Conversie van BibTeX naar HTML, TeX en LaTeX

Erik Frambach

Samenvatting

BibTeX-bibliografieën kunnen met behulp van een Perl-script vertaald worden naar HTML. Een variant op dat script levert een compleet LaTeX-bestand. Nog een andere variant levert een plain TeX-bestand dat naar believen bijgeschaafd kan worden. Als gereedschap is enkel Perl en (La)TeX nodig.

Keywords: bibTeX, Perl, HTML, conversie

1 Aanleiding

Ooit is door de MAPS-redactie een begin gemaakt met het opzetten en bijhouden van een MAPS-bibliografie waarin gegevens van alle artikelen die in de MAPS zijn gepubliceerd zijn opgenomen. Helaas is dat mooie initiatief verwaterd, zodat er slechts een schamele en verre van complete versie beschikbaar was. Tijd dus voor een grote inhaalslag waarin alle achterstallige onderhoud wordt uitgevoerd. In die slag zijn meteen alle artikelen uit alle MAPSen als afzonderlijke PDF-bestanden beschikbaar gemaakt (zie ook het artikel over de nieuwe NTG WWW-pagina's elders in deze MAPS).

Uitgaande van de bestaande fragmenten en van alle PDF-versies van de MAPS-artikelen heb ik de MAPS-bibliografie in ruwe vorm gecompleteerd. Vervolgens heb ik de auteurs van artikelen waarin *keywords* en/of *abstract* ontbraken gevraagd die gegevens aan te vullen. Met weinig succes overigens. Daarom nogmaals een oproep: stuur aanvullingen!

Uiteraard moest die MAPS-bibliografie beschikbaar zijn op onze WWW-site, met alle artikelen in PDF die erin genoemd worden. Echter, BibTeX is niet voor iedereen even praktisch. Daarom ben ik gaan zoeken naar mogelijkheden om BibTeX te converteren naar HTML. Echt goed gereedschap heb ik daar niet voor kunnen vinden, dus ben ik eens rond gaan vragen.

Taco Hoekwater stuurde al snel een Perl-script op dat ongeveer deed wat ik wilde. Vervolgens kwam Hans van Mourik met zijn versie. Wybo Dekker heeft toen die twee bewerkt tot één.

Vanuit deze bib2html-convertor was het eenvoudig om met enkele aanpassingen ook plain TeX of LaTeX te genereren.

2 Gereedschap

BibTeX-bestanden zijn van nature strak gestructureerd, zodat een 1-op-1 conversie naar HTML/TeX/LaTeX mogelijk moet zijn.

Als gereedschap kwam al snel 'Perl' naar voren, omdat het uitermate geschikt is voor tekstmanipulatie, en omdat het op alle computer-platforms beschikbaar is. Andere opties hebben we niet eens meer onderzocht. Met 'AWK' zou het ook kunnen, of zelfs met TeX, maar met die laatste zou wel erg lastig en bijzonder traag worden.

3 Conversie naar HTML

HTML is redelijk eenvoudig van opzet zodat conversie niet al te moeilijk moet zijn.

Om te beginnen definiëren we het stramien waarbinnen alles gebeurt:

```
<html>
<body>
anything
</body>
</html>
```

'anything' wordt ingevuld als een 'definition list', die als volgt gestructureerd is:

```
<dl>
  <dt>term</dt>
  <dd>descriptie</dd>
  <dt>nog een term</dt>
  <dd>nog een descriptie</dd>
  etc.
</dl>
```

3.1 Problemen

Het grootste probleem bij conversie naar HTML is het feit dat in HTML geen macro's gedefinieerd kunnen worden. Kortom, alles moet helemaal uitgeschreven worden. Daarmee gaat dus onherroepelijk informatie verloren. Terugconverteren van HTML naar BibTeX is daardoor niet meer zonder meer mogelijk, maar in de regel is dat geen probleem.

Andere problemen zijn meer des TeX's: accenten worden in HTML heel anders geschreven dan in (La)TeX. In HTML zijn bovendien bepaalde accenten die TeX kent niet gedefinieerd, bv. '`\'n`'. Ook kunnen sommige accenten in TeX op verschillende manieren geschreven worden:

'\"{\i}' en '\"\i' en '{\"\i}' zijn equivalent. De convertor moet alle varianten aankunnen.

Speciale tekens zoals '\OE', '\dots', '\l{}' en '\&' vragen ook om een speciale behandeling. De verbatim-omgeving is helemaal een verhaal apart, omdat de eenvoudigste truc die bib2html uithaalt neerkomt op het simpelweg verwijderen van alle '\', '{' en '}'. Dat mag in verbatim-modus natuurlijk niet.

We hebben er bewust voor gekozen mathematische tekens niet te ondersteunen omdat dat te veel problemen geeft. In een bibliografie (zeker die van de MAPS) komen die niet of nauwelijks voor zodat dat geen problemen geeft.

Ook andere constructies moeten vermeden worden in het BibTeX-bestand. Bib2html ondersteunt geen '{\bf ...}', '\textit{...}' en dergelijke. Het zal duidelijk zijn dat TeX-macro's definiëren geheid problemen geeft.

4 Conversie naar LaTeX

Bij de conversie naar LaTeX worden de velden uit het BibTeX-bestand overgenomen (met kleine wijzigingen, bv. '\speak' i.p.v. '\language'). Voor het zetten van die velden worden macro's gedefinieerd, zodat het LaTeX-bestand in één keer gecompileerd kan worden en meteen fraai geformateerde uitvoer levert.

5 Conversie naar TeX

De opzet is hier in principe dezelfde als bij conversie naar LaTeX, maar de macro's voor opmaak zijn 'leeg' gelaten. De uitvoer zal in eerste instantie niet fraai zijn. De TeX-uitvoer is meer bedoeld voor specialistische (database-achtige) bewerkingen zoals Hans Hagen die demonstreert in http://www.ntg.nl/maps/pdf/maps.pdf.

Dat document is gegenereerd met de naar TeX geconverteerde MAPS-bibliografie als invoer. De uitvoer is een interactief document waarin door de bibliografie gewandeld kan worden via verschillende 'linked-lists' of dimensies. Voorbeelden: alle artikelen van één MAPS op een rijtje; alle artikelen van één auteur op een rijtje; alle artikelen met een bepaald *keyword* op een rijtje. Uiteraard zijn er ook verschillende indexen beschikbaar. En dit alles wordt volautomatisch gegenereerd!

6 Voorbeelden

Als demonstratie volgt hier een klein stukje BibTeX-code, met daarna de uitvoer van de conversie naar HTML, plain TeX en LaTeX. De BibTeX-code is geplukt uit de MAPS-bibliografie en is 'verrijkt' met enkele grapjes om te laten zien hoe bib2html met 'moeilijke' gevallen omgaat.

```
@ARTICLE{2-6,
  author = {{Theo de Klerk}},
  names = {klerktde},
  title = {{Boeken over \TeX}},
  year = {{1989}},
  language = {{Dutch}},
  journal = {{MAPS}},
  volume = {{2}},
  pages = {{19-20}},
  size = {{71}},
  keywords = {{boekbespreking \& \OE{}uvre,
    \verb+\def\TeX{difficult}+}},
  abstract = {{Bespreki\'ng 'Einf\"uhrung in
    \TeX\dots' (Norbert Schwartz); '\TeX\ f\"ur
    Fortgeschrittene' (Wolfgang Appe\l{}t);\\
    '\LaTeX\ eine Einf\"uhrung' (Helmut Kopka);
    'Kompaktf\"uhrer \LaTeX' (Reinhard Wonneberger)}}
}
```

Alle velden zijn voorzien van dubbele accolades. De reden hiervoor is dat ik in dit geval niet wil dat BibTeX zelf met hoofdletters en kleine letters gaat stoeien. Meer gebruikelijk is het om in zo'n geval bv. te schrijven 'Dit {I}s {\TeX}'. Dan zal de hoofdletter 'I' beslist een hoofdletter blijven (ongeacht de voorschriften in de BibTeX-style) en zal de TeX-compiler niet gaan klagen dat ie het commando '\tex' niet kent. Omdat dit een erg onhandige manier van schrijven vereist heb ik ervoor gekozen alles af te schermen, maar toch BibTeX-compatibel te blijven.

Het veld 'names' is uitsluitend bedoeld als sorteersleutel en wordt in de regel niet afgedrukt.

6.1 HTML

(De formattering is iets aangepast om de code hier netjes te kunnen afdrukken. Regelovergangen hebben echter voor HTML dezelfde betekenis als spaties, dus hindert dat niet.)

```
<html>
<body>
<dl>
<dt class="ARTICLE" id="2-6"><strong>Theo
  de Klerk</strong></dt>
<dd>
  <em>Boeken over TeX</em>, Dutch, MAPS
  <strong>2</strong>, 1989, pp. 19-20<br>
  <strong>keywords:</strong> boekbespreking &amp;
  Oeuvre, <tt>\def\TeX{difficult}</tt><br>
  <strong>abstract:</strong> Bespreking
  'Einf&uuml;hrung in TeX...' (Norbert Schwartz);
  'TeX f&uuml;r Fortgeschrittene' (Wolfgang Appelt);<br>
  'LaTeX eine Einf&uuml;hrung' (Helmut Kopka);
  'Kompaktf&uuml;hrer LaTeX' (Reinhard Wonneberger)<br>
<a href="pdf/2_6.pdf">2-6 in PDF</a> (71 KB)
</dd>
</dl>
</body>
</html>
```

6.2 LaTeX

```
\documentclass[a4paper]{article}

% macros:
```

```
\newif\ifseteditor
\def\editstring{}
\def\startentry#1{%
  \def\id{#1}
  \seteditortrue\par\noindent\hangindent 10mm}
\def\author#1{%
  \seteditorfalse#1\\}
\def\names#1{}
\def\editor#1{%
  \ifseteditor
    #1 (editor)\
  \else
    \def\editstring{#1 (editor)}
  \fi}
\def\title{%
  \begingroup\bf}
\def\speak#1{%
  \endgroup
  \\ #1,}
\def\journal#1{%
  {\sl#1}}
\def\volume#1{%
  {\bf#1},}
\def\series#1{%
  #1,}
\def\year#1{%
  #1,}
\def\booktitle#1{%
 in:~{\sl#1},}
\def\publisher#1{%
  \editstring, #1,}
\def\pages{%
  pp.~}
\def\keywords{
  \\ {\bf keywords: }}
\def\abstract{%
  \\ {\bf abstract: }}
\def\size#1{%
  (\id.pdf: #1 KB)}
\def\stopentry{}

% end of macros

\begin{document}

\startentry {2-6}
\author {Theo de Klerk}
\names {klerktde}
\title {Boeken over \TeX}
\speak {Dutch}
\journal {MAPS}
\volume {2}
\year {1989}
\pages {19-20}
\keywords {boekbespreking \& \OE{}uvre,
  \verb+\def\TeX{difficult}+
\abstract {Bespreki\'ng `Einf\"uhrung in \TeX\dots'
  (Norbert Schwartz); `\TeX\ f\"ur Fortgeschrittene'
  (Wolfgang Appe\l{}t);\\
  `\LaTeX\ eine Einf\"uhrung' (Helmut Kopka);
  `Kompaktf\"uhrer \LaTeX' (Reinhard Wonneberger)}
\size {71}
\stopentry
\end{document}
```

6.3 plain TeX

```
\def\LaTeX{La\TeX}
\def\\{\hfil\break}

\def\startentry{}
\def\author{}
\def\names{}
\def\editor{}
\def\title{}
\def\speak{}
\def\journal{}
\def\volume{}
\def\series{}
\def\year{}
\def\booktitle{}
\def\publisher{}
\def\pages{}
\def\keywords{}
\def\abstract{}
\def\size{}
\def\stopentry{}

\def\type{} % !!! macro for typesetting
            % !!! verbatim text

\startentry {2-6}
\author {Theo de Klerk}
\names {klerktde}
\title {Boeken over \TeX}
\speak {Dutch}
\journal {MAPS}
\volume {2}
\year {1989}
\pages {19-20}
\keywords {boekbespreking \& \OE{}uvre,
  \type{\def\TeX{difficult}}
\abstract {Bespreki\'ng `Einf\"uhrung in \TeX\dots'
  (Norbert Schwartz); `\TeX\ f\"ur Fortgeschrittene'
  (Wolfgang Appe\l{}t);\\
  `\LaTeX\ eine Einf\"uhrung' (Helmut Kopka);
  `Kompaktf\"uhrer \LaTeX' (Reinhard Wonneberger)}
\size {71}
\stopentry
\bye
```

7 Conclusie

BibTeX-documenten zijn strak gestructureerd. Dit maakt het relatief eenvoudig om ze te converteren naar andere systemen die gestructureerd werken mogelijk maken. LaTeX ligt daarbij voor de hand, maar ook plain TeX doet het prima en geeft uiteindelijk zelfs meer flexibiliteit, zij het dat daarvoor navenant meer programmeerwerk nodig is.

8 Beschikbaarheid

De convertors zullen binnenkort op CTAN gezet worden.