subparagrafen enz. te zien. Bij interactieve teksten zijn de lijsten automatisch doorgelust, zodat men snel door een tekst kan lopen (van inhoud naar subinhoud naar subsubinhoud enz.).

Omdat in dit document een structuur in termen van hoofdstukken (secties) niet erg voor de hand ligt, is gekozen voor het genereren van ruim vijftig afzonderlijke lijsten.

Via de tussenstap *publicaties* (een lijst) komt men bij de afzonderlijke lijsten met publicatienummers. Een kritische lezer zal zich afvragen of het tonen van nummers zin heeft. Welnu, veel van de gebruikers zijn gewend om met nummers om te gaan en vertalen nummers automatisch naar de betekenis.

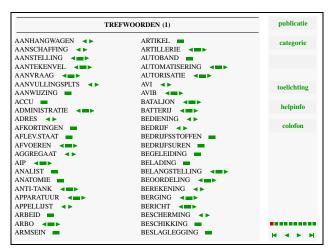
Bij de publicatie vinden we links onder de categorie en rechts het trefwoord. Omdat er meerdere publicaties per categorie en trefwoord mogelijk zijn, wordt de mogelijkheid geboden naar verwante publicaties te springen. In een interactief document van deze omvang is het namelijk vrij zinloos om achter een ingang in een register enige tientallen paginanummers te tonen. Als we op het woord klikken keren we terug naar de relevante bladzijde in het register.





Overigens geldt voor registers een vergelijkbaar verhaal als voor lijsten. Er kan een onbeperkt aantal registers worden gedefinieerd, die op elk niveau als deelregister kunnen worden opgeroepen (per deel, hoofdstuk, enz. of per zelf te definiëren sectie).

Het eerste register meldt de categorieën en is dus een variant op de eerder besproken lijst.



Het tweede register beslaat tien bladzijden, wat te zien is aan de twee na onderste navigatiebalk. Het blokje dat bij de actuele bladzijde hoort is rood, de rest is groen. De balkjes krijgen een andere vorm als de afstand ertussen onder een kritische grens komt. De reden hiervoor ligt in het afnemend oplossend vermogen van het aanklikbare deel en de onnodig grote overhead in termen van bytes die dergelijke sprongen met zich meebrengen.

Er zijn een, twee of drie opties mogelijk. Als er slechts één verwijzing is, dan tonen we een blokje. Als er twee zijn, dan tonen we twee pijltjes, waarmee men naar het eerste of laatste trefwoord kan springen. Bij drie of meer verwijzingen gaan we met het middelste blokje naar het midden van de reeks bij het trefwoord behorende publicaties. Daar aangekomen kan men stapsgewijs de publicaties langs lopen, wat handiger is dan steeds weer terug te moeten keren naar het register.

Bij omvangrijke documenten komt ACROBAT in de problemen als we gebruik maken van de /dest operator onder pdfmark's. Omdat de tekst geschikt moet zijn om onder MSDOS te raadplegen, is echter gebruik gemaakt van het springen naar bladzijden. Omdat de MSDOS versie achterloopt in de ontwikkeling van pdfmark's, is springen naar andere documenten niet mogelijk.

We kunnen dus niet werken in termen van hier ben

ik en *ga daar naar toe*, maar moeten het doen met *ga naar bladzijde*. Dit betekent dat alle relevante paginanummers tijdens de verwerking door T_EX beschikbaar moeten zijn. Met wat kunstgrepen lukt dat vrij aardig, zij het dat Huge EMT_EX onmisbaar is.

Het tweede document

Het tweede document is niet alleen veel omvangrijker (bijna 40.000 records) maar ook veel gecompliceerder. Het betreft hier namelijk in projecten gegroepeerde publicaties, waarbij binnen een groep onderscheid wordt gemaakt in hoofdpublicaties en 'verborgen' publicaties.

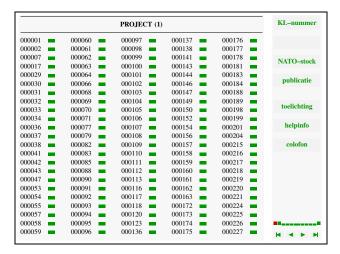


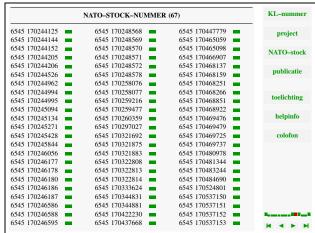
Omdat het niet realistisch is ruim 40.000 bladzijden te genereren, is besloten de gegevens te groeperen. Daarbij is onderscheid gemaakt in algemene informatie, hoofdpublicaties en verborgen publicaties. Het bestand is daartoe gesorteerd, waarna aan TEX de zorg voor de nadere ordening en controles op consistentie is overgelaten. In tegenstelling tot het eerste bestand kan een groep publicaties zich hier uitstrekken over meerdere bladzijden. Ondanks de compacte vorm komen we toch nog op bijna 8.500 pagina's uit.

In dit document zijn alleen registers gebruikt. Veel ruimte is daarbij bespaard door te werken in kolommen. Het aantal kolommen wordt binnen ConTEXT alleen begrensd door het maximaal aantal te alloceren boxen. Om te voorkomen dat de registers onoverzichtelijk lang worden, zijn verborgen publicaties niet opgenomen.

Bij documenten met een dergelijke omvang, komen we uit bij iets anders vormgegeven navigatie-instrumenten. Ook hier zien we een uit blokjes opgebouw-

4 PRAGMA ADE





de balk rechts-onder. In dit geval is de resolutie echter (automatisch) aangepast aan de beschikbare ruimte. Men kan steeds naar de eerste en laatste bladzijde van een subpaginabereik springen. Bovendien kan men naar de aangrenzende bladzijden gaan en in grote sprongen naar de overige. Al bladerend zien we in zo'n balk de wat hogere blokjes van links naar rechts schuiven.

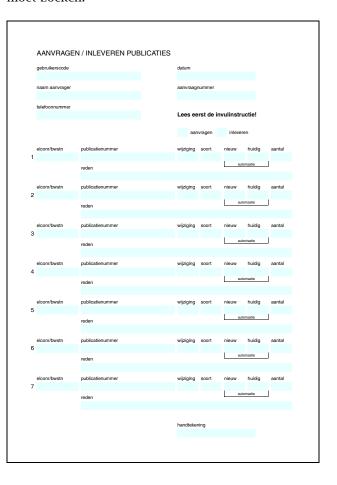
KL-nummer PUBLICATIE-NUMMER (3) 1/2DL 6230-17-107-4708 1/2DL 5810-17-030-9001 5895 - 010166961 project 1/2DL 5985-17-036-6397 5810 - 170309001 - 1/2DL 5810-17-043-0301 6230 - 171074708 1/2DL 6515-17-046-8293 5985 - 170366397 NATO-stock 5810 - 170430301 = 1/2DL 5810-17-046-1980 1/2DL 5985-17-036-9860 6515 - 170468293 - 170369860 1/2DL 6530-17-055-8748 publicatie 5810 - 170461980 1/2DL 6130-17-047-2070 6530 - 170558748 1/2DL 5815-17-035-7787 5815 - 170357787 - 170472070 1/2DL 6605-17-047-7448 1/2DL 6130-17-103-2847 toelichting 1/2DL 5815-17-035-7788 6130 - 171032847 1/2DL 6625-01-063-6628 5815 - 170357788 1/2DL 6130-17-103-3175 6625 - 010550802 1/2DL 5825-17-106-0045 helpinfo 6130 - 171033175 - 010636628 5825 - 171060045 1/2DL 6130-17-106-6194 6625 - 011134262 1/2DL 5855-17-035-2846 5855 - 170352846 1/2DL 5855-17-049-5863 6130 - 171066194 **■** 1/2DL 6145-00-226-8812 6145 - 002268812 6625 - 170540600 1450 - 010274616 2350 - 010274623 1/2DL 6145-17-036-9410 1/2DL 6625-17-038-8807 2350 - 010274624 1/2DL 6230-17-050-7591 1/2DL 6625-17-046-8863 2350 - 010290637 6760 - 170507591 6625 - 170468863 2350 - 010310311 6760 - 170540838 1/2DL 6625-17-051-6134 5855 - 170495863 6780 - 170529689 6625 - 170516134 1/2DL 5895-01-016-6961 1/2DL 6230-17-056-1980 1/2DL 6625-17-051-6656 5820 - 170352178 6230 - 170561980 6625 - 170516656

Een van de registers heeft sub-ingangen. Dit wat cryptische register is voor kenners wel hanteerbaar, omdat de gebruikers bij de nummers een soort document voor de geest kunnen halen.

Het bestelformulier

Tot slot is er nog het bestelformulier. De getoonde variant is een vrijwel een-op-een vertaling van het bestaande document en wordt dan ook in A4 formaat getoond. Klikt men op een veld, dan wordt de bijbehorende helpinformatie getoond. Van daaruit kan men weer terugspringen naar het formulier. Helaas gaat binnen

ACROBAT READER de cursor niet op de bestemming staan, zodat de gebruiker (met zijn ogen) weer het veld moet zoeken.



Enkele opmerkingen

Tijdens het ontwerpen van dergelijke documenten ontkomt men niet aan het op punten uitbreiden en optimaliseren van de functionaliteit van de gebruikte systemen. We gaven reeds aan dat 'huge' T_EX noodzakelijk is. Hetzelfde geldt voor het hulpprogramma dat de registers sorteert. Nog niet vermeld is dat ConT_EXT slechts één hulpfile gebruikt. Dat wil zeggen dat alle lijsten, registers, synoniemlijsten, verwijzingen en andere noodzakelijke informatie, zoals subpaginaclusters in een file worden geplaatst. Deze file wordt door T_EXUTIL omgezet in een tijdens de volgende verwerkingsslag in te lezen hulpfile, waarin de in aanmerking komende lijsten en registers zijn gesorteerd. ConT_EXT haalt op gezette tijden de nodige informatie uit deze file.

Aangezien de behoefte de capaciteit van T_EXUTIL ruim te boven ging is de capaciteit aanzienlijk vergroot. De bijna veertigduizend te verwerken ingangen in de registers van het tweede document liggen momenteel ruim onder de eenvoudig te verhogen limiet van 250.000 in T_EXUTIL.

Het karakter van ConT_EXT brengt met zich mee dat relatief veel rekentijd wordt opgeofferd aan de gebruikersinterface. Desondanks blijft onze voorkeur uitgaan naar gebruik van standaard ConT_EXT mogelijkheden boven speciaal ontwikkelde macro's. Wel is zo'n project een goede gelegenheid de kritische onderdelen van een macropakket te optimaliseren. Zo leverde bijvoorbeeld de optimalisatie van het mechanisme dat de actuele kleuren over de paginagrenzen tilt, een verdubbeling van de verwerkingssnelheid op.

Het zoekmechanisme van ACROBAT READER is bij dergelijke files nauwelijks te gebruiken. Ten eerste is de omvang van de files te groot, daarnaast is het lastig zoeken in de bonte mix van woorden, afkortingen en getallen. Om eenzelfde reden kan ook geen gebruik worden gemaakt van ACROBAT CATALOG. (Wat onder MSDOS overigens niet beschikbaar is.) Een lijst met niet in een index op te nemen woorden is nauwelijks op te stellen. Bovendien is de omvang van deze lijst beperkt. Een ander nadeel is dat de gegenereerde bestanden bijna net zoveel ruimte in beslag nemen als de tekst zelf. Het blijft dus T_FX-werk.

Om de gegevens te sorteren is een programma geschreven waarmee zeer omvangrijke bestanden kunnen worden gesorteerd op meerdere te specificeren velden. Gelukkig zijn dergelijke programma's acceptabel snel te programmeren.

Tegenover het voordeel van platform-onafhankelijkheid en typografische kwaliteit staat het nadeel van omvangrijke files. De verwerkingstijd van dergelijke omvangrijke bestanden is niet onaanzienlijk, ook bij gebruik van snelle systemen. Bovendien worden grote tussenprodukten gegenereerd, zodat ook veel schijfruimte nodig is. Vandaar dat de meeste verwerkingslagen 's nachts in batch hebben plaatsgevonden.

bestand	DVI	POSTSCRIPT	PDF	blz
debkl-ep	19.578	22.131	15.351	3.046
debkl-vs	53.688	62.611	42.258	8.477
debkl-ef	30	55	29	4
gezipped	3.458	3.505	20.602	

In de bijgaande tabel is in KiloBytes de omvang van de verschillende files vermeld. De omvang van de DVI-files is mede bepaald door de aanwezigheid van de vele specials voor kleur en hyperlinks. Daarbij kan worden opgemerkt dat zowel ACROBAT als DVIWINDO wordt ondersteund. In de POSTSCRIPT-file zijn alleen pdfmark's opgenomen. De POSTSCRIPT-file wordt aan ACROBAT DISTILLER aangeboden zonder font-outlines.

De PDF-file is, wat we ook mogen verwachten, kleiner dan de POSTSCRIPT-file. Omdat we 'partial font embedding' toepassen, zijn hierin wel de fonts (deels) opgenomen. De omvang van PDF-files wordt echter hoofdzakelijk bepaald door het aantal sprongen. Op dit moment neemt ACROBAT DISTILLER pdfmark's bijna verbatim in de PDF-file over, terwijl andere informatie wordt gecomprimeerd. Mijns inziens is een besparing van 50 bytes per sprong zonder problemen te realiseren. Wat dit oplevert bij vele tienduizenden sprongen laat zich eenvoudig uitrekenen.

Een aardige bijkomstigheid is dat de PDF-file zich veel minder efficiënt laat comprimeren dan de DVI- en POSTSCRIPT-files. Waarschijnlijk heeft dit te maken met de wijze waarop binnen het PDF-formaat de gegevens worden gestructureerd.

Wellicht kan de toegankelijkheid worden vergroot door gebruik te maken van gerichte plug-ins, bijvoorbeeld om registers te doorlopen. Meer dan een toevoeging kunnen deze echter niet zijn, omdat we afhankelijk zijn van de beschikbaarheid ervan.

Tot slot

Kijkend naar de aard en omvang van het document en

6 PRAGMA ADE

het aantal verwijzingen dat daarvan het gevolg is, kunnen we onszelf de vraag stellen wanneer we de grens van het redelijke overschrijden. Met name het tweede hierboven beschreven document zit dicht tegen deze grens aan. De doorslaggevende redenen om gebruikte maken van ACROBAT liggen vooral in de beschikbaarheid op meerdere platforms en het onderliggende bestands-formaat. TeX voegt daar nog een meer dan gemiddelde typografische kwaliteit aan toe.