FAKSIMIL

”SGML: En enkel och leverantörsoberoende dokumentstandard”,
IMT bulletinen. vol. XI, nr 4 1996, pp. 12–15.
Utges av Institutet för Medieteknik, KTH, Stockholm

Donald Broady intervjuad av Mats B. Andersson.

IMT
BULLETINEN

Utgiven av Institutet för Medieteknik
Årgång 11, nr 4/96

200 KVALITETSKALIBRERADE FÖRETAG ETT MÅL

BOKBINDERIFORSKNING TAR NY FART

DIGITAL BILD

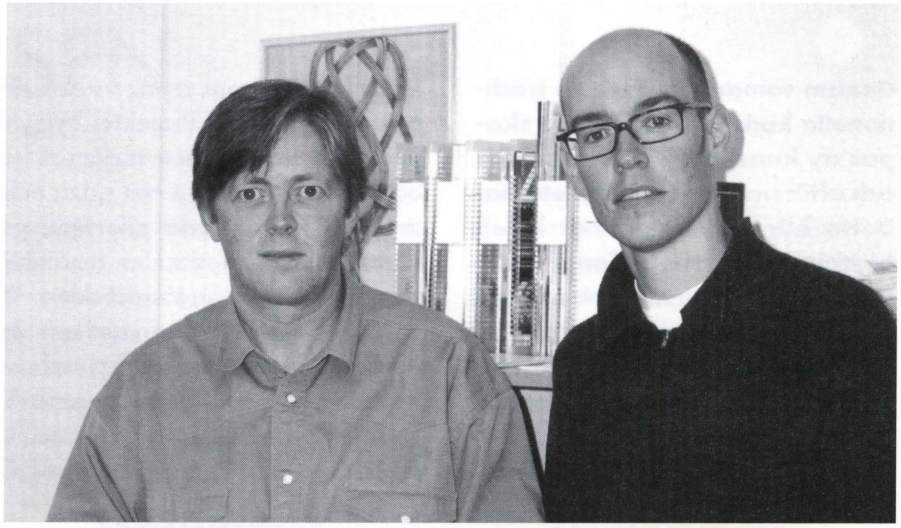
SGML EN DOKUMENTSTANDARD

PFT-PROGRAMMET UTVÄRDERAT

SGML:

En enkel och leverantörsoberoende dokumentstandard

Nu finns den enkla och leverantörsoberoende dokumentstandarden som gör parallellpublicering lättare. Med SGML-kodning kan dokumenten dessutom enkelt tas emot i all tänkbar utrustning och system över hela världen. Information kan lagras i många år och återanvändas i kommande generationers system och programvaror. Och det är betydligt lättare att uppdatera dokument om de är SGML-kodade. SGML= Std General Markup Language.



Artikelförfattaren Mats B. Andersson, IMT, tillsammans med Greger Henriksson från Norstedts Juridik (till vänster).

Hanterar du stora dokument och stora dokumentvolymer i din verksamhet ska du definitivt överväga att introducera SGML-kodning. Bäst är det naturligtvis om texten kodas samtidigt som den skrivs in. Men det är också relativt okomplicerat att SGML-koda äldre informationsmaterial.

Det finns en rad mycket intressanta exempel på företag som med stor framgång börjat arbeta med SGML-kodning.

Norstedts Tryckeri SGML-kodar SOU, Sveriges Offentliga Utredningar. Tidningarnas Telegrambyrå, TT håller på att utveckla ett enkelt system för SGML-kodning av det nyhetsmaterial som sprids till redaktioner över hela landet. Telekommunikationsföretaget Ericsson har omfattande manualer för service och underhåll. SGML har där kommit att bli ett värdefullt hjälpmedel inte minst för att underlätta och möjliggöra parallellpublicering. Dessutom menar Ericsson att informationen blivit mycket lättare att finna nu när den SGML-kodats.

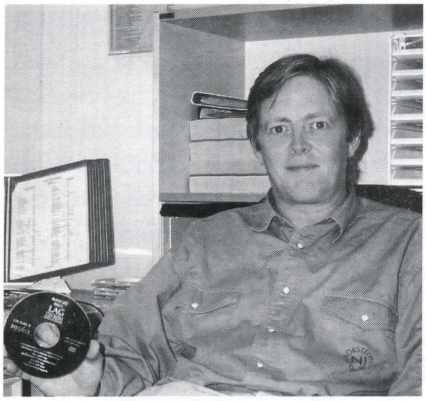
Nyheter från TT

TT har sedan 1994 intresserat sig för SGML och de har numera bestämt sig för att implementera det i sitt nyhetssystem. TT beräknar starta med det SGML-baserade informationssystemet kring årsskiftet 1996–97. Till en början enbart med textinskrivning men kring sommaren 1997 också för andra delar som radio/TV-tablåer, börstabeller och sport.

TTs kunder ställer sig mycket positiva till det nya systemet och de ökade möjligheter till intelligent kodning som detta ger. Problemet med dagens format IPTC 7901 är att den innehåller sätteri-information. Layout och struktur är ihopvävt vilket försvårar parallellpublicering. Formatet har dålig intelligens och det är svårt att lägga in information om informationen, s k metainformation.

TT kommer att bygga sin SGML-verksamhet kring en DTD (Document Type Definition) som kallas NITF, News Industry Text Format. En svårighet i projektet har varit att sätta sig in i en så komplex DTD som NITF. Ibland kan den kännas litet väl stor men det skapar också många möjligheter. NITF är framtagen av IPTC helt med utgångspunkt i nyhetsförmedling. IPTC är ett standardiseringsorgan vars medlemmar är ett antal stora nyhetsbyråer och ett antal större världsledande tidningar. De arbetar med standardiseringsfrågor som bl a rör språk, bild, kommunikation, kodformat. Även multimedia försöker de komplettera som en påbyggnad på NITF.

Projektet kan sägas ligga i teknikens framkant och TT har haft det goda omdömet att ta visst stöd av extern SGML-kompetens. De har haft lärare från KTH och Ericsson och har diskuterat med annan SGML-exper-



Greger Henriksson, Norstedts Juridik, har lagt in juridisk dokumentation på CD. Denna omfattande dokumentering har blivit lätthanterlig med hjälp av SGML.

tis. För textinskrivning tänker de använda en egenutvecklad editor som lägger på märkningen för SGML när filen sparas, samt tar bort märkningen när en SGML-fil importeras. Olika fält som motsvarar olika märkningar ska göra det enkelt för journalisten att skriva telegram utan att behöva bry sig om SGML-märkning.

TT lovordar enkelheten

TT räknar med att det inte kommer att behövas någon speciell utbildning för de journalister som ska hantera systemet. Genom att använda ett gränssnitt med fält i vilka man fyller i de olika uppgifter som hör till ett dokument, blir hanteringen mycket enkel. Dokumenten SGML-uppmärks automatiskt vid lagring tack vare fälten och man slipper momentet att SGML-märka.

Norstedts Juridik

Greger Henriksson, vid Norstedts Juridik har valt att arbeta med SGML:

– Skälet var främst att det krävdes en struktur för att navigera i författningar och för att kunna genomföra sökningar på olika dokumentdelar, berättar Greger Henriksson. Det gällde också att bestämma ett sätt att kunna hantera filer över en mycket lång tid. Vi kan inte vara beroende av leverantörers olika nycker när vi ska underhålla lagtexter under de år som kommer. Det var också med sikte på CD-ROM-produktion som kodningen inleddes, vi såg att själva produktionen skulle bli snabbare och enklare om vi använde SGML.

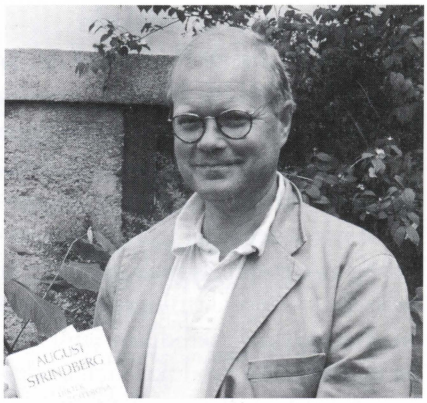
August Strindberg

Donald Broady är projektledare för en försöksverksamhet vid NADA/CID på KTH som bl a inneburit SGML-stöd för Strindbergs samlade verk. Bland medarbetarna finns också Hasse Haitto, en välkänd person i SGML-sammanhang. Försöket innebar SGML-stöd vid produktion av två volymer i serien Nationalupplagan av August Strindbergs samlade verk, nämligen romanen Svarta Fanor och diktboken Dikter på vers och prosa.

Arbetet innebar att i samarbete med redaktionen producera SGML-filer och att från dessa filer generera tryckfärdigt original. De ansvarade för hela den tekniska produktionen av volymerna, från redaktörernas manuskript till tryckfärdigt original (PostScript-filer) som levereras till tryckeriet. Märkningen baseras på SGML och följer TEI:s riktlinjer (Text Encoding Initiative). Arbetsgruppen vid CID ombesörjde typograferingen med hjälp av program de själva utvecklat samt LaTeX som formaterare. Arbetet har närmast inneburit att utveckla en omfattande "regelbok" i form av datorprogram konstruerade med utgångspunkt i dels utgivarnas kunskaper om textkritisk vetenskaplig utgivning, dels formgivarnas kunskaper om typografi.

Av flera skäl valdes SGML i produktion. Donald Broady berättar:

– Vi ville ge redaktörerna ökad kontroll över typografin. Därför försökte vi så långt som möjligt fånga upp den typografiskt relevanta informationen redan i SGML-filerna, dvs vi försökte minimera det resterande manuella arbetet i formateringsprogrammet LaTeX. I jämförelse med traditionellt sättningsförfarande, där korrektur sänds fram och åter, innebar detta förfarande att redaktionen gavs bättre kontroll. Redaktörerna kunde vara säkra på att generella ändringar i SGML-filen eller skripten slog igenom överallt (såsom den typografiska utformningen av olika typer av rubriker och stycken, nonbreaking space mellan initial och efternamn, avstavningspreferenser, principer för horungekontroll och alla andra tänkbara typografiska regelbundenheter). Redaktörerna och formgivarna kunde när som helst ta del av skärmpresentation eller laserutskrift för kontroll av enskilda sidor, de kunde pröva sig fram genom att skriva ut alternativ och även i övrigt blev arbetsgången utpräglat explorativ.



Donald Broady, NADA/CID har valt SGML i produktion av August Strindbergs verk.

– Och framför allt ville vi skapa SGML-filer som är möjliga att återanvända i framtiden, för tryck i andra sammanhang samt för en digital utgåva, framhåller Donald Broady.

Den största fördelen i detta sammanhang var den regelbaserade arbetsgången. Många problem kan lösas en gång för alla. SGML-standarden ger en garanti för att konsekvent kodning (som innebär att samma sak alltid kodas på samma sätt) ger konsekvent layout. En sådan regelbaserad layout tillåter att tidigare arbete kan tas tillvara och att inga överraskningar tillkommer.

Ta sig över tröskeln

Den största svårigheten är att det är svårt att ta sig över tröskeln. Specialistkompetens behövs på områdena textkritik, bokformgivning, SGML, samt beträffande formateringsprogrammet (i detta fall LaTeX) och justering av typsnitt. Initialkostnaden är avsevärd. Uppskattningsvis krävde utarbetandet av metodiken samt själva den tekniska produktionen av dessa två volymer flera personmånader. Själva SGML-märkningen tog i sig bara några veckor när rutinerna förelåg.

Men när väl arbetsgången fungerar är förfarandet förhållandevis billigt förutsatt att skrifterna är tämligen enhetliga i fråga om den typografiska formgivningen. Vinsterna blir självfallet större ju fler utgåvor av samma slag som produceras. De kunde exempelvis använda samma skript för produktion av avsnitten med ordförklaringar i de två volymer vi framställde. Här var typografin identisk, och framställningen av den andra volymens ordförklaringar gick mycket snabbt.

Billig programvara

Donald Broady menar att det inte krävs särskilt kostsam programvara.

– Vi gjorde ganska enkelt bruk av de program vi hade tillgång till, och även enklare program hade dugt bra. Kommersiella standardprogram eller public domain-program räcker långt. För att skapa översättningsskript använde vi OmniMark eftersom vi fick en akademisk licens på förmånliga villkor, men vi utnyttjade egentligen inte detta programs fulla potential. Det hade gått bra med gratisprogram, t ex ett enkelt översättningsprogram (t ex standardkommandon under unix) i förening med ett valideringsprogram, t ex James Clarks beprövade parser SP.

Det måste dock tillfogas att TEIs DTD är utomordentligt väl genomtänkt. Det är välförtjänt att den på senare tid blivit ett slags internationell rikslikare för kodning och utbyte av humanistisk digital litteratur. Många behov är redan förutsedda i DTDn, och om man vill göra något därutöver är detta lätt ordnat tack vare den modulära uppbyggnaden och möjligheterna att enkelt göra egna ändringar och tillägg.

Materialet SGML-kodades i samarbete med redaktionen, antingen i vanlig editor (emacs, MS World) med efterföljande validering, eller i specifik SGML-editor (Author/Editor). Den generella översättningsprogramvaran OmniMark användes för validering och finputsning av kodningen. Som formaterare (layout-program) valdes LaTeX. OmniMark-skript skapades för indata till formateraren. För LaTeX skapades även styrfiler som tillät att layoutinstruktioner infogades antingen redan i SGML-filen eller i OmniMark. Utmatningen från LaTeX skedde i form av PostScript-filer som sändes till tryckeriet. PostScript-utmatningen användes även under arbetets gång för skärmpresentationer eller laserutskrifter som redaktörerna och bokformgivarna kunda ta ställning till.

– För närvarande har vi enbart medverkat i publiceringen av ordinära böcker, men siktet är inställt på en digital utgåva. Som ett förarbete har vi genomfört en rudimentär SGML-märkning av det fyrtiotal volymer av Nationalupplagan som hittills utgivits, avslutar Donald Broady. ■

Mats B Andersson

DATA OM DATORUTVECKLINGEN

Gates/chip	ökning 2,5 ggr/3 år
DRAM bits/chip	ökning 4 ggr/3 år
Mikroprocessorer/chip storlek	ökning 2 ggr/1,5 år
MIPS/watt	ökning 3 ggr/1,5 år
Komponenter/chip	ökning 10 ggr/10 år
Klockfrekvens	ökning 4 ggr/10 år
Minneskostnad	minskning 10 ggr/10 år
Datahastighet	ökning 2 ggr/2 år
Minnesstorlek	ökning 4 ggr/2 år
Prestanda : pris	ökning 1.000.000/20 år

källa: Pira International

CD-ROM-UTVECKLINGEN

CD-ROM marknadens värde

1993	325 miljoner dollar
1995	793
1997	2.500

CD-ROM spelare på marknaden

1992	1,5 milj CD-ROM spelare
1993	6,7
1995	20,0

källa: Pira International

TEKNIKFAKTA

Uppmärkning

System för ord- och textbehandling kräver ofta att extra information läggs in i den ursprungliga texten. Uppmärkning för att tala om hur systemet ska formattera de olika delarna typografiskt, har den grafiska industrin sedan lång tid varit väl förtrogen med.

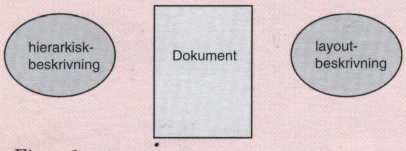
Den största skillnaden mellan SGML och andra lagringsformat är att de märkord, som ligger insprängda i texten, anger vilken karaktär innehållet i texten har. Märkorden placeras dels i början av informationselementet, dels i slutet. De har inget att göra med information om textens typografiska utseende teckensnitt, grad osv.

Ett informationselement kan se ut på följande sätt:
<markering> innehåll </markering>

Dokumenten lagras i bitar som informationselement och utan layoutkoder knuten till sig. Bitarna samlas, enligt en uppsättning regler i DTDn, ihop till ett dokument som kan publiceras med önskad layout. Den önskade layouten kan då innebära att all text som är markerad med <markering> ska presenteras med times, *italic*, 12 punkter.

Dokumentet

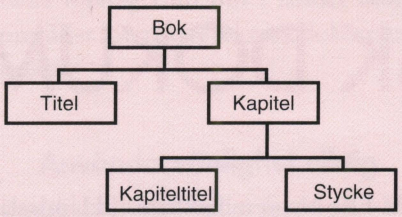
Idén är lika enkel som genialisk. Det fullständiga dokumentet delas upp i tre delar se figur 1.



Figur 1

- Innehåll (Det som sägs i texten)
- Utseende (Hur texten är presenterad, d v s layout, typsnitt etc)
- Struktur (Vilka olika delar som ingår i texten och hur dessa är relaterade till varandra). Se figur 2.

Dokumentstruktur



Figur 2. Grafisk beskrivning av ett dokuments struktur.

Dokumenttypsdefinition, DTD

Den hierarkiska strukturen i dokumentet anges i dokumenttypsdefinitionen (DTD = Document Type Definition) för definition av dokumentets hierarkiska struktur. (Se figur 3.) DTDn är en uppsättning regler som anger vad ett dokument får innehålla och i vilken ordning. Exempelvis kanske en rubrik måste komma före en paragraf. I DTDn anges vilka delar som bygger upp en viss typ av dokument. Varje typ av dokument har sin egen DTD. SGML märker upp texten med koder som beskriver innehållet i de delar som bygger upp dokumentet. Det är dessa märkord som texten uppmärks med som deklareras i DTDn.

Dokumenttypsdefinition

```
<!ELEMENT bok (titel+)>
<!ELEMENT kapitel (kapitel, stycke+)>
<!ELEMENT titel (#PCDATA)>
<!ELEMENT kapitel (#PCDATA)>
<!ELEMENT stycke (#PCDATA)>
```

Figur 3. En dokumenttypsdefinition, DTD på den bok vars struktur beskrivs i figur 2.

YTTERLIGARE INFO OM SGML

Litteratur

För den intresserade finns litteratur i ämnet, exempelvis:

"The SGML handbook"
av Charles F. Goldfarb, 1990

"Practical SGML"
av Erik van Herwijnen, 1994

"Using SGML"
av Martin Colby & David S. Jackson, 1996

På Internet finns en hel del material, bland annat svenska användarföreningens sidor som kan vara värda ett besök.

SGML användarförening i Sverige

Sedan 1993 finns en ideell användarförening med syfte att vara ett forum och kontaktnät för spridande av kunskap och information kring SGML. Föreningen arrangerar seminarier och konferenser samt stödjer sådan utveckling som leder till ett aktivt användande av SGML. Föreningen är indelad i fyra grupper med inriktningar mot: Publicering, databashantering, HyTime och DTD-utveckling. Mer information om föreningen finns att hitta på Internet:
http://info.admin.kth.se/SGML/

Kontaktperson för styrelsen är Helena Antbäck, tel 08-708 80 15.

Leverantörer/utbildare

Exempel på leverantörer och utbildare inom SGML-området: CPS Computer Publishing Systems AB, IBM Svenska AB, Publishing Development AB, SDR Scandinavian Data Resource AB och Texcel AB.

Institutet för Medieteknik, IMT

IMT bistår industrin och dess organisationer med konsultationer beträffade dokumentstandarden SGML.
Kontakta IMT på tel: 08-453 57 00.

MBA