PDF genereren voor e-readers

Abstract
NotuDoc is een commerciële internet applicatie die ConTEXt gebruikt voor het on-the-fly genereren van pdf documenten, onder andere voor de e-readers van iRex technologies. Dit artikel geeft een blik achter de schermen.

Inleiding

Dit artikel gaat over het genereren van PDF bestanden voor e-readers. Voordat we daarop dieper ingaan, eerst wat uitleg over de applicatie (NotuDoc) waar dit proces een onderdeel van is, en een korte introductie van de gebruikte e-readers.

Het genereren van een PDF document vanuit NotuDoc is slechts een relatief klein onderdeel van de applicatie, maar toch komt daar al wel het een en ander bij kijken.

NotuDoc

Het bedrijf NotuBiz (http://notubiz.nl) houdt zich bezig met alles wat er komt kijken bij het vastleggen en publiceren van (raads)vergaderingen via de moderne media. NotuBiz verzorgt bijvoorbeeld live streaming en ook digitale verslagen kunnen via NotuBiz op het internet beschikbaar gesteld worden. De klantenkring bestaat uit de lokale overheden, uiteraard met name in Nederland.

In de praktijk bleek dat de informatiestroom *voorafgaand* aan de vergaderingen vaak ook aanzienlijk verbeterd kon worden. NotuDoc is uit dit idee ontstaan: het is een internet applicatie die (al dan niet voorlopige) vergaderagenda's koppelt aan de bijbehorende vergaderstukken zoals commissierapporten, presentaties en offertes. Na deze koppeling wordt het resultaat beschikbaar gesteld aan de relevante partijen via het netwerk en/of via PDF bestandsexport.

Door alle benodigde vergaderstukken te combineren op één plaats wordt het makkelijker voor de deelnemers aan de vergadering om zich voor te bereiden. Bovendien staat na afloop van de vergadering nu alles al klaar voor officiële publicatie, wellicht na een koppeling met het verslag van de vergadering.

NotuDoc is een kant-en-klaar commercieel product, en heeft vrij uitgebreide configuratiemogelijkheden. Dat is met name van belang omdat de interface moet aansluiten bij de vormgeving van de website van de klant, maar ook een het gaat nog iets verder: niet alle klanten hebben dezelfde vergader-structuur en zeker niet allen hebben dezelfde opzet voor intern databeheer. Verschillende variaties daarvan worden standaard aangeboden als onderdeel van NotuDoc, ingrijpendere veranderingen (denk aan koppelingen met document management systemen) zijn mogelijk op offertebasis. NotuDoc is geschreven en wordt onderhouden door Elvenkind BV in samenwerking met NotuBiz, en maakt gebruik van Elvenkind's development framework dat geschreven is in perl 5.

E-readers

Op dit moment is NotuDoc voorbereid op het genereren van PDF documenten voor twee verschillende e-readers, beide ontwikkeld door iRex Technologies (http://www.irextechnologies.com), een spin–off van Philips. Naast deze twee apparaten (iLiad en DR1000) is het uiteraard ook mogelijk om PDFs te genereren voor printing of voor interactief gebruik op een computer.

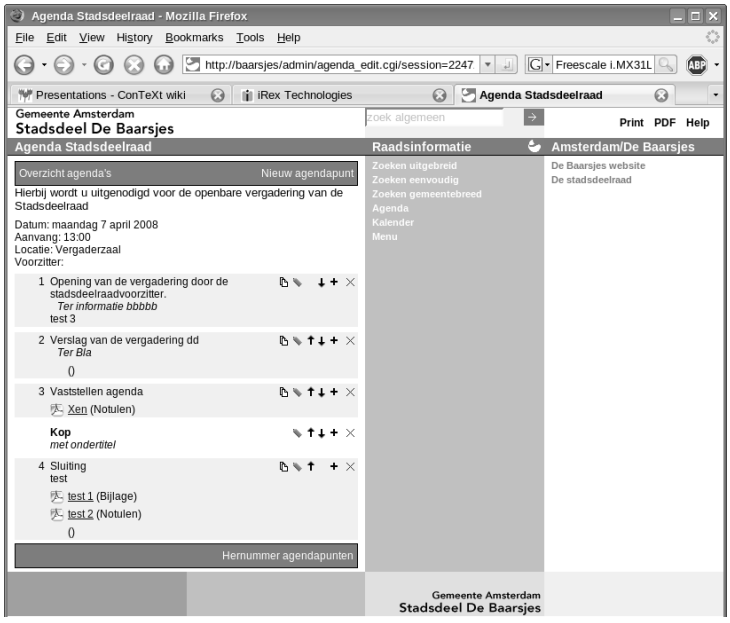


Figure 1. Hoofdscherm van de internet applicatie

Zowel de iLiad als de nieuwere DR1000 zijn gebaseerd op dezelfde basis-technologie. De iLiad bestaat nu al enkele jaren, en wordt onder andere gebruikt voor de digitale versie van het NRC Handelsblad. De DR1000 is een maand geleden gelanceerd. Zoals te verwachten is, is de DR1000 wat groter en sneller dan zijn voorganger, maar verder zijn er op de vormgeving na weinig technologische verschillen.

Beide apparaten zijn gebaseerd op zogenaamd `elektronisch papier', een technologie waarbij het getoonde zichtbaar blijft ook als het scherm *niet* vele malen per seconde ververst wordt.

Een belangrijk voordeel van deze technologie is dat er hierdoor veel minder stroom nodig is, waardoor de levensduur van de batterijen veel langer is dan bij de gewone TFT of LCD schermen. Een bijkomend voordeel is dat het scherm niet constant verlicht hoeft te worden, wat veel rustiger is voor het oog van de lezer.

Anderzijds zijn er natuurlijk nadelen aan elektronisch papier. De twee grootste daarvan zijn dat de reactiesnelheid van het scherm veel lager is dan bij gewone computerschermen en (het meest in het oog springend) dat de huidige versies alleen in staat zijn om grijstonen te tonen, geen kleur.

Beide apparaten gebruiken tevens de Wacom Penabled technologie (http://www.wacom.com/tabletpc/what_is_penabled.cfm) die het mogelijk maakt om rechtstreeks op het scherm aantekeningen en schetsen te maken, zodat je bijvoorbeeld correctie-aantekeningen in een PDF kunt maken. De bijgeleverde (windows) software is in staat zulke aantekeningen te combineren met de originele PDF, bijvoorbeeld voor verzending per email.

Beide apparaten bieden ondersteuning voor PDF, HTML, mobipocket, en enkele bitmap formaten. Alle door iRex gebruikte en ontwikkelde software is open source die werkt op een linux versie die speciaal bedoeld is voor consumentenapparaten. Verbinding maken met de PC voor het uploaden van bestanden gebeurt via USB of via een optioneel tcp/ip (wireless) netwerk. Het opslagmedium is een verwisselbaar SD (DR1000) of CF/MMC (iLiad) kaartje.

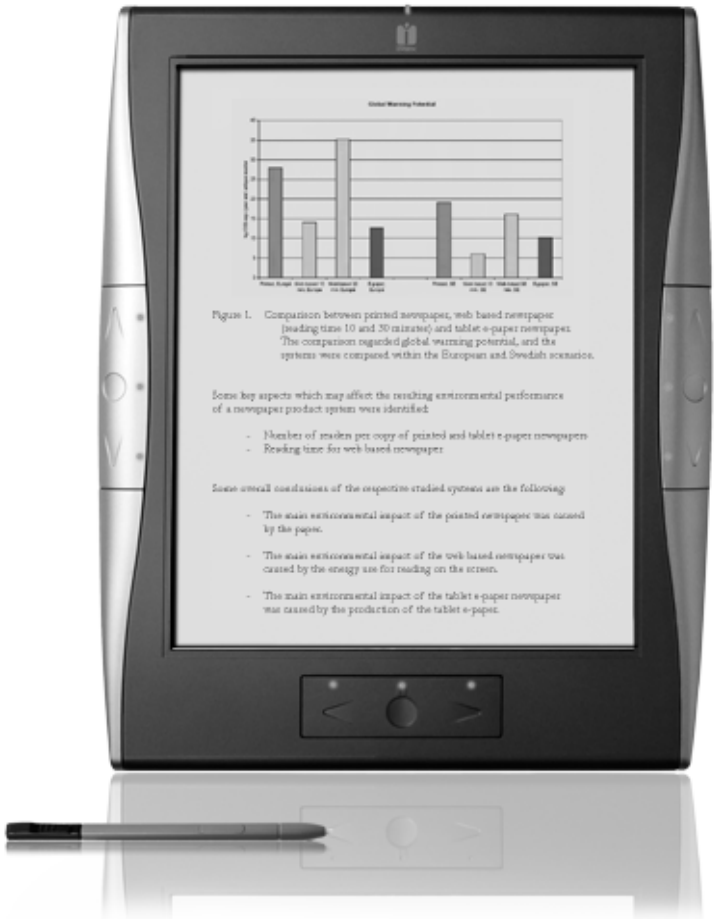


Figure 2. De iLiad (links) en DR1000 (rechts).

PDF generatie

De PDF generatie in NotuDoc wordt gedaan door een perl script dat volledig template-gestuurd is. Het gebruikt vrijwel identieke code voor het genereren van TEX als voor de generatie van de HTML pagina's, alleen de character escape functies en de bestandsnamen zijn specifiek voor TEX aangepast. Net als de website worden de PDF documenten ook runtime gegenereerd via een aanroep van `texexec`. De gebruikte distributie is ConTEXt minimals (http://minimals.contextgarden.net/).

ConTEXt templates

De applicatie heeft per klant een instelling opgeslagen voor de gewenste PDF uitvoer layout. Als voorbeeld neem ik de `iLiad', maar het kan ook iets anders zijn zoals `DR1000' of gewoon `A4'. Los van deze globale voorkeur is het mogelijk om per klant de opmaak te configureren zodat die aansluit bij de huisstijl.

De opmaak instellingen worden gedaan via ConTEXt macros en zijn gescheiden van de afhandeling vanuit de web applicatie. De applicatie draagt alleen zorg voor het omwerken van de database gegevens van de agenda naar een ConTEXt invoerbestand en het exporteren van de bijbehorende vergaderstukken naar PDF bestanden. Al het andere wordt gedaan door de ConTEXt macro files die door het perl script alleen maar gekopieerd worden.

`agenda-iliad.tex`
Dit is het hoofd bestand, en hierin vinden twee verschillende soorten vervangingen plaats.

In de listing hieronder zie je twee regels staan die lijken op de HTML syntax voor zogenaamde `server side includes'. Dat is geen toeval: zoals hierboven al vermeld werd gebruikt het systeem voor de generatie van de TEX bestanden dezelfde code als voor het aanmaken van de HTML pagina's.

De twee `#include` bestandjes worden ingelezen door het perl script en op die plek in de TEX uitvoer tussengevoegd. De exacte inhoud van deze bestanden wordt in de volgende paragraaf uitgelegd.

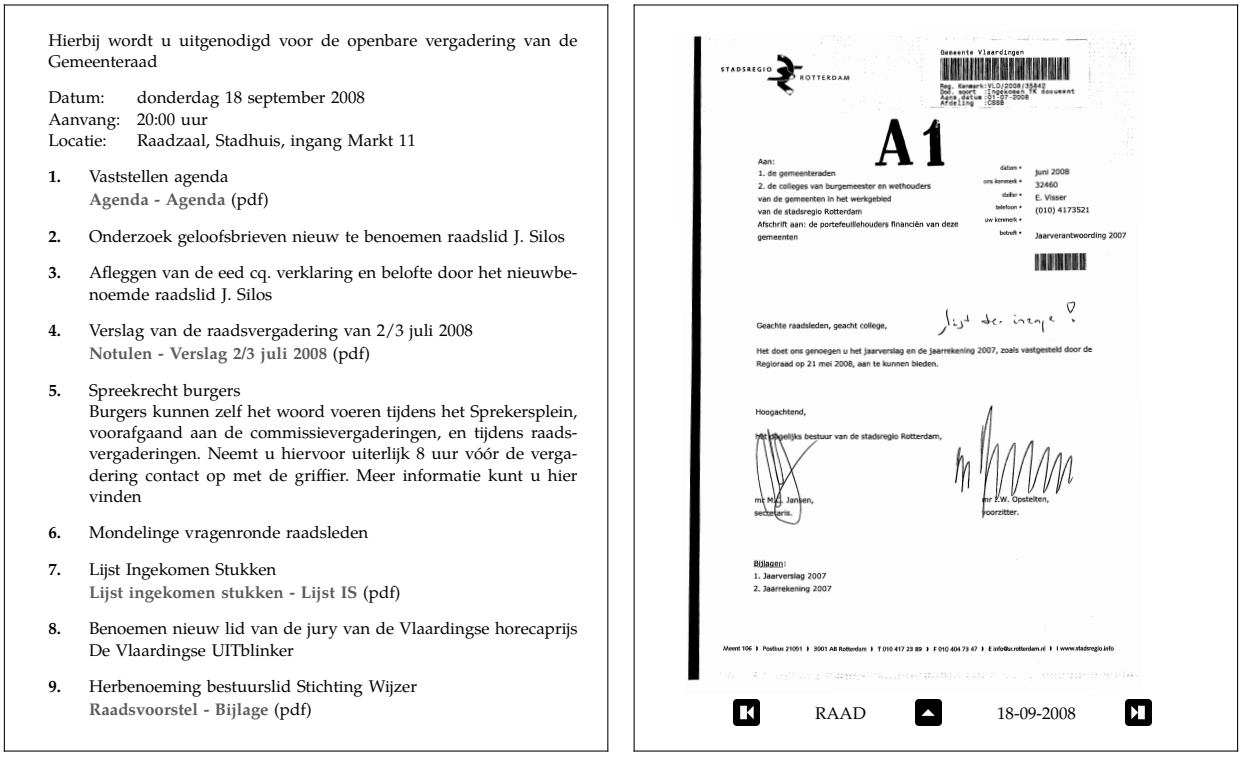
De tweede soort vervanging gebruikt de trefwoorden in hoofdletters en tussen # tekens. Die trefwoorden worden vervangen door de feitelijke inhoud (en metadata) van de agenda. Het trefwoord `#LIST#` is daarbij het belangrijkst omdat daarin zich effectief de hele inhoud van de agenda bevindt, die wordt namelijk recursief wordt opgebouwd.

Een agenda bestaat uit meta-informatie zoals plaats en tijd, en een aantal agendapunten. Agendapunten kunnen eventueel gerangschikt zijn in categoriën, en optioneel kan er per agendapunt een aantal agendastukken zijn in verschillende bestandsformaten. Dit alles wordt aangestuurd door kleine template bestandjes die op diverse niveaus worden aangeroepen.

```
\unprotect
<!--#include src='agenda-macros-00.tex' -->
<!--#include src='agenda-macros-iliad.tex' -->
\protect

\starttext
\startagenda[Gremium={#GREMIUM#},
             Datum={#DATUM#},
             Datumkort={#DATUMKORT#},
             Categorie={#CATEGORIE#},
             Aanvang={#AANVANG#},
             Locatie={#LOCATIE#},
             Aanhef={#AANHEF#},
             Koptitel={#TITLE2#},
             Titel={#TITLE#}]

\startpunten
#LIST#
\stoppunten

\stopagenda
\stoptext
```

`header_line.tinc`
Dit template wordt gebruikt voor de tussenkopjes die behoren bij eventuele categoriën van vergaderpunten.

```
\startheaderline
  [Titel={#TEXT#},Pagina=#PAGE#,Aard={#AARD#}]
\startheaderbody
#BODY#
\stopheaderbody
\stopheaderline
```



Hierbij wordt u uitgenodigd voor de openbare vergadering van de Gemeenteraad

Datum: donderdag 18 september 2008
Aanvang: 20:00 uur
Locatie: Raadzaal, Stadhuis, ingang Markt 11

1. Vaststellen agenda
   Agenda - Agenda (pdf)
2. Onderzoek geloofsbrieven nieuw te benoemen raadslid J. Silos
3. Afleggen van de eed cq. verklaring en belofte door het nieuwbenoemde raadslid J. Silos
4. Verslag van de raadsvergadering van 2/3 juli 2008
   Notulen - Verslag 2/3 juli 2008 (pdf)
5. Spreekrecht burgers
   Burgers kunnen zelf het woord voeren tijdens het Sprekersplein, voorafgaand aan de commissievergaderingen, en tijdens raadsvergaderingen. Neemt u hiervoor uiterlijk 8 uur vóór de vergadering contact op met de griffier. Meer informatie kunt u hier vinden
6. Mondelinge vragenronde raadsleden
7. Lijst Ingekomen Stukken
   Lijst ingekomen stukken - Lijst IS (pdf)
8. Benoemen nieuw lid van de jury van de Vlaardingse horecaprijs De Vlaardingse UITblinker
9. Herbenoeming bestuurslid Stichting Wijzer
   Raadsvoorstel - Bijlage (pdf)

STADSREGIO ROTTERDAM

Gemeente Vlaardingen

A1

Aan:
1. de gemeenteraden
2. de colleges van burgemeester en wethouders van de gemeenten in het werkgebied van de stadsregio Rotterdam

Afschrift aan: de portefeuillehouders financiën van deze gemeenten

datum • juni 2008
ons kenmerk • 32460
steller • E. Visser
telefoon • (010) 4173521
uw kenmerk •
betreft • Jaarverantwoording 2007

Geachte raadsleden, geacht college,

Het doet ons genoegen u het jaarverslag en de jaarrekening 2007, zoals vastgesteld door de Regioraad op 21 mei 2008, aan te kunnen bieden.

Hoogachtend,

het dagelijks bestuur van de stadsregio Rotterdam,

mr M.J. Jansen, secretaris.

mr F.W. Opstelten, voorzitter.

Bijlagen:
1. Jaarverslag 2007
2. Jaarrekening 2007

Meent 106 • Postbus 21051 • 3001 AB Rotterdam • T 010 417 23 89 • F 010 404 73 47 • E info@sr.rotterdam.nl • I www.stadsregio.info

RAAD 18-09-2008

Figure 3. De eerste en één van de vervolgpagina's van een voor de iliad gegenereerde PDF

`puntnr_line.tinc`
Dit is het template voor elk van de aparte agendapunten. `#BODY#` bevat de verklarende tekst bij dit punt, `#DOCS#` bevat de lijst van bijbehorende stukken. Deze laatste is zelf weer een programmatisch opgebouwde lijst omdat er meer dan één vergaderstuk per agendapunt kan zijn.

```
\startpunt[Nummer={#NR#},Titel={#PUNT#},Aard={#AARD#}]
\startpuntbody
#BODY#
\stoppuntbody
\startpuntdocs
#DOCS#
\stoppuntdocs
\stoppunt
```

`puntdoc_line.tinc`
Dit is het eerste van drie mogelijke templates voor een vergaderstuk. Deze is voor vergaderstukken (d.w.z. geëxporteerde PDF bestanden) die zullen worden meegenomen als appendices in de te genereren PDF.

```
\agendadocument[#ICON#]{#LINK#}{#LABEL#}
```

`puntnodoc_line.tinc`
Deze template wordt gebruikt voor vergaderstukken die eigenlijk zouden moeten worden meegenomen als appendices (omdat het PDFs zijn), maar die op grond van configuratie parameters te groot zijn bevonden voor daadwerkelijk gebruik. Er wordt daarom een aparte macro gebruikt die een gepaste meldingstekst kan tonen.

```
\agendanodocument[#ICON#]{#LABEL#}
```

`puntdoc_line_noembed.tinc`
Dit is de derde mogelijkheid, deze is bedoeld voor non-PDF vergaderstukken zoals Microsoft Office bestanden en powerpoint presentaties. Omdat bestanden in die formaten niet kunnen worden getoond is er geen hyperlink mogelijk, en daarom is het trefwoord `#LINK#` in dit geval niet aanwezig.

```
\agendadocument[#ICON#]{}{#LABEL#}
```

ConTEXt macros

Zoals hierboven al werd vermeld zijn de gebruikte ConTEXt macros opgesplitst in twee verschillende bestanden.

Het eerste bestand heeft de naam `agenda-macros-00.tex`, en dit bestand wordt ongewijzigd gebruikt door alle klanten en alle PDF layouts. Het bevat een generieke implementatie van de macros die we eerder zagen in de template bestanden. Deze macros zorgen alleen voor de infrastructuur en doen zelf geen vormgeving. Voor de vormgeving zijn er aanroepen van `\directsetup`.

Typerend voor de inhoud van dit bestand zijn macro definities zoals deze:

```
\def\dostartagenda[#1]%
  {\getparameters
      [Agenda]
      [Gremium=,Datum=,Datumkort=,
       Categorie=,Aanvang=,locatie=,
       Aanhef=,Titel=,Koptitel=,
       Voorzitter=,
       #1]%
   \pagereference[firstpage]
   \directsetup{agenda:start}}

\def\stopagenda
  {\directsetup{agenda:stop}}
```

en deze:

```
\def\agendadocument[#1]#2#3%
  {\doifnotempty
     {#2}
     {\doglobal \appendtoks \addimage{#2}{#3}\to \everyendagenda }%
   \def\DocumentType{#1}%
   \def\DocumentFile{#2}%
   \def\DocumentBody{#3}%
   \pagereference[#2-referer]
   \directsetup{agenda:document}}
```

De macro `\addimage` is de interessantste macro in dit bestand. Hij krijgt als argument de naam van een geëxporteerd PDF bestand door, en zorgt ervoor dat zo'n PDF pagina voor pagina wordt ingelezen via `\externalfigure`. In een wat versimpelde vorm ziet die er als volgt uit:

```
\unexpanded\def\addimage#1#2{%
  \pagereference[#1]
  \xdef\previouspdf{\currentpdf}%
  \gdef\currentpdf{#1}%i
  \getfiguredimensions[#1.pdf]%
  \imgcount=\noffigurepages
  \dorecurse
```

```
      {\the\imgcount}
      {\externalfigure
          [#1.pdf]
          [page=\recurselevel,
           factor=max,
           size=cropbox]%
       \page}%
    \pagereference[#1-last]
}
```

In de appendices van de gegenereerde PDF (zie de figuur) is er een extra interactie–regel onderaan de pagina met daarop drie buttons die springen naar de eerste pagina van de huidige appendix, de eerste pagina van de volgende appendix, en naar de referentie naar deze appendix in de agenda zelf. Deze hyperlinks gebruiken de waarden van `\currentpdf` en `\previouspdf`.

De gevraagde `setups` en de algemene layout definities staan in `agenda-macros-iliad.tex`. Dit bestand kan specifiek gedefinieerd zijn voor een bepaalde klant, of er kan een generieke vorm gebruikt worden: er is een default bestand voor één voor elk van de voorgedefinieerde PDF layouts.

De PDF layouts voor de e-reader zijn verschillend van de layouts voor een PC scherm of printer, maar de meeste verschillen zijn voor de hand liggend. Uiteraard is er een eigen (kleiner) papierformaat. De ruimte op een e-reader is schaars, dus die moet zo goed mogelijk gebruikt worden, dus er worden heel kleine marges gedefinieerd. PDF object compressie wordt uitgeschakeld, omdat de hardware van de e-readers erg licht is in vergelijking met een PC. Het Kleuren–subsysteem van ConTEXt wordt aangezet, maar in grijswaarden. Et cetera.

Het meest ingrijpende verschil is dat de vergaderstukken die bij `normaal' gebruik aparte bestanden zouden blijven hier worden ingebed in het hoofdbestand. Dit maakt het overzetten van de bestanden naar de e-reader eenvoudiger, maar belangrijker is dat dit de gebruikersvriendelijkheid van het resultaat verbetert: Externe PDF hyperlinks worden danwel niet ondersteund (iLiad) of zijn erg langzaam (DR1000).

Een greep uit de inhoud van `agenda-macros-iliad.tex`:

```
\definepapersize [iliad] [width=124mm,height=152mm]

\setuppapersize [iliad] [iliad]

\enableregime[utf8]

\pdfminorversion = 4

\setuplayout[height=14.5cm,
             footer=12pt,
             footerdistance=6pt,
             width=11cm,
             topspace=12pt,
             header=0pt,
             backspace=24pt,
             leftmargin=12pt,
             rightmargin=12pt]

\setupcolors[state=start,conversion=yes,
             reduction=yes,rgb=no,cmyk=no]

\definecolor[papercolor][r=1,b=1,g=1]
```

```
...

\setupbackgrounds[page][state=repeat,
                       background=color,
                       backgroundcolor=papercolor]

...

\startsetups agenda:start
  \blank
  \setupfootertexts[\dofooteragenda]
  \setupfooter[state=high]
  \AgendaGremium
  \blank
  \starttabulate[|l|p|]
  \NC Datum:   \NC \ss\AgendaDatum\NC \NR
  \NC Aanvang: \NC \ss\AgendaAanvang\NC\NR
  \NC Locatie: \NC \ss\AgendaLocatie \NC\NR
  \stoptabulate
  \blank
\stopsetups

\startsetups agenda:stop
  \page
  \the\everyendagenda
  \everyendagenda={}
\stopsetups

\startsetups punten:start
  \startitemize[width=24pt]
\stopsetups

\startsetups punten:stop
  \stopitemize
\stopsetups

....

\startsetups agenda:nodocument
  {\DocumentBody {\tfx bestand te groot voor inclusie}\par }%
\stopsetups
```

Tenslotte

De generatie van PDF bestanden is een klein maar belangrijk onderdeel van NotuDoc. We hebben gekozen voor gebruik van TEX vanwege de hoge kwaliteit van de uitvoer en specifiek voor ConTEXt vanwege het gemak waarmee de vormgeving gescheiden kan worden van de gegevens.

Taco Hoekwater
Elvenkind BV
taco@elvenkind.com