TEX en SGML bij Elsevier Science

Simon Pepping

Elsevier Science, Physics & Materials Science,
Sara Burgerhartstraat 25, Postbus 103, 1000 AC Amsterdam
s.pepping@elsevier.nl

1 LATEX-projekt

Op de TEX90-conferentie [1] kondigde Nico Poppelier aan dat *Elsevier Science* ging beginnen met het aannemen van LATEX-files van auteurs; deze zouden worden verwerkt tot de in het wetenschappelijke tijdschrift gepubliceerde tekst zonder dat het artikel opnieuw gezet werd. Er waren op dat moment documentstijlen voor 4 tijdschriften geschreven, waarmee men de mogelijkheden en moeilijkheden van zo'n onderneming wilde verkennen. De eigenlijke publikatie werd voorlopig nog via konventioneel zetwerk vervaardigd. Dit bleek de voorzichtige start te zijn van een succesvol projekt.

Twee jaar lang werd er experimenteel gewerkt. Selektief vroegen we auteurs om ons hun LATEX-file toe te sturen. Af en toe deden zich onverwachte problemen voor. Zo kwam ik bij het bewerken van een overzichtsartikel van 225 gedrukte pagina's voor de volgende verrassing te staan. Urenlang had ik ieder hoofdstuk afzonderlijk geedit en door TEX gehaald. Eindelijk kon ik het artikel in zijn geheel draaien; verwachtingsvol zag ik de paginacijfers van de afgewerkte bladzijden over het scherm trekken. Plotseling verscheen daar de mededeling dat ik de TEX-capaciteit overschreden had: ik had teveel strings gebruikt. Een narigheidje van LATEX op een PC was de boosdoener: LATEX moet een publikatie als één geheel beschouwen vanwege de mogelijkheid van kruisverwijzingen; bij een grote publikatie kan dat de beperkte mogelijkheden van het op de PC draaiende TEX-programma te boven gaan. In de loop van dit jaar heb ik aan de hand van deze ervaring nog een vraag op TEX-NL kunnen beantwoorden; het maakte de vraagsteller niet gelukkiger. Zo hebben wij allen onze portie verrassingen gehad die ons vertwijfeld deden afvragen waar we aan begonnen waren. Gelukkig bleken ze nooit fataal te zijn; steeds hebben we de problemen met kennis, inzet en kreativiteit op kunnen lossen.

Na een experimentele periode van 2 jaar volgde eind 1992 de openbare aankondiging dat Elsevier Science voor vele tijdschriften auteurs uitnodigde om hun LATEX-files in te sturen; daarbij kregen ze de garantie dat de ingezonden file zou worden gebruikt voor de publikatie als die gekodeerd was in standaard LATEX en er geen TEX-fouten in zouden zitten. Dit bleek te voorzien in een behoefte bij onze auteurs. In de loop van 1993 steeg de stroom ingestuurde compuscripten (manuscripten in electronische vorm) gestaag en dit jaar heeft die toename zich voortgezet. In 1993 hebben we ruim 30 000 gedrukte pagina's in 38 wetenschappelijke tijdschriften gepubliceerd uitgaande van door auteurs ingestuurde LATEX-compuscripten. Op dit moment werken 45 bureauredakteuren verspreid over 4 uitgeefafdelingen geheel of gedeeltelijk met LATEX-compuscripten.

We verzorgen zelf de training van onze bureauredakteuren via eigen kursussen, aangevuld door coaching op de werkplek door kollega's met ervaring. Technisch wordt het projekt ondersteund door onze afdeling Informatietechnologie, die zorgt voor de installatie en het onderhoud van het TEX-systeem en voor het schrijven en bijhouden van gevarieerde software die ons helpt om de stroom compuscripten zo efficiënt en foutloos mogelijk te verwerken. Bovendien hebben op de diverse afdelingen enkele bureauredakteuren een diepgaande kennis van LATEX opgebouwd, zodat hun kollega's op hen kunnen terugvallen bij de velerlei LATEX-problemen die de verschillende artikelen kunnen oproepen. De coördinatie van het projekt berust eveneens bij de afdeling Informatietechnologie.

In de loop van de tijd hebben we de volgende procedure ontwikkeld. Slechts enkele tijdschriften, zoals het tijdschrift *Nuclear Physics*, handelen het refereeingproces via netwerkkommunikatie af. Daardoor ontvangen we nog bijna alle voor publikatie aanvaarde artikelen op papier. Als het lettertype (Computer Modern) of de stijl van het manuscript suggereren dat het in TEX of LATEX is geschreven, vragen we de auteur of hij of zij ons de file wil opsturen. Gelukkig is het een wijd verbreide gewoonte van auteurs geworden om het e-mailadres op het artikel te vermelden; dit maakt ook voor ons de kommunikatie efficiënt en snel.

De meeste auteurs reageren positief, willen het zelfs graag. Ze hebben veel werk gestoken in de juiste presentatie van hun manuscript, niet alleen naar de editors van het tijdschrift toe maar ook naar kollega's. In deze Internettijden is een wetenschappelijk tijdschrift namelijk nooit meer het forum waar een artikel wereldkundig gemaakt wordt. In preprint-vorm is het allang rondgestuurd naar een grotere of kleinere kring kollega's, ter kennisneming en voor reakties. Veel auteurs vinden het prettig dat al dat werk resultaat heeft tot in de gedrukte versie van het artikel. Bovendien geeft het hen het gevoel dat ze meer greep hebben op het gedrukte resultaat en niet meer afhankelijk zijn van bureauredakteuren of proeflezers.

Net als vroeger wordt de tekst na ontvangst van het artikel door een bureauredakteur doorgelezen en bewerkt. Oor-

spronkelijk deden we dat met dezelfde filosofie en normen als vroeger: de auteur levert alleen de tekst; layout, spellingsnormen en allerlei andere kleine zaken worden door ons bepaald in overeenstemming met de stijlrichtlijnen van het tijdschrift. Geleidelijk aan zijn we tot de konklusie gekomen dat met het veranderen van de techniek ook deze normen moeten veranderen: we zetten het artikel niet meer helemaal opnieuw; dan is het ook zinloos om alle keuzen die een auteur al gemaakt heeft nog eens over te doen. Bovendien, nu auteurs een betere vormgeving tot hun beschikking hebben dan de typemachine van de sekretaresse en hun eigen vulpen voor de formules, besteden ze daar ook veel meer aandacht aan dan vroeger. We laten de teksten dus veel meer zoals ze zijn en grijpen alleen daar in waar dat naar ons oordeel beslist nodig is.

De stijl van het tijdschrift wordt in een goed LATEX-artikel wel heel eenvoudig verkregen: in de `\documentstyle`-regel wordt de door de auteur gebruikte documentstijl vervangen door de documentstijl van het tijdschrift.

Na deze werkzaamheden produceert de bureauredakteur een dvi-file en leest deze na op onverwachte effekten die zouden kunnen optreden bij de omzetting van documentstijl en op fouten die bij het editen gemaakt kunnen zijn. Na de nodige korrekties uitgevoerd te hebben levert hij of zij de dvi-file van het artikel bij de zetterij in. Daar wordt een uitdraai op hoge resolutie gemaakt en worden eventueel de figuren gemonteerd (ingeplakt).

Van de ontvangst van de LATEX-file tot de uitvoer op papier wordt de TEX-omgeving dus niet verlaten. Dit garandeert een hoge grafische kwaliteit, vermijdt de kans op konversiefouten, vooral in de formules, en bespaart veel werk;

Elsevier Science heeft nooit Computer Modern als lettertype voor haar tijdschriften geaccepteerd. Hoewel het in het kader van TEX een zeer succesvol lettertype is geweest, hebben wij altijd de voorkeur gegeven aan het in de grafische wereld meer gebruikelijke Times Roman lettertype. Daarom heeft ons LATEX-projekt van het begin af aan gedraaid met de zetletter van onze fotozetter; op dit moment is dat Adobe Times. Een van de medewerkers van onze zetterij, een grafikus, heeft de fonts geschikt gemaakt voor gebruik door TEX door de font dimen parameters aan te vullen en te korrigeren waar nodig. Een dergelijke onderneming was al eerder voltooid door de American Mathematical Society (AMS), die daarover uitvoerig gerapporteerd heeft op de TUG bijeenkomst in augustus 1990 [2]. Mede naar aanleiding van de daarbij opgedane ervaringen beschreef Knuth in datzelfde jaar [3] het idee van virtuele fonts. Net als in andere implementaties van PostScript-fonts, zoals PSNFSS, is ook voor ons het mechanisme van virtuele fonts een essentieel hulpmiddel geweest om de Adobe fonts op een beheersbare manier in TEX in te brengen.

Een ander aspekt van de filosofie die reeds in [1] naar voren is gebracht is het volgende: wij willen de auteur niet lastig vallen met speciale instrukties voor het voorbereiden van compuscripten voor Elsevier-tijdschriften. Dit zou de flexibiliteit van de refereeingprocedure aantasten: als de auteur in de loop van die beoordelingsprocedure bij een ander tijdschrift van een andere uitgever terecht zou komen, zou hij zijn LATEX-file aan moeten passen aan de specifieke instrukties van die uitgever. Daarom hebben wij nooit van auteurs geeist dat zij hun artikelen in een of ander voor Elsevier specifiek formaat zouden koderen. Wij voeren een breed acceptatiebeleid: artikelen in de standaard documentstijl `article` of zelfs in de stijl `revtex` van onze medeuitgevers van de American Physical Society (APS) worden door ons verwelkomd. Daardoor zijn wij zelfs pas vrij laat met onze eigen documentstijl `elsart` naar buiten gekomen.

Met deze brede acceptatiepolitiek staan we vrijwel alleen. De meeste van onze konkurrenten vragen van hun auteurs dat ze de artikelen intypen met gebruikmaking van hun specifieke macropakket of zelfs van macro's die specifiek voor één tijdschrift zijn geschreven. Het is een onmiskenbaar feit dat dit beleid extra eisen aan ons als uitgever stelt: ieder artikel kan komen met zijn eigen macro's die op hun eigen wijze met onze gewenste tijdschriftstijl in konflikt kunnen komen. Dat betekent dat we de artikelen niet op een standaardmanier aan kunnen pakken en dat onze medewerkers over meer kennis van LATEX moeten beschikken om de situatie op te vangen. (Er zijn natuurlijk grenzen aan wat we kunnen accepteren; de aanwezigheid van teveel eigen programmeerwerk in een artikel is een reden voor behandeling van het artikel op de konventionele manier.) Aan de andere kant denk ik dat deze brede benadering ook bijgedragen heeft aan de bereidheid van auteurs om ons hun LATEX-files toe te sturen en daarmee tot het succes van ons LATEX-projekt.

Een strenge grens aan ons acceptatiebeleid is dat we ons in principe alleen tot LATEX-artikelen beperken; daarmee wijzen we Plain-TEX files, `phyzzx`-TEX files en `harvmac` files af. Waarom hebben we voor LATEX gekozen?

- LATEX werkt met documentstijlen. Dit geeft de uitgever de gelegenheid om in principe het artikel in de stijl van zijn tijdschrift af te drukken door alleen maar een andere documentstijl voor het artikel aan te geven.
- Zoals Lamport in zijn boek [4, p. 6] schrijft moet het de primaire funktie zijn van bijna alle LATEX-opdrachten die men intypt om de logische struktuur van een dokument te beschrijven. Daarmee is een LATEX-dokument geschikt om naar een SGML-dokument omgezet te worden. Verderop zal ik aangeven waarom wij dat belangrijk vinden.

Terugkijkend over de laatste 4 jaar mogen we konstateren dat het voorzichtige begin van ons LATEX-projekt met 4 tijdschriften, parallel aan de konventionele produktie, succesvol is uitgegroeid tot een dagelijks gebruikte produktiemethode.

2 Andere compuscripten

Het succes van ons LATEX-projekt is deels te danken aan het feit dat LATEX bij onze auteurs erg populair is: bijna iedereen gebruikt het. Dat is echter alleen zo binnen bepaalde vakgebieden in de academische wereld: natuurkunde, wis-

kunde, kunstmatige intelligentie. Hoe platter de tekst in een bepaald vakgebied, d.w.z. hoe minder formules er in de artikelen en rapporten voorkomen, hoe minder populair LATEX is. Dat is b.v. duidelijk te zien in het vakgebied materiaalkunde: hoewel dit vak dicht tegen de natuurkunde aan ligt, is het meer beschrijvend en minder wiskundig. Het gevolg is dat we hier veel meer artikelen krijgen die met tekstverwerkers als WordPerfect geschreven zijn.

In ons biomedische uitgeefprogramma is men al in de jaren tachtig gestart met het opvragen en verwerken van zulke compuscripten. Volgens dezelfde filosofie als eerder geschetst, willen we de auteurs zo weinig mogelijk beperkingen opleggen bij het schrijven van hun artikel. Daarom accepteren we compuscripten uit een zo breed mogelijke reeks van tekstverwerkers. In de praktijk moeten we een grens leggen bij tekstverwerkingsprogramma's die te weinig gebruikt worden om er een konversie voor te vervaardigen.

Al die verschillende binnenkomende auteursfiles worden direkt gekonverteerd naar een standaardformaat. Daarvoor is een brede reeks aan konversieprogramma's geschreven. Met de zo vervaardigde file doet de bureauredaktie hetzelfde als bij LATEX-files: de tekst wordt gekontroleerd en bewerkt. Het uiteindelijke doel is echter meer dan een geredigeerde file. De file moet geschikt zijn om door een zetter naar zijn zetsysteem vertaald te worden. Dat gaat het beste als de file gestruktureerd is, d.w.z. dat bij de verschillende onderdelen niet alleen staat hoe ze vorm gegeven zijn, maar vooral wat hun funktie in het dokument is. Aan het einde van de redaktieproces wordt daarom de tekst omgezet in een SGML-file volgens de door Elsevier opgestelde document type description. Wanneer het niet mogelijk is een syntaktisch korrekte SGML-file te produceren door inconsistenties in de oorspronkelijke file, wordt dit door het konversieprogramma gesignaleerd en wordt de bureauredakteur gevraagd om korrektie. De resulterende, syntaktisch gekontroleerde SGML-file wordt naar de zetter gestuurd, die het artikel vervolgens op papier zet.

Hier treedt wel een duidelijk verschil tussen TEX en SGML naar voren: TEX is een grafische taal en een TEX-file bevat dus reeds de opmaak op papier. SGML is een documentbeschrijvingstaal; een SGML-document bevat geen enkele opmaakinformatie voor de vormgeving voor welk medium dan ook. Die opmaak wordt door de zetter vervaardigd uitgaande van de beschrijving die de SGML-file biedt. LATEX tracht beide te doen: het document wordt gekodeerd met de grafische taal TEX, maar het macromechanisme daarvan wordt zodanig gebruikt dat de koderingen in de file tevens een beschrijving van het dokument geven in de zin van 'Generalized Markup'.

3 Toekomst – SGML

In de nabije toekomst is ons doel niet meer om uitsluitend de binnengekomen artikelen op papier te zetten en te verspreiden onder de abonnees, wat eeuwenlang de taak van tijdschriftuitgevers is geweest. In het electronische tijdperk willen wij informatie in gestruktureerde vorm klaarzetten voor meervoudig gebruik, zowel in uitgeefprojekten, waarvan het tijdschrift een voorbeeld is, als op verzoek van klanten ('on demand publishing').

In [1] is de basisfilosofie waarmee we dit doel willen bereiken al uiteengezet. Onze fundamentele uitgangspunten en overwegingen zijn sindsdien weinig of niets veranderd. Ik maak die filosofie daarom duidelijk aan de hand van dezelfde figuur die in [1] is gebruikt (Fig. 1).

Zoals Fig. 1 laat zien, staat in de Elsevier-filosofie SGML centraal. Onze redenen daarvoor kunnen in de volgende vier punten worden samengevat:

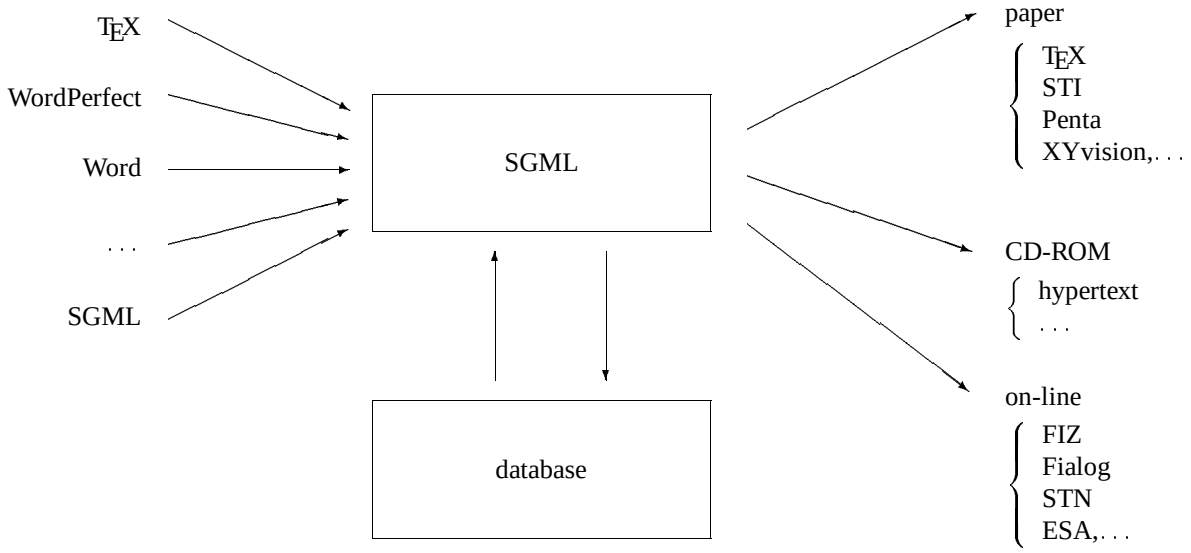
1. SGML stelt ons in staat om de logische struktuur van een dokument te scheiden van de visuele struktuur.
2. SGML stelt ons in staat om compuscripten van verschillende origine na konversie te verwerken in een uniforme omgeving.
3. SGML stelt ons in staat om compuscripten te verwerken onafhankelijk van het uitvoermedium of het uitvoerformaat.
4. SGML stelt ons in staat om de artikelen in een database op te slaan en later in een andere vorm van publikatie, eventueel op een ander medium, opnieuw te gebruiken.

Het zou natuurlijk prachtig in deze filosofie passen als auteurs in staat zouden zijn om ons een SGML-file aan te leveren. Zover is het echter nog niet. SGML-invoer van auteurs zal pas mogelijk zijn als de tekstverwerkingssoftware in staat is achter de schermen een SGML-file te produceren van hetgeen de auteur op voor hem of haar begrijpelijke wijze invoert. De ideale tekstverwerker van de toekomst kombineert dan ook de volgende vaardigheden:

- De invoer is WYSIWYG, zelfs voor formules.
- De uitvoer is in een hoogwaardige grafische taal, b.v. TEX.
- Tevens wordt er van de tekst een SGML-file geproduceerd.

Totdat die ideale tekstverwerker bij onze auteurs gemeengoed zal zijn geworden, is de strategie zoals geschetst in Fig. 1, met een SGML-file als uniform uitgangsprodukt voor de verwerking en opslag, voor LATEX-artikelen niet haalbaar; het is nog steeds erg moeilijk om formules uit b.v. een LATEX-file naar SGML te vertalen en later, na bewerking, weer naar een grafische taal, b.v. TEX, terug te vertalen met behoud van de grafische kwaliteit. Wel streven we ernaar om aan het einde van het trajekt, als de LATEX-file binnen de TEX-omgeving geredigeerd en voor publikatie gereed gemaakt is, een SGML-versie van het artikel te produceren voor opslag in een database en hergebruik in andere publikatievormen. Voor niet-LATEX-compuscripten, in het algemeen met geen of weinig formules, wordt de in Fig. 1 geschetste procedure al grotendeels gerealiseerd.

Een ander belangrijk punt van de toekomstige behandeling van artikelen vormen de illustraties. Tot voor kort vielen ze buiten het kader van de tekstverwerking en moesten ze na afloop met de hand ingeplakt worden. In deze situatie komt snel verandering: integratie van tekst en beeld



Figuur 1: *Konversie naar en van SGML.*

wordt steeds beter mogelijk. De veelheid van programma's waarmee die integratie bereikt wordt maakt het de uitgevers echter niet gemakkelijk. In de LATEX-wereld is een snelle konvergentie te zien naar PostScript. Het probleem van PostScript-files voor de uitgever is dat ze nauwelijks bewerkt kunnen worden. Zelfs kleine aanpassingen zijn niet altijd mogelijk. Gelukkig wordt de manipuleerbaarheid van PostScript-files steeds beter, door de komst van allerlei hulpprogramma's. Dat maakt het steeds beter mogelijk voor uitgevers om PostScript-files te accepteren als vorm waarin beeldmateriaal voor een artikel wordt aangeleverd.

4 TEX en nieuwe electronische produkten

Nieuwe electronische produkten worden samengesteld uit informatie die in een database is opgeslagen in SGML-gekodeerde vorm. Zoals boven al vermeld is SGML geen formatteringstaal. Dus de ruwe gegevens uit de database moeten gekonverteerd worden naar een taal die wel formatteert. TEX is daarvoor geschikt, zeker als de gegevens veel formules bevatten; het feit dat een TEX-file geheel in ASCII is is natuurlijk een voorwaarde om de teksten via e-mail te kunnen distribueren. Daarom hebben we konversieprogramma's ontwikkeld die de in SGML gekodeerde databaseuitvoer omzet in TEX-teksten die aan onze afnemers gepresenteerd kunnen worden.

Het zwaartepunt bij de nieuwe produkten heeft de afgelopen jaren gelegen bij het opzetten van zgn. alerting services. In deze tijd met zijn gigantische informatiestroom, die bovendien nog steeds groeit, is het moeilijk voor onderzoekers om op de hoogte te blijven van wat er in de vakliteratuur omgaat. Onze alerting services, Contents Alert geheten, helpen hen hierbij door hen regelmatig de verzamelde inhoudsopgaven toe te sturen van de verschillende tijdschriftnummers die Elsevier in de komende weken op hun vakgebied uit gaat brengen. Via electronic mail wordt deze informatie rechtstreeks op het bureau van de abonnees afgeleverd. Deze Contents Alert uitgaven worden gepresenteerd als plainTEX-teksten die op een zodanige wijze zijn getypt dat een lezer die niet over TEX beschikt er toch goed van kan kennisnemen.

Een stap verder gaat onze dienst *Nuclear Physics Electronic*. Het toonaangevende kernfysica tijdschrift *Nuclear Physics* zit in een hoek van de fysica waar snelheid een overheersende rol speelt en waar men bij uitstek ingespeeld is geraakt op het gebruik van netwerkkommunikatie en de modernste informatietechnologie. Nuclear Physics Electronic biedt kernfysici een database van de in het tijdschrift Nuclear Physics gepubliceerde artikelen. De database kombineert bibliografische informatie met de volledige gepubliceerde tekst. De eerste komt uit onze eigen database, en wordt gepresenteerd in LATEX. De volledige tekst van de artikelen komt van de auteurs en bestaat uit de TEX-files van de manuscripten.

5 Conclusie

Al vanaf het begin van de jaren tachtig is het bij Elsevier Science duidelijk geweest dat de opkomende informatietechnologie de werkwijze en de rol van de wetenschappelijke uitgever grondig zouden veranderen. Gedurende de laatste jaren hebben bij Elsevier Science de volgende informatietechnologische projekten duidelijk gestalte gekregen:

- In de vakgebieden wis- en natuurkunde publiceren we zoveel mogelijk uitgaande van de door auteurs in LATEX geschreven artikelen. In sommige afdelingen worden meer dan 50% van de artikelen, oplopend tot 90%, op

deze manier gepubliceerd. Het traditionele zetwerk vanaf manuscript is daardoor sterk afgenomen.

- In de andere vakgebieden verwerken we de auteursfiles van een zo breed mogelijk spectrum van tekstverwerkers; daarbij maken we een essentieel gebruik van SGML als intermediaire kodering en voor database opslag.

Ongetwijfeld zal de informatietechnologie ook in de komende decennia snel verder evolueren. Daarmee zullen het gezicht van de wetenschappelijke informatievoorziening en de rol en plaats daarin van de wetenschappelijke uitgever sterk veranderen, veel meer dan in het afgelopen decennium. In die komende situatie zal het een belangrijke taak van de uitgever zijn om informatie in gestruktureerde vorm op te slaan en klaar te zetten voor gebruik in meerdere vormen, zowel in allerlei uitgeefprojekten als direkt toegankelijk voor individuele klanten ('on demand publishing'). SGML is het gereedschap waarmee die strukturering bereikt kan worden. Voor het presenteren van een deel van de uitgeefprodukten is TEX een geschikte taal.

Referenties

[1] N.A.F.M. Poppelier, SGML and TEX in scientific publishing, TUGboat 12 (1991) 105–109.

[2] R.E. Youngen, W.B. Woolf and D.C. Latterner, Migration from Computer Modern fonts to Times fonts, TUGboat 10 (1989) 513–519.

[3] D.E. Knuth, Virtual fonts: more fun for grand wizards, TUGboat 11 (1990) 13–23.

[4] L. Lamport, LATEX, a document preparation system, user's guide and reference manual (Addison-Wesley, 1985, 1994).