Rob Schrauwen
DTD Development & Maintenance, Elsevier Science
Sara Burgerhartstraat 25
1055 KV Amsterdam
r.schrauwen@elsevier.com

Online publishing

XML workflow in de praktijk

Sinds zijn promotie in de wiskunde in 1991 werkt Rob Schrauwen bij Elsevier Science. Begonnen in de bureauredactie van wiskundetijdschriften, is hij nu verantwoordelijk voor de computersystemen en de op XML gebaseerde 'content-standaarden' die voor de productie van tijdschriften en boeken nodig zijn.

Sinds 1995 worden alle tijdschriftartikelen die bij Elsevier Science gepubliceerd worden, opgeslagen in SGML. SGML, wijd verspreid in de uitgeverswereld, staat voor *standard generalized markup language* en is herkenbaar aan *tags* zoals <author>, die structuur aan het document verlenen. Welke tags in een document zijn toegestaan of verplicht zijn, in welke volgorde zij mogen voorkomen, hoe zij samenhangen, en ook welke symbolen gebruikt mogen worden, wordt in een DTD, *document type definition*, vastgelegd. Een voorbeeld van zo'n DTD is de definitie van HTML, de taal waarin webpagina's zijn geschreven. Elk SGML document moet tegen de DTD gevalideerd worden, zodat software het document gegarandeerd kan verwerken. Een DTD kan het bijvoorbeeld onmogelijk maken om artikelen zonder titel of zonder auteursnamen te publiceren, wat af en toe consternatie geeft bij auteurs die dat willen.

Elsevier Science, net als de meeste andere wetenschappelijke uitgevers, heeft sinds 1992 zijn eigen, publiek beschikbare DTD ontwikkeld [5] voor wetenschappelijke artikelen. Belangrijk is dat artikelen 'media neutraal' worden opgeslagen, met zoveel mogelijk informatie over de structuur van het artikel, maar zo weinig mogelijk indicaties over de presentatie. Het is bijvoorbeeld niet mogelijk om een paginagrens aan te geven. Dit heeft ook geen zin, want dezelfde bron wordt gebruikt voor de diverse formaten waarop een en hetzelfde artikel gepubliceerd wordt: op papier (of, equivalent, in PDF-formaat), online (geconverteerd naar HTML) of bijvoorbeeld op palmtop computers — populair in de medische wereld. Zo wordt ook de referentielijst volledig presentatie-neutraal gestructureerd op een manier die wel met de structuur van BibTeX valt te vergelijken. Zo'n gedetailleerde structuur maakt het bijvoorbeeld mogelijk om dynamisch links te creëren naar de artikelen van 152 uitgevers via CrossRef [1].

Inmiddels is XML in zwang geraakt en de nieuwste standaarden voor opslag van documenten zijn op XML gebaseerd. XML, *extensible markup language*, is een variant van SGML waarin onder andere obscure mogelijkheden van SGML niet zijn toegelaten. Wat voor velen in de Internetwereld nieuw was, namelijk de mogelijkheid om presentatieonafhankelijk documenten te stuctureren, was in de uitgeefwereld al gemeengoed via SGML. Door het grootschalige gebruik ontwikkelt XML zich snel en komen er vele standaarden tot stand. De recente Elsevier XML DTD 5.0, onderdeel van een familie XML DTDs voor tijdschiften en boeken, maakt gebruik van MathML [4] in plaats van het oude zelfontwikkelde model voor formules. Het gebruik van Unicode [7] — dat beoogt een uniek nummer te geven aan ieder bestaand letterteken of symbool — maakt het mogelijk om talloze symbolen standaard te vertonen in webbrowsers terwijl tot nu toe gebruik gemaakt moest worden van kleine plaatjes. Voor Elsevier Science is het belangrijk deel te nemen aan de ontwikkeling van dergelijke standaarden. Samen met grote society publishers zoals de AMS is in het STIX project [6] gezorgd dat een grote verzameling wetenschappelijke symbolen in Unicode beschikbaar is en dat een vrij beschikbare font set gemaakt wordt met deze symbolen.

Het begin van dit artikel zou er in de Elsevier Science Journal Article DTD 5.0 ongeveer uitzien als in figuur 1 is weergegeven.

Workflow

Dagelijks komen 900 artikelen binnen bij de productieafdelingen van Elsevier. Elk artikel is al geaccepteerd door de wetenschappelijke redactie. Gegevens zoals naam en adres van de corresponderende auteur en de titel worden in het workflow-systeem ingevoerd, en het artikel krijgt een unieke identificatie. Daarna gaat het artikel naar een van de toeleveranciers om in SGML of XML te worden omgezet en in de tijd-

schriftstijl te worden gegoten. Dit duurt vijf à tien dagen, waarna een drukproef in PDF naar de auteur wordt ge-emaild. Deze PDF-file is rechtstreeks gegenereerd vanuit de SGML/XML-file, die de bron is van alle producten (het *SGML first* principe). Tegelijk worden de SGML/XML-file met bijbehorende figuren en de PDF-file naar Elsevier teruggestuurd. Afhankelijk van de tijdschrift-policy plaatsen sommige tijdschriften de ongecorrigeerde proef meteen online op de web-platformen waarop die tijdschriften verschijnen. Hiervan is ScienceDirect de belangrijkste, maar er zijn ook grote bibliotheken die de artikelen rechtstreeks in SGML/XML willen ontvangen (zie ook [3]). Indien het preprint-nummer bekend is, kan deze versie al linken aan de preprint zoals verschenen op arXiv (in een later stadium linkt arXiv ook terug) en bovendien worden de bibliografische referenties via CrossRef [1] gekoppeld aan de artikelen of de abstracts waarnaar ze verwijzen. Deze functionaliteit staat en valt met de beschikbaarheid van alle benodigde gegevens in de brondocumenten van de auteurs.

Wanneer de auteurs hun correcties hebben aangebracht en teruggestuurd, wordt hun informatie direct naar de toeleverancier doorgestuurd, waarna binnen een dag of vijf de gecorrigeerde versie wordt aangeleverd aan Elsevier, zodat deze versie online beschikbaar gemaakt kan worden. In feite is dit al de definitieve versie, waarvan alleen het volume-nummer en de paginacijfers nog onbekend zijn. Naar deze versie kan al verwezen worden door middel van de *DOI*, de digital object identifier [2], al gebeurt dat in de praktijk helaas nog weinig.

Een aantal weken later — in wiskundige tijdschriften soms heel wat later — wordt het artikel opgenomen in een nummer van het tijdschrift. De paginering is nu bekend en de toeleverancier levert opnieuw de PDF-files aan, die nu een exacte representatie zijn van de pagina's zoals ze in het issue zullen verschijnen. Tegelijk stuurt de toeleverancier uit dezelfde bron gegenereerde PDF-files met een behoorlijk hogere resolutie naar de drukker.

Van de artikelen in de ruim 1600 tijdschriften die Elsevier Science produceert, wordt het merendeel volledig in SGML/XML gestructureerd. Van ongeveer vijftig tijdschriften worden echter alleen de 'heads' (titel, auteursnamen, affiliaties) en 'tails' (referenties) gestructureerd — het gehele artikel is natuurlijk wel in PDF-formaat beschikbaar. Een deel van de tijdschriften uit de wiskunde en theoretische informatica valt hieronder vanwege de moeilijkheden om bijvoorbeeld commutatieve diagrammen, natuurlijke-deductiediagrammen of zelfgemaakte symbolen weer te geven. In LaTeX is het eenvoudig zelf symbolen te creëren en van deze mogelijkheid wordt veelvuldig gebruik gemaakt, maar om deze op alle mogelijke platformen getrouw weer te geven in diverse groottes gaat met vele problemen gepaard.

```
<head>
  <dochead>Online publishing</dochead>
  <article-title>XML workflow in de praktijk
  </article-title>
  <author-group>
    <author>
      <given-name>Rob</given-name>
      <surname>Schrauwen</surname>
      <e-address>r.schrauwen@elsevier.com</e-address>
    </author>
    <affiliation>
      <textfn>DTD Development &amp;amp; Maintenance,
        Elsevier Science, ...
      </textfn>
    </affiliation>
  </author-group>
</head>
<body>
  <sections>
    <section>
      <para>Sinds zijn promotie...
```

Figuur 1 Fragment van een document gestructureerd in XML

LaTeX en XML

Het merendeel van de artikelen dat binnenkomt, is geschreven in Microsoft Word. Zulke artikelen worden probleemloos naar XML omgezet. In de wiskunde is LaTeX uiteraard veel gebruikelijker. Afgezien van de problemen met speciale symbolen, is het goed mogelijk een LaTeX-file automatisch naar XML om te zetten. Aangezien het gaat om de structuur, is het irrelevant voor de auteur om het artikel zo goed mogelijk op het eindproduct te laten lijken; het is des te belangrijker om gestructureerde LaTeX-documenten te maken. Ingewikkelde macro's maken de conversie moeilijker.

Vooralsnog zijn SGML en XML typische opslagformaten, ongeschikt om rechtstreeks artikelen in te schrijven. 'XML support', wat sommige tools bieden, zegt weinig: het is onwaarschijnlijk dat er ondersteuning is voor de juiste door de uitgever gebruikte DTD. Voor wiskundige artikelen blijven LaTeX en BibTeX de aangewezen formaten, al wordt verwacht dat aangeleverde *MathML*-formules in de toekomst gebruikt kunnen worden in het productieproces.

Referenties

1 CrossRef, The Central Source for Reference Linking, http://www.crossref.org.

2 DOI, The Digital Object Identifier System, http://www.doi.org.

3 E-Journal Archival DTD Feasibility Study, prepared for the Harvard University E-Journal Archiving Project, http://www.diglib.org/preserve/hadtdfs.pdf.

4 MathML, The Mathematical Markup Language, http://www.w3.org/Math.

5 Simon Pepping and Rob Schrauwen, Tag by Tag, Documentation of the Elsevier Science DTDs 4.1–4.3, http://www.elsevier.com/locate/sgml.

6 The STIX Project, http://www.ams.org/STIX.

7 Unicode, http://www.unicode.org.