Detta dokument är ett underlag som i förväg
sändes till deltagarna i mötet
på Kungl. Biblioteket 1996-09-19
om ev framtida Strindbergutgivning på nätet.

Digitala utgåvor av August Strindbergs verk

Inledande överväganden

Version 3
27 augusti 1996

Donald Broady
Nada/CID, KTH
100 44 Stockholm

Epost broady@nada.kth.se
URL http://www.nada.kth.se/~broady

Innehåll

1. Inledning

Vid Institutionen för numerisk analys och datalogi, KTH, har vi sedan 1993 arbetat med att utforska de tekniska möjligheterna för framtida digitala utgåvor av August Strindbergs verk. Arbetet har skett inom ramen för de FRN-finansierade projekten ”Det nya handbiblioteket” och ”Textkritisk digital litteratur”.[1]

Nedan har jag sammanfattat en del erfarenheter och preliminära överväganden. Det handlar i första hand om tekniska lösningar, som tål att diskuteras. Många andra överväganden behövs, främst synpunkter från Strindbergforskare och andra litteraturvetare och humanister, men även från specialister från arkiv- och biblioteksvärlden, samt från dem som förstår sig på juridiska, ekonomiska och förlagspolitiska förutsättningar. Därför är det följande ett diskussionsunderlag, inget färdigt förslag.

Den läsare som är obekant med SGML och TEI kan läsa den här texten bakifrån. Avsnitt 5 är en kort introduktion till SGML, avsnitt 7 ett exempel på TEI-märkning.

2. Urval av innehåll

En heltäckande digital utgåva, eller om man så vill ett digitalt Strindbergarkiv, borde helst innefatta:

- Texten i Nationalupplagan av August Strindbergs samlade verk (nedan betecknad ”Saml. Verk”), Almquist & Wiksell Förlag (1981—1985) samt Norstedts Förlag (1986 ff). Enligt planerna skall upplagan bestå av 72 textvolymer, varav 42 är utgivna, var och en bestående av etablerad text, kommentar om tillkomst och mottagande samt ordförklaringar. Därtill planeras 10 ännu inte färdigställda volymer med textkritisk kommentar. Huvudredaktör är Lars Dahlbäck.

- Texten i *August Strindbergs brev* (nedan betecknad”Brev”), utg. Torsten Eklund (vol. I—XV) och Björn Meidal (vol. XVI ff), Bonniers Förlag.

- Förarbeten, utöver dem som enligt planerna kommer att publiceras i kommentarvolymerna till Saml. Verk. De flesta förarbeten finns i den hos Kungl. Biblioteket deponerade samlingen ”Gröna säcken”, särsk. kartongerna 1—9.

[1] Se forskningsprogrammen D. Broady, *Det nya handbiblioteket*. Kungl. Tekniska Högskolan, Institutionen för numerisk analys och datalogi, Reports from Interaction and Presentation Laboratory 73, april 1993 (omtryck s. 83—107 i *Biblioteken, Kulturen och den sociala intelligensen. Aktuell forskning inom biblioteks- och informationsvetenskap*. Red. Lars Höglund. Forskningsrådsnämnden/Valfrid, Göteborg 1995) samt ”Textkritisk digital litteratur” (Ansökan till Forskningsrådsnämnden, programmet Biblioteks- och informationsvetenskaplig forskning, 9 okt. 1995).

Redan från början var arbete med Strindbergmaterial ett centralt inslag i dessa projekt, som framgår av ansökan till FRN för projektet ”Det nya handbiblioteket” 1993-02-15: ”Som en fallstudie prövas metoder för att framställa en SGML-märkt maskinläsbar version av Nationalupplagan av August Strindbergs samlade verk i enlighet med de riktlinjer som utarbetats av TEI (Text Encoding Initiative).” Lars Dahlbäck, huvudredaktör för Nationalupplagan av August Strindbergs samlade verk, var medsökande i samband med denna ansökan och arbetet kom att utföras i nära samverkan med Strindbergsprojektet vid Stockholms universitet.

Vid KTH har vi prövat olika verktyg och märkningsalternativ, och genomfört konvertering och en första rudimentär SGML-märkning av de 42 hittills utgivna volymerna av Saml. Verk. Under 1995 ansvarade vi för den tekniska produktionen — från märkningen av redaktörernas manuskrip till tryckfärdigt original — av de två senaste volymerna, *Svarta Fanor* (vol. 57) och *Dikter på vers och prosa. Sömngångarnätter på vakna dagar och strödda tidiga dikter* (vol. 15).

- Annat av Strindbergs hand som saknas i Saml. Verk och Brev. Eventuellt även teckningar och målningar.

- Brev till Strindberg.

- Litteratur som Strindberg citerade eller på annat sätt förhållit sig till.

- Fotografier, kartor o.likn.

- Relevanta kataloger, förteckningar, auktoritetslistor för stavning av namn, etc. Det handlar om både nyskapat material och återutgivning av exempelvis Barbro Ståhle Sjönells katalogisering av innehållet i ”Gröna säcken” och Hans Lindströms *Strindberg och böckerna* (del I, 1977, förteckn. över Strindbergs bibliotek, samt del II, 1990, förteckn. över Strindbergs boklån och läsning).

- Sekundärlitteratur, handbokstext, material för undervisningsbruk.

- Viss programvara som mottagarna kan behöva för att utnyttja materialet.

I denna önskelista finns ingen särskild punkt för textkritisk kommentar. Sådan bör infogas i texterna. Inte heller index, namnregister, innehållsförteckningar, konkordanser, ackumulerade ordförklaringar etc bör behandlas som fristående information, utan skall helst vid behov kunna extraheras fram ur den märkta texten.

En heltäckande utgåva som den ovan skisserade är självfallet ett gigantiskt och kostsamt projekt som i bästa fall kan förverkligas långt in i nästa århundrade. Men även till omfång och ambitionsnivå blygsammare projekt bör helst lämna bidrag till framväxten av ett mer beständigt digitalt Strindbergarkiv. Materialet skall för det första tillfredsställa vetenskapliga krav på urval, textetablering och kommentar, för det andra utformas (se avsnittet om arkivformat nedan) så att det är flyttbart och följer (eller kan konverteras till) internationella standarder.

Det förefaller rimligt att digitala versioner av bokutgåvorna Saml. Verk och Brev utgör kärnan i ett framtida digitalt Strindbergarkiv, men på kortare sikt är det mer realistiskt att tänka sig publicering av enskilda verk.

En ambitiös utgåva av Inferno skulle kunna innehålla Strindbergs franska originalmanuskript (i faksimil samt transkribering), utgivarnas kommentar, ordförklaringar, de tre först publicerade versionerna av texten (Fahlstedts översättning till svenska 1897, den av Strindbergs korrigerade andra upplagan från samma år, den första franska utgåvan 1898), transkriberingar av ytterligare Strindbergtexter från samma period (brev, Ockulta Dagboken, outgivet material), återgivningar av de texter som Strindberg hänvisar till på eller mellan raderna (Gamla Testamentet, teosofiska och andra ockulta skrifter, Swedenborg, Balzacs *Séraphita)* samt relevant sekundärlitteratur. En sådan komplett utgåva skulle dels tjäna specialisternas intressen, dels kunna utgöra underlag för en normaliserad förenklad ”folkupplaga”.

En mer anspråkslös utgåva av ett enskilt verk kunde innehålla ungefär detsamma som t.ex. en volym i serien Saml. Verk utan andra tillägg än kanske en textkritisk kommentar och självfallet interna länkar.

En annan möjlighet är en samlad utgivning av några besläktade verk; förslagsvis en samling samhällskritisk prosa omfattande Röda rummet, Götiska rummen, Svarta Fanor samt Tal till svenska nationen jämte kommentar och relevanta förarbeten och brev.

Eller en utgåva med Stridbergs självbiografiska verk i linje med hans egna notering (SgNM 1:1, 20) enligt vilken han önskade sig en postum utgåva under rubriken ”Tjänstekvinnans son” innefattande Tjänstekvinnas son del 1—4, brev, Ockulta dagboken, Inferno, Ensam m.fl..

Man kan förstås även tänka sig populära utgåvor utan vetenskaplig apparat, förslagsvis inför kulturhuvudstadsåret 1998 en utgåva ”Strindbergs Stockholm” med lämpligt textmaterial och många

kartor och fotografier och andra illustrationer. Det är ofrånkomligt att en multimedial presentation över nätet eller på CD-ROM innefattar skräddarsydda texter, bilder, videosekvenser och interaktiva möjligheter som inte låter sig återanvändas i andra sammanhang. Men även i ett sådant fall bör helst det centrala text- och bildmaterialet färdigställas på sådant sätt att det uppfyller de fordringar man har rätt att ställa på ett framtida digitalt Strindbergarkiv, och så att — vilket är samma sak — läsarna kan hämta ut detta material och göra eget bruk därav i trygg förvissning om att det är kvalitetsgranskat.

3. Krav på kritiska digitala utgåvor

Den grundläggande uppgiften är att framställa material som kan ingå i ett digitalt Strindbergarkiv av bestående värde, dvs. en motsvarighet till vetenskapliga utgåvor i bokform. Sådana bokutgåvor består av transkriberad och kommenterad text och/eller av faksimil. Principerna bakom urval, attribuering och textetablering redovisas. Läsaren får besked om källornas beskaffenhet och varianter. Vetenskapliga utgåvor är en gemensam tillgång för alla som arbetar inom området. De tjänar som en förutsättning både för individuellt arbete — idealet är att informationen skall vara så tillförlitlig och fullständig att läsaren i normalfallet slipper konsultera originalmanuskript eller tidigare utgåvor — och för samarbete och debatt. Man vet att man talar om samma sak när man citerar, hänvisar och tolkar.

För dessa tryckta kritiska utgåvor existerar en lång tradition betingad av skriv- och tryckteknologin. Det har tagit hundratals år att utveckla de redigeringsprinciper som tillåter att man presenterar attribueringar, varianter och tolkningar, anger från vilka utgivare olika ingrepp stammar, sammanställer konkordanser etc — och inte minst har man nått fram till en konsensus om hur allt detta skall åskådliggöras på papper i enlighet med ändamålsenliga och vedertagna typografiska principer.

För kritiska digitala utgåvor, som blott funnits under några få år, har ett motsvarande ”standardiseringsarbete” nätt och jämt inletts. Situationen påkallar ödmjukhet. Grundregeln borde vara att digitala utgåvor skall bevara de landvinningar som filologer, texttolkare, utgivare och typografiska formgivare åstadkommit genom seklen beträffande utgåvor på papper. I dessa avseenden bör digitala utgåvor inte vara sämre än tryckta. Eftersom digitala utgåvor inte är bundna till pappersmediet bör tvärtom ur många aspekter vara överlägsna. I tryckta kritiska utgåvor använder man för att spara plats en komprimerad och för läsaren ofta svårgenomtränglig notation för att hålla reda på varianter, för hänvisningar, för textkritisk kommentar etc. Man tvingas nöja sig med att av trycket utge en mindre del av det material som utgivarna arbetat med. Digital teknik besväras inte av dessa begränsningar. Förkortningarna kan för den läsare som så önskar vara tillgängliga i klartext. I stället för att bara hänvisa till texter kan man in vissa fall infoga dessa in extenso (och/eller länkar till dem). Man kan ur texten vid behov extrahera skräddarsydda konkordanser, index, innehållsförteckningar och andra orienteringshjälpmedel i stället för att publicera separata sådana som blir inaktuella så snart texten revideras eller utökas. Man kan presentera varianter — eller text och kommentar, eller text tillsammans med översättning till främmande språk — sida vid sida på skärmen med länkar mellan motsvarande passager. Man kan foga in kommentar, förklaringar eller varianter (eller länkar till dessa) på de ställen i texten i de hör hemma — ungefärligen som om redaktören av en bokutgåva kunde tillåta sig ett obegränsat antal obegränsat långa fotnoter på varje sida, fotnoter som dessutom kan innehålla andra fotnoter. Man kan välja många olika presentationsformer för en och samma text: faksimil av manuskriptsidor, för forskningsbruk en transkriberad diplomatarisk version som ligger så nära ett originalmanuskript som möjligt, versioner avsedd för databashantering, förenklade och normaliserade versioner avsedda som ”folkupplagor”, andra upplagor rikligt försedda med länkar till kommentar och och sekundär- och referenslitteratur etc.

I dag finns en olycklig och i någon mån onödig tendens till tudelning av den humanistiska digitala litteraturen. Å ena sidan skapas kvalificerade arkiv och verkutgåvor för den akademiska humanistiska världen, å andra sidan populära CD-ROM-produktioner, WWW-sidor eller arkiv med råtext för den breda publiken och för skolbruk.

De vetenskapliga digitala textarkiv som hittills byggts upp inom den akademiska världen brukar vara ganska svårtillgängliga och kräva en hel del utrustning och kunnande, inte minst av teknisk art. Ofta har man använt egna märkningsscheman som inte följer några standarder och som användaren därför har svårt att konvertera till sitt eget system. Förlagens utgåvor för den akademiska publiken är lättare att använda men i gengäld ibland så dyra att bara välbemedlade bibliotek och utbildningsinstitutioner har råd att skaffa dem. Det engelska förlaget Chadwyck-Healey, som är marknadsledande när det gäller publicering av kvalificerad digital litteratur för humanister, begär 25 000 pund för fem CD-ROM innehållande ”The English Poetry Full-Text Database”, en omfattande (165 000 dikter) samling engelsk poesi fram till slutet av 1800-talet. Ett annat problem är att förlagen brukar publicera materialet i en form som är hårt bunden till de presentations-, indexerings- och sökprogram med vilka det levereras. Det fungerar väl så användaren nöjer sig med att söka på sin CD-ROM, läsa på skärmen eller skriva ut isolerade partier, men man stöter på problem så snart man vill införliva materialet med sitt eget system för att göra något annat därmed än vad förlaget föreställt sig.

Förlagens publicering för den breda publiken har hittills prioriterat formgivningen framför den textkritiska kvaliteten. Det är gränssnittets attraktionskraft, bilderna och video- och ljudsekvenserna som skall sälja varan. Många CD-ROM-produkter kallas ”interaktiva”, men det är en interaktivitet som innebär att läsaren bereds tillfälle att följa på förhand upptrampade stigar, inte att utveckla egna sätt att umgås med materialet. Den läsare som vill arbeta med materialet och inte bara få det presenterat för sig tvingas lägga ned åtskilligt arbete på konvertering och redigering innan han eller hon kan göra något eget med texten, vilken för övrigt sällan tillfredsställer ordinära textkritiska krav. I andra ändan av skalan finner vi en förkärlek för asketiska presentationsformer hos de många arkiv — projekt Gutenberg är det mest kända — som gör råa textfiler tillgängliga över internet. Här saknas varje slags märkning och länkning, det är alltså fråga om återgivningar av texter bokstav för bokstav — i bästa fall. Alltför ofta är alstren eländiga ur textkritisk synvinkel. Läsaren lämnas i okunnighet om redigeringsprinciper, ibland t.o.m. om vilken tryckt utgåva som utgjort underlaget. Än värre är de många rena avskrivningsfelen. Oftare än man tror är texterna ett snabbt hopkommet resultat av skanning och OCR-tolkning utan tillräcklig efterföljande kontroll.

Ytterligare ett slag av digital litteratur avsedd för både forskare och den breda publiken är den som återger en förlagas utseende. Humanister har nytta av de framväxande bildarkiven med avfotograferade manuskript, kartor eller konstföremål. Så kommer exempelvis Wittgensteinforskarna säkert att välkomna den faksimilutgåva av Wittgensteins 20 000 efterlämnade manuskriptblad — varav hälften aldrig förut publicerats — som nu färdigställes i Bergen och till sommaren 1997 skall utges på fyra CD-ROM av Cambridge University Press. Vissa tidskrifter publicerar en parallell digital version som återger pappersupplagans typografi. Men när det gäller ett författarskap som Strindbergs förefaller det mer rimligare att att låta den etablerade texten vara huvudsak, självfallet gärna försedd med länkar till bilder av manuskript. (Den nämnda faksimilutgåvan av Wittgensteins papper utgör en mindre del av verksamheten vid Wittgensteinarkivet i Bergen; det stora arbetet består i skapa transkriberade diplomatariska versioner av samma material vilka kan filtreras till olika utmatningsformat).

Jag behöver knappast orda om vikten av att Strindbergs författarskap göres lätt tillgängligt i forskning och undervisning och för allmänheten. En kulturnation borde sörja för att de viktigaste gemensamma skatterna på bästa sätt göres tillgängliga för medborgarna. När vi har att göra med digitalt material innebär tillgängligheten något mer än bara att det tillhandahålles i vilket skick som helst. Det skall vara åtkomligt i beständigt och flyttbart format. Det får med andra ord inte vara inlåst i speciella tillämpningar som bara är tillgängliga på vissa plattformar och medier och som ganska snart blir föråldrade. En Strindbergproduktion på CD-ROM eller på WWW-sidor är inte särskilt tillgänglig, och inte särskilt länge, om materialet är anpassat till märkningsscheman, programvara, plattformar eller skärmstandarder som kommer att vara föråldrade om några år. Det finns alltid en risk för att vi låter oss styras av de för ögonblicket mest tilltalande presentationsformaten (i dag multimediaproduktioner på CD-ROM, samt för WWW-presentation den för ögonblicket populäraste varianten av HTML-märkning). Men att anpassa arkivformatet därefter vore som om varje bok i

bibliotekens samlingar vore avsedd att läsas genom speciella glasögon som förmodligen inte kommer att fungera om några år. Kulturarvet får inte hanteras som färskvara med bäst före-datum.

Det förefaller troligt att digitala versioner av bokutgåvorna Saml. Verk och Brev under lång tid framöver kommer att utgöra kärnan i ett digitalt Strindbergarkiv. Principer för urval, textetablering, attribuering och presentation kunde kanske övertas därifrån och anpassas till de digitala mediernas krav. Detta är dock en sak som Strindbergforskarna måste bedöma. Ur teknisk synvinkel skulle ett bibehållande av de existerande bokutgåvornas disposition bl.a. innebära att förarbeten behandlas som separata dokument i stället för att infogas i bastexten (men självfallet länkade till denna). Efter hand kan andra och med tanke på digital publicering elegantare redigeringsprinciper övervägas, men det vore enklast att slippa alltför stora omöbleringar i förhållande till de existerande bokutgåvorna, bl.a. eftersom tekniken hur som helst måste utformas så att läsarna ges möjlighet att använda böckerna och digitala versioner parallellt. Jag kan tänka mig att märkningsarbetet skulle kunna följa principen att de textsträngar som återstår om all märkning skalas bort är identisk med den publicerade texten i bokutgåvorna. Vidare skulle ett väl definierat urval av bokutgåvornas typografiska egenskaper (sid- och kanske radbrytning, betydelsebärande teckenformatering, åtskillnader mellan olika slag av stycken och rubriker etc) fångas upp i märkningen. Sådana ambition ställer bestämda krav på märkningsschemat. En ortodox SGML-filosofi, enligt vilken typografin skall lämnas därhän, är inte tillämplig när man har befintlig text på papper att ta hänsyn till (Strindbergs manuskript, tidiga tryckta versioner, samt Saml. Verk och Brev). Tillvägagångssättet måste i stället ansluta till TEIs centrala ambition: att troget beskriva egenskaper, även visuella, hos text som redan föreligger på papper. En helt annan sak är att det är viktigt att undvika låsningar till typograferings- och övriga presentationsfunktioner hos de system till vilka Strindbergarkivets material skall levereras.

Vissa omdisponeringar förefaller dock befogade. Ordförklaringarna i Saml. Verk bör inte fördelas på de olika volymerna utan samordnas (här finns hos Språkdata i Göteborg en ackumulerad samling ordförklaringar att utgå från) och länkas in i texten. Viss textkritisk kommentar bör infogas i texten i stället för att lagras separat: när det föreligger varianter på detaljnivå eller när Strindberg skrivit med rödpenna bör detta anges i anslutning till textstället i fråga.

Fortsatta överväganden om redigeringsprinciper kräver samverkan mellan dels Strindbergforskare och andra som är förtrogna med textetablering och redigering av vetenskapliga utgåvor på papper, dels specialister som känner till möjligheterna och begränsningarna hos dagens och morgondagens informationsteknik. En rejäl utredning skulle behövas, motsvarande den som Lars Dahlbäck genomförde som grundval för bokutgåvan Saml. Verk.[2]

[2] Lars Dahlbäcks utredning resulterade i de redigeringsprinciper som redovisas i Saml. Verk, vol 1, s. 275—333 (*Ungdomsdramer I*, 1989).

4. Arkivformat, utbytesformat, presentationsformat

Även om publiceringen under de närmaste åren bara kan gälla mindre delar av författarskapet, bör arbetet genomföras på så sätt att det bereder marken för mer fullständiga utgåvor. Viktigast inför framtiden är att skapa digitala arkiv av bestående värde, dvs. arkiv med material som dels fyller vetenskapliga krav, dels lagras i format som tillåter att det revideras och återanvändes för kommande behov.

Det är väsentligt att skilja mellan arkivformat, utbytesformat och presentationsformat.

Med arkivformat (en alternativ benämning är ”lagringsformat”) avser jag den ”definitiva” form i vilken materialet lagras hos ett framtida digitalt Strindbergarkiv. Arkivformaten måste vara fåtaliga och tämligen beständiga. I arkivformaten skall alla av redaktörena godkända kompletteringar och revideringar införas. Helst skall varje beståndsdel i materialet (text, bild, länkinformation) föreligga i en och endast en ”masterversion”, från vilken konvertering sker till andra format vilka kan vara mer kortlivade och anpassade till stundens krav.

Utbytesformaten används för att sända materialet till andra institutioner eller andra datorsystem, och måste i viss mån utformas med tanke på mottagarnas behov och tekniska förutsättningar. Antalet utbytesformat bör dock begränsas och så långt som möjligt åstadkommas genom automatisk utmatning från arkivformaten.

Presentationsformaten kan vara hur många som helst och snabbt föränderliga. De är avpassade för en särskild publik eller de för tillfället lämpligaste distributionsteknikerna eller medierna (över internet eller intranet, på papper, CD-ROM, World Wide Web etc). Medan jag tänker mig att arkivformaten och utbytesformaten skall kontrolleras av den eller de institutioner som ansvarar för ett framtida digitalt Strindbergarkiv (det kan självfallet bli fråga om ett virtuellt arkiv vars samlingar är spridda över flera institutioner), kommer presentationsformaten att kontrolleras av andra: av förlag, forskningsinstitutioner och bibliotek, samt förstås läsarna själva.

4.1 Arkivformat

Det viktigaste avgörandet gäller arkivformatet. Det gäller att finna ändamålsenliga format för tre slag av information: text, bilder och länkar. Ett väl valt arkivformat innebär att materialet blir beständigt och mångsidigt användbart. Ett och samma material kan levereras och presenteras på olika sätt för skilda läsare (specialister, amatörer, skolelever), filtreras för att sändas mellan institutioner och införlivas med olika slag av datorsystem, matas ut på skilda sätt för leverans över nät eller på CD-ROM eller och för presentation på alternativa medier: papper, skärm, talsyntes, braille...

När det gäller Strindbergs författarskap förfaller det ändamålsenligt att utforma arkivformatet så att det närmast motsvarar en vetenskaplig tryckt utgåva, innefattande etablerad text och faksimil, kommentar, korshänvisningar, konkordanser etc. Denna samling kan sedan bearbetas, filtreras och matas ut på olika sätt. För ”folkupplagor” kan kommentarer, varianter o.likn. skalas bort. För läromedel kan länkar till handbokstext eller instuderingsuppgifter tillföras. För visuell presentation på internet kan HTML- eller pdf-filer matas ut. För speciella utgåvor för synskadade kan arkivformatets ICADD-attribut utnyttjas vid utmatning till braille, talsyntes eller stor stil. För bokproduktion används den del av märkningen som har typografisk relevans.

Det finns ingen anledning att hårdra motsättningen mellan digital litteratur och litteratur på papper. För egen del föredrar jag att betrakta litteratur på papper som ett sätt — utan tvivel det viktigaste men ändå ett bland andra — att presentera digital litteratur. Att digitala utgåvor eller digitala arkiv är obundna av papprets begränsningar utesluter inte att papper används som medium vid utmatning. Det är ju så papperslitteratur nu för tiden framställs, från digitala dokument via sättningssystem till tryckning. Men det finns väsentliga skillnader vad gäller arkivformat.

I traditionell bokproduktion är standardutgåvan den som utkommit av trycket. Alla digitala versioner blir mer eller mindre överspelade i och med att boken färdigställts. De filer som härrör från layoutprogrammet eller sättningsprogrammet är bundna till en specifik version av en specifik programvara på en specifik plattform med specifika teckenuppsättningar och blir svåra att återanvända så snart man byter systemmiljö eller t.o.m. uppgraderar till en ny programversion. Åtskilliga sent tillkomna ändringar och redaktionella arrangemang brukar brukar saknas i de filer som bevaras. Somliga ingrepp som åstadkommits manuellt genom att sätteriet klippt och klistrat kan helt enkelt inte sparas i sättningsfilerna. Möjligen kan man för framtiden bevara PostScript-filer som använts vid tryckningen, men dessa är slutprodukter som svårligen låter sig återanvändas för andra ändamål än för ett oförändrat nytryck. Den som vill vara på den säkra sidan konsulterar den tryckta utgåvan. Det är boken som är akivformatet.

Däremot har digital litteratur arkivformat som är obundet av papperspresentationen. Arkivformaten hos ett Strindbergarkiv skall väljas så att samlingarna kan tjäna som eller ligga till grund för en digital standardutgåva. Man måste kunna lita på, hänvisa till och citera därur på ungefärligen samma sätt som man använder en vetenskaplig bokutgåva eller en arkivsamling. Det bör också kunna användas för inmatning i samband med produktion av utgåvor på papper.

Det innebär självfallet inte att ett digitalt Strindbergarkiv skulle sätta de existerande kritiska pappersutgåvorna av Strindbergs författarskap ur spel. Brevutgåvan har i det närmaste fullbordats och serien Saml. Verk är färdig till tre femtedelar. Att dessa två utgåvor under många år kommer att förbli standardutgåvor är ett sakförhållande som måste beaktas när det digitala arkivformatet utformas. Märkningen måste i alla viktiga avseenden fånga in relevanta aspekter av pappersutgåvorna, inklusive vissa typografiska egenskaper såsom sidbrytning, helst även radbrytning, samt kursiver, fetstil, stycken av olika karaktär, angivande av korshänvisningar etc. Vidare bör texten i det digitala arkivformatet slaviskt återge den etablerade texten i bokutgåvorna, även eventuella felaktigheter (som dock läsaren inte behöver se om han eller hon inte vill). Med andra ord: vid behov av rättelser eller kompletteringar skall bokutgåvornas ursprungliga text bevaras jämsides med nytillskottet. Den som använder en digital version skall inte behöva ha bokutgåvorna till hands för att identifiera en bestämd passage i dessa, och omvänt skall en läsare med tillgång till böckerna kunna göra fullt bruk av Strindbergarkivets digitala material.

4.1.1 Text

Det i dag mest självklara arkivformatet för text är en märkning som följer den internationella standarden SGML (Standard Generalized Markup Language, ISO 8879:1986) i enlighet med riktlinjerna från TEI (Text Encoding Initiative), vilket blivit en norm för kvalificerad humanistiskt digital litteratur. Det finns i dag ingen konkurrerande standard, och om en sådan skulle uppstå i framtiden torde SGML-märkningen erbjuda bästa tänkbara förutsättningar till konvertering.

TEIs riktlinjer föreslår ett stort antal märkord (ca 400) och attribut, varav en mindre del kommer att behövas i fallet Strindberg. Därför är ett viktigt förarbete inför framtida digitala Strindbergutgåvor att avgränsa en lämplig delmängd av alla dessa märkord och attribut. Det existerar en bantad version av TEIs DTD, kallad ”TEI Lite”, som är tänkt att täcka de flesta behov, men det är en öppen fråga om just denna delmängd är lämpad för Strindbergs verk.

I ett avseende tror jag att man redan från början bör göra en avgränsning. Man bör avstå från de märkord och attribut som är avsedda för lingvistisk analys. En märkning som identifierar varenda satsdel, böjningsform, stilvalör etc skulle bli mer omfattande än den egentliga texten (som är fallet med t.ex. den gigantiska samlingen British National Corpus), inskrivningsarbetet skulle bli synnerligen krävande, och dessutom skulle det bli omständligt att revidera och hantera materialet om varenda ord vore inkapslat i märkord med vidhängande komplicerade attribut. Jag föreslår alltså att lingvisternas behov tills vidare lämnas därhän och att arkivformatet primärt skall vara till nytta för användare med litterära eller historiska intressen. Detta innebär förstås inget slutgiltigt ställningstagande, lingvistisk märkning kan tillföras när resurser och behov föreligger.

Sannolikt kommer TEIs riktlinjer att behöva revideras eller kompletteras för att passa Strindbergmaterialet, vilket är förhållandevis enkelt tack vare att TEIs DTD är modulärt uppbyggd

och lätt att utöka med egendefinierade tillägg. Erfarenheten visar dock att TEIs riktlinjer räcker långt. Det finns all anledning att vänta med hemsnickrade lösningar intill dess att man uttömt de möjligheter som TEIs riktlinjer erbjuder.

Man kan även tänka sig att arkivformatet innehåller vissa systemberoende egenskaper, exempelvis länkar som hänvisar till det lokala filsystemet eller vissa systemberoende entitetsdefinitioner. I så fall krävs konvertering innan materialet kan exporteras till andra system. Förmodligen kommer inte all information att kunna exporteras, vilket inte behöver innebära någon katastrof. Det är snarare regel än undantag att kvalificerade digitala arkiv för eget bruk håller sig med ett arkivformat som är rikhaltigare i jämförelse med det material som levereras till andra system. Men självfallet bör informationen vara så flyttbar som någonsin är möjligt.

Beträffande filsystem är det nog klokt att redan från början så långt som möjligt eftersträva flyttbara lösningar som är oberoende av filsystem. Man skall inte för tid och evighet behöva bestämma om den etablerade versionen av Rödas Rummet skall utgöra en enda stor fil eller fördelas på en fil per kapitel, eller kanske bestå av informationsfragment lagrade i en databas. Filsystemen orsakar alltid problem eftersom de skiljer sig åt för olika plattformar, och det är osäkert vad en fil kommer att vara i framtiden. Därför bygger SGML-standarden inte på filer utan i stället på ”entiteter”, som kan vara en fil men som även kan vara något annat, t ex en samling filer eller informationsfragment förbundna med pekare.

Inte minst viktig är den bibliografiska informationen eller ”metainformationen”, avsedd bl.a. för automatisk katalogisering och som underlag för bibliografiskt arbete. Enligt TEIs riktlinjer måste varje TEI-anpassat dokument inledas med en ”TEI header” med sådan information, och arbetet med att anpassning till relevanta standarder inom arkiv- och biblioteksvärlden fortsätter.[3] Det finns anledning att pröva möjligheten att utnyttja TEI headers som behållare för metainformation även rörande material som inte är i SGML-format. Detta kan vara ett sätt för ett digitalt Strindbergarkiv att härbärgera och leverera material som (ännu inte) överförts till arkivformat.

De texter som etablerats i Brev och Saml. Verk kommer mänskligt att döma att under åtskilliga decennier utgöra kärnan av textmaterialet i ett digitalt Strindbergarkiv. Här finns digitala versioner att utgå från. Kerstin Dahlbäck har ombesörjt en digital version av hela brevutgåvan (exkl. kommentarerna) och även outgivna brev, och tillfogat viss märkning. Med hjälp av sätteriet och Språkdata i Göteborg har vi vid KTH fått tillgång till maskinläsbara versioner av de hittills utgivna volymerna (dock i vissa fall enbart textdelen, inte kommentar och ordförklaringar) i Saml. Verk. Vi har försökt spåra och korrigera fel, översatt sättningskoderna till mer flyttbart format och försett texten med grov SGML-märkning, men många felaktigheter återstår. Här krävs korrekturläsning som jämför de digitala versionerna med bokutgåvornas definitiva text. Problemet är som nämnts att i ordinär bokproduktion blir utgåvan på papper den enda slutgiltiga versionen. Ofta sägs att den ena eller andra texten existerar i ”maskinläsbar” version, men den kan ändå vara bemängd med fel och oklarheter, i synnerhet i de fall när den knappast redigerats utan skapats genom skanning eller genom enkel konvertering från sättningsfilerna eller från de ordbehandlingsfiler som matats in bokproduktionen.

4.1.2 Bilder

Vad gäller bilder är det viktigaste att finna former för fångst och lagring av Strindbergs handskrivna sidor. Detta är självfallet väsentligt inte bara med tanke på digital publicering utan även för att bevara faksimil av original med begränsad livslängd. Enligt Lars Dahlbäck är visserligen merparten av Strindbergs efterlämnade papper i någorlunda gott skick, men där förekommer (liksom i bl.a. Bonniers korrespondens till Strindberg) kopiebokskopior vars skrift bleknar, och även väl bevarade papper kan skadas av nötning eller råka ut för andra olyckor.

[3] Bl.a pågår ett projekt som syftar till koordinering av TEI och MARC, se URL http://www.columbia.edu/cu/lib…/sgmlmarc/davis.9603.text.html. Tanken är att metainformation så långt som möjligt skall kunna flyttas fram och åter mellan TEI header-formatet och MARC-format, för att stödja t.ex. automatisk katalogisering av digitala dokument.

Det har funnits delade meningar om de bästa teknikerna för sådan digitalisering, men på sistone tycks specialisterna inom arkivvärlden ha blivit någorlunda eniga om att digital kamera är att föredra — förutsatt att tiden, pengarna och de personella resurser räcker till — och att man skall utnyttja så hög upplösning och framför allt så stort färgdjup som över huvud taget är möjligt. Erfarenheten visar att färgdjupet (dvs mängden färginformation knuten till varje pixel) i många hänseenden är viktigare än upplösningen (det totala antalet pixlar per ytenhet). Även om de nu tillgängliga presentationsteknikerna och bildskärmarna inte utnyttjar den bildkvalitet som de bästa kamerorna åstadkommer, har dagens digitalkamera-teknik nu nått så långt att den förefaller erbjuda ett säkert sätt att lagra materialet för framtida behov.

Alternativa billigare och snabbare tekniker (skanning av originalet, eller mikrofilmning med åtföljande skanning) ger resultat som kan duga gott för CD-ROM produktioner o. likn., där dagens begränsade lagringskapacitet ändå inte ger plats för stora bildfiler och där presentationsfunktionerna anpassats till den för tillfället dominerande skärmtekniken, men kraven måste ställas högre på arkiv som skall överlämnas till kommande generationer.

Lagring av digitaliserade bilder bör ske i ett format som utan någon som helst informationsförlust kan konverteras till andra relevanta format. Det är inget som hindrar att arkivformatet utnyttjar komprimeringsalgoritmer som inte gör annat än minskar redundansen. Däremot bör man undvika sådana som faktiskt förstör informationen, exempelvis tekniker som i likhet med JPEG avlägsnar information som är oväsentlig för det mänskliga ögat, ty det obeväpnade ögat har sina begränsningar. Även inom det synliga spektrumet varierar dess känslighet. I framtiden kommer säkert Strindbergs manuskript att utforskas med hjälp av verktyg som är mycket mer sofistikerade än dagens.

4.1.3 Länkar

Beträffande länkar är frågan mer öppen. Det finns olika slag av länkar att ta hänsyn till, och ingen etablerad dominerande standard.

Man bör man skilja mellan å ena sidan länkar som så att säga hårdlödes inne i själva materialet, å andra sidan länkar som lagras separat. Med andra ord kan man antingen inne i textmassorna infoga länkar som knyter samman, låt säga, en Strindbergtext med förarbeten, eller också kan man samla länkinformationen för sig. Det senare alternativet — som har fördelen att själva texterna inte korrumperas och att det kan vara lättare att reorganisera och underhålla länkinformationen — är möjligt tack vare HyTime (Hypermedia/Time-based Structuring Language, ISO 10744:1992), en SGML-baserad standard för bl.a. länkning.

Även TEI har lanserat en egen kraftfull och kompakt länkningsmekanism, ”TEI extented pointers”, kallade”tei links”, som är värd att beakta. I dag finns dock inte särskilt mycket programvara som stödjer dessa länkar.

Under alla omständigheter bör man hålla isär två olika frågor, dels frågan om länkarnas ankarfästen (dvs. punkter eller regioner som de förbinder), dels frågan om själv länkningsfunktionen. När man exempelvis märkt början och slutet av det textparti i en bastext som avviker från en varianttext förfogar man över ett ankarfäste. Länkfunktionerna som gör bruk av ankarfästena bör utformas så att de vid behov lätt kan revideras, vilket förenklas om länkinformationen sparas separerad från de sammanlänkade dokumenten. HyTime-standarden möjliggör till och med att man länkar samman dokument som s.a.s inte vet om att de är länkade till varandra. I så fall behöver man inte ens märka ut ankarfästen inne i dokumentet; HyTime-länken håller själv reda på att i detta dokument har jag mitt ankarfäste i fjärde kapitlet, tredje stycket, etc.

Den egentliga SGML-standarden (ISO 8879) saknar mekanismer för länkning mellan dokument. Däremot erbjuder SGML en mekanism för länkning inom ett dokument, men det är nog klokt att så långt som möjligt bör undvika den lösningen. Begrepp som dokument eller fil kan komma att bli obsoleta i takt med att system utvecklas som snarare organiserar informationsobjekt (eller med SGMLs terminologi: entiteter), vilka ibland är en fil, ibland ingår i en fil, ibland är spridda över många filer. (Vilket är rimligt även från en redaktionell utgångspunkt: det förefaller fel att en gång för alla avgöra om, låt säga, en dramautgåva skall fördelas på en fil för varje scen, en för varje akt, en för varje pjäs eller en för hela utgåvan.)

Det slags adressering och länkning som är beroende av filernas fysiska lokalisering — exempelvis HTTP-länkar så som dessa används på WWW i dag — är självfallet oanvändbar som arkivformat. Sådana länkar förstörs så snart filerna döps om eller flyttas från en katalog till en annan eller från en server till en annan. Men det pågår, även i HTML-världen, ett intensivt arbete med adresserings- och länkningsmekanismer vilket förhoppningsvis så småningom kommer att resultera i flyttbara standarder.

Länkar till och från bilder är givetvis viktiga, t.ex. sådana som förbinder en transkriberad text med faksimil av handskrivna sidor. Här tror jag att man till vidare måste nöja sig med att låta en bild i dess helhet bli ankarfäste för länken. Det vore visserligen ur många avseenden en fördel att kunna definiera ankarfästen som utgör punkter eller regioner inne i bilder, men för sådana mekanismer existerar ännu ingen dominerande standard. Det torde i dag inte finnas någon annan lösning än att lagra varje bild, exempelvis ett faksimil av varje manuskriptsida, som en egen fil och att låta länkar hänvisa till denna fil i dess helhet. Från varje sida i exempelvis ett romanmanuskript kunde man då skapa länkar som pekar till motsvarande region i den transkriberade texten, för att ge läsaren möjlighet att öppna ett fönster som visar den aktuella manuskriptsidan; medan det tyvärr inte är genomförbart att i arkivformatet förbereda för tillämpningar som länkar ett bestämt ställe på manuskriptsidan, låt säga en rad, till motsvarande region i den transkriberade texten. (Dock finns förstås inget som hindrar att materialet när det väl hämtas från arkivformatet bearbetas vidare för presentation i något specifikt system som utnyttjar koordinater inuti bilder och som exempelvis i två fönster intill varandra låter den transkriberade texten och det avbildade manuskriptet följas åt när läsaren förflyttar sig i ettdera.)

Även om det är svårt att förutse framtidens länkningsstandarder, är det viktigt att redan nu välja ett arkivformat för länkningsinformation som ger goda utsikter för konvertering till kommande standarder.

4.2 Utbytesformat

Från arkivformaten skall man, helst automatiskt, kunna generera material i ett begränsat antal utbytesformat avpassade för mottagarens behov och system.

Enklast vore förstås om arkivformatet vore sådant att materialet kunde levereras i befintligt skick till vissa mottagare, dvs. anpassat till relevanta standarder och med ett innehåll som kan spridas till omvärden. I dessa fall skulle inget särskilt utbytesformat behövas. Det innebär att redaktörerna håller sitt eget arbetsmaterial åtskilt från det offentligt tillgängliga digitalt Strindberg-arkivets samlingar. Efterhand som redaktörena finner att en version är tills vidare definitiv placeras den i det egentliga Strindbergarkivet, vars samlingar kanske är direkt tillgänglig på nätet (ev. via behörighets- och betalningssystem) och fungerar som den samlade officiella textkritiska digitala utgåvan av Strindbergs verk. En sådan organisation skulle innebära att Strindbergarkivets samlingar helt enkelt *är* standardutgåvan, vilken kan ligga till grund för olika slags separatutgåvor.

Men möjligt är att materialet i arkivformatet kommer att innehålla information som det är olämpligt eller onödigt att sprida till omvärlden. Det kan röra sig om information för internt bruk, såsom redaktörernas information om den egna ”ärendehanteringen” eller ofärdigt *work in progress*, eller systemberoende information såsom länkar till dokument eller program vilka bara är tillgängliga inom det lokala nätet. I så fall är det snarast de ur arkivformaten genererade utbytesformaten som definierar standardutgåvan.

En ytterligare komplikation kan vara att det textmaterial som lagras i SGML är utformat så att det i ett eller annat avseende avviker från TEIs riktlinjer eller t.o.m. från SGML-standarden. Då är det sannolikt förnuftigt att så långt som möjligt låta utbytesformat följa TEIs riktlinjer. Det är bl.a. så riktlinjerna är avsedda att användas, för att möjliggöra att material skickas mellan arkiv eller forskningsinstitutioner vilka var för sig har egna interna format.

Dock är det sist nämnda problemet, att den egna institutionen använder ett unikt format som inte omvärlden kan utnyttja, mest kännbart för de arkiv som varit pionjärer och utvecklat sina märkningsrutiner innan det senaste decenniets utveckling av standarder och systemoberoende

lösningar. Ett exempel är märkningsschemat MECS som skapades under 1980-talet i den miljö som 1990 omvandlades till Wittgensteinarkivet vid universitetet i Bergen. Här har man under lång tid transkriberat och märkt Wittgensteins flertalet av efterlämnade papper och föredrar naturligt nog att fortsätta enligt samma principer med de resterande. Dessutom anser man att det egna schemat MECS innebär vissa fördelar i jämförelse med TEI.[4]

Snart sagt alla nystartade humanistiska digitala publiceringsprojekt av någon dignitet strävar i dag efter att i möjligaste mån anpassa formatet för lagrad text till det internationella standardiseringsarbetet, dvs. i praktiken till TEIs riktlinjer. Denna omsvängning märks när nya publiceringsprojekt presenteras på konferenser: den som inte använder TEIs riktlinjer känner sig tvungen att ursäkta sig och förklara varför. (Av praktiska och ekonomiska skäl börjar projektarbetet ofta med inskrivning eller inskanning som resulterar i tämligen sparsamt märkta textsamlingar, men då brukar paperförfattaren alltid försäkra att ”senare tänker vi använda TEI”.)

Det är tänkbart att ”TEI Lite”, dvs. TEIs förenklade DTD, kan duga som utbytesformat i många sammanhang. TEIs ordinära DTD är ganska trög att använda eftersom den består av många beståndsdelar av vilka flertalet inte användes men som ändå måste läsas in för att mottagaren skall kunna använda sin editor eller sitt presentationsverktyg.

Det är svårt att förutse de utbytesformat som kan komma att krävas i framtiden. Med dagens teknik förefaller det t.ex. rimligt att bilder lagras i bästa tänkbara kvalitet men utbytes med lägre upplösning och färgdjup och kanske komprimerade och konverterade — allt för att garantera en någorlunda snabb nätöverföring och för att de skall kunna hanteras av mottagarens system. Ett inför framtiden intressant alternativ erbjuder de nya tekniker som innan överföringen registrerar den mottagande datorns tekniska förutsättningar och därefter över nätet levererar exakt det mått av pixelinformation som grafikkortet och skärmen klarar av. Då skulle Strindbergarkivet kunna innehålla högklassiga bilder vilkas information vid nätöverföring reduceras så mycket som krävs för att passa de tekniska förutsättningarna hos på mottagarsidan.

[4] Klaus Hvitfeldt, upphovsmannen till MECS, har uppmärksammat flera i samband med märkning av Wittgensteinmaterialet besvärande begränsningar hos TEI i dess nuvarande utformning. TEI (och SGML) fungerar bäst vid strikt hierarkiskt uppbyggda texter. I Wittgensteins egensinniga manuskript bryts ofta den hierarkiska ordningen: en rubrik kan uppträda mitt inne i ett stycke, kanske en som författaren plötsligt kom att tänka på kunde passa för hela kapitlet. Där finns många överlappande element — t. ex. till varandra relaterade stycken som uppträder här och där längs olika grenar av den hierarkiska strukturen, eller citat som börjar i slutet av ett stycke och fortsätter en bit in i nästa stycke —, vilka helst skall inkapslas av samma märkord. Vidare förutsätter TEI (och SGML) en på förhand definierad DTD, som visserligen inte i alla avseenden behöver följa TEIs riktlinjer och som förstås kan ändras med tiden, men varje större ändring av DTDn är en ganska stor sak som i allra värsta fall kan kräva revision av tidigare märkt material. Det är inte möjligt att mitt under pågående märkning helt enkelt uppfinna och stoppa in ett nytt märkord. För Wittgensteinarkivets behov erbjuder därför MECS vissa fördelar genom att vara mindre hierakiskt orienterat, genom att medge överlappande element och genom att tillåta att man under arbetets gång tillför nya märkord utan hänsyn till någon DTD. I gengäld saknar förstås MECS många av TEIs fördelar, vad gäller både märkningens precision och flyttbarheten.

Här finns flera vägar att gå. För det första finns vissa generella möjligheter att hantera sådant som överlappande element i SGML-tillämpningar, vilket dock förutsätter en hel del programvaruutveckling. Det finns t.ex. än så länge (och så kanske det förblir) knappast några tillämpningar som stöder CONCUR, den SGML-mekanism med vars hjälp man kan knyta flera DTDer till ett och samma dokument. Man kan i stället laborera med ”marked sections”, men det ett tungrott förfarande som innebär att man måste underhålla parallella textpartier med olika märkning. För det andra finns möjlighet att påverka den fortsatta utvecklingen av TEI (just Wittgensteinarkivets problem anförs ofta i den pågående diskussionen inom TEI). För det tredje finns förstås möjligheten att förfina och försöka sprida särskild programvara som stödjer sådant som SGML förbjuder. (Den till MECS knutna programvaran är av det slaget men den är inte spridd utanför Bergen. Forskare som vill arbeta mer intensivt med samlingarna måste bege sig till Bergen och sätta sig ned vid arbetsstationerna där. Mer praktiskt vore om materialet i grunden vore SGML-märkt och i princip tillgängligt för alla, men kanske kompletterat med icke flyttbar märkning som kräver särskilda installationer.) För det fjärde kan man givetvis göra TEIs DTD mer tillåtande, men om man löser upp denna alltför mycket blir märkningen föga användbar; det vore att gå alldeles för långt att tillåta rubriker mitt inne i ett stycke. Och för det femte kan man om så krävs lagra materialet i ett eget unikt format och använda TEI som utbytesformat. Det är den sistnämnda vägen som Wittgensteinarkivet tills vidare bestämt sig för.

4.3 Presentationsformat

Även om gränsen mellan utbytesformat och presentationsformat är suddig, tror jag att det är en nyttig distinktion. Presentationsformaten kan vara av otaliga slag och ständigt föränderliga. Framtidens presentationsformat kan vi inte ens föreställa oss. Ett digitalt Strindbergarkiv har inte i uppgift att föreskriva vilka presentationsformat som skall användas. Förlag eller forskningsinstitutioner kommer att utforma egna lösningar för publicering på papper, över nät, på CD-ROM och så småningom på i dag okända medier. I många fall kommer läsarna att föredra egna bearbetnings- och presentationsverktyg och i stället för en prefabricerad produkt önska sig ett ”råmaterial” som kan stoppas in en *viewer*, ett redigeringsprogram, ett ordbehandlingsprogram, ett layoutprogram, eller speciella program för lingvistisk eller tematisk analys. I så fall avgöras presentationsformatet av den programvara användaren förfogar över.

Om textmaterialet lagrats i SGML kan läsare med tillgång till SGML-baserade presentations- eller bearbetningsverktyg hämta materialet som det är, i arkivformatet eller kanske ett utbytesformat anpassat till TEI Lite. Man kan således tänka sig att materialet publiceras helt enkelt genom att läggas upp på en internetserver för nedladdning över nätet som okompilerade SGML-filer, vilka läsarna får hantera efter behag. Samtidigt som läsarna laddar ned materialet, kanske inifrån sin webbrowser, kan de erbjudas att hämta hem programvara (i första hand en viewer med utskriftsfunktion) och annat (DTD, SGML-deklaration, stilmallar, länkstrukturer) som de kan behöva.

Ett alternativ är att kompilera SGML-materialet och göra en färdig ”elektronisk bok” därav. Det är vanligen så dagens kvalificerade digitala humanistiska litteratur publiceras av förlagen. Läsaren ska inte behöva se eller bekymra sig om SGML-märkningen, som enbart användes för att styra layout, sökningsfunktioner etc. Ur läsarens synvinkel är nackdelen är att materialet är instängt i en speciell tillämpning och därmed svårt att använda för andra syften än dem producenten tänkt sig. Förlagen tycks uppfatta denna begränsning som en fördel. Men det ena behöver inte utesluta det andra. Så länge ursprungsmaterialet är i SGML kan man både publicera skräddarsydda elektroniska böcker och samtidigt behålla ”originalet” i ett flyttbart arkivformat. Det är så exempelvis den av Peter Robinson redigerade *Wife of Bath’s Prologue* (Cambridge University Press, 1996), den till dags dato förmodligen mest avancerade textkritiska utgåvan på CD-ROM, producerats.

Ett tredje alternativ är utmatning från arkivformatet till HTML för publicering på World Wide Web. En hel del information går därmed förlorad och det är svårt att hitta sätt att arrangera mer komplexa och omfattande texter med hjälp av HTML, som dock är ett så spritt märkspråk att denna möjlighet bör erbjudas åtminstone för visst textmaterial (dikter eller kortare prostycken, men inte hela romaner) som lämpar sig för websidor. Tillsammans med HTML-dokumenten kan faksimil av manuskript och andra bilder levererades konverterade till formatet GIF, som nästan alla webbrowsers kan visa, eller JPEG som många kan visa.

Det är viktigt att hålla i minnet att WWW-publicering inte behöver begränsas till HTML och GIF eller JPEG. Det finns många tilläggsprogram som låter användaren bruka sin webbrowser som ett sökverktyg varefter en annan tillämpning ta över när materialet väl är funnet. Man kan därmed publicera det mest skilda material, som inte alls behöver vara SGML-märkt, genom att göra det åtkomligt från en WWW-server.

Den i dag närmast till hands liggande lösningen vore att göra Strindbergmaterial tillgängligt på en server som besökaren hittar med hjälp av Netscape eller någon annan webbrowser. Här ser besökaren en välkomstsida med instruktioner och länkar till andra sidor med olika nedladdningsalternativ. Somligt material kan matas ut som HTML och visas i Netscapes ordinära fönster. Om besökaren i stället väljer att läsa in en SGML-fil startar automatiskt en SGML-viewer (som besökaren ges tillfälle att ladda ned om han eller hon saknar en sådan). Nedladdade bilder eller material med fixerat typografiskt utseende kan aktivera tilläggsprogram för olika grafikformat eller för pdf. Eller också kan besökaren från en meny välja att ladda ned filer i andra format (PostScript, råa textfiler, rtf, Ms Word…) för fortsatt hantering i det egna lokala systemet. I alla dessa fall är det självfallet önskvärt, men kanske inte alltid realiserbart, att varje slag av information (text, bilder, länkinformation) lagras i ett och endast ett format, arkivformatet, varifrån det via filter matas ut när det efterfrågas. Annars

tvingas man dels lägga ned tid på manuellt konvertering och redigering, dels underhålla ett antal parallella uppsättningar av samma informationsmängd med åtföljande dubbelarbete och felrisker.

Man kan också i de fall där formgivingsaspekterna är centrala tänka sig utmatning till formatet pdf som ger en flyttbar typografi. Därmed skulle läsaren på skärmen eller utskrift erhålla ett typografiskt arrangemang och en bildpresentation som ungefärligen kunde motsvara, låt säga, en boksida i Saml. verk eller Brev. Ytterligare ett alternativ om man inte är intresserad av återanvändning är leverans av PostScript-filer som troget återger en tryckt boksida.

Sedan tillkommer förstås många slags presentationsformat som kan krävas för typografering av en bokutgåva, eller för framställning av en i formgivningshänseende mer genomarbetad digital utgåva etc. Detta är dock senare frågor. Våra egna erfarenheter av deltagande i publiceringen av ett par Strindbergvolymer i bokform tyder på att det är olyckligt att alltför mycket blanda ihop å ena sidan arkivformat, å andra sidan specifika typografiska krav för en viss utgåva — samtidigt som arkivformatet självfallet bör utformas på så sätt att det på bästa sätt kan utnyttjas i samband med t.ex. framställningen av tryckta utgåvor eller multimediaprodukter.

Jag kommer i huvudsak att uppehålla mig vid arkivformaten, och i synnerhet format för lagring av text.

5. Om beskrivande märkning och SGML, kort introduktion

Detta avsnitt[5], hämtat ur ett annat sammanhang, finns med här som kort introduktion för den som är obekant med SGML.

SGML (Standard Generalized Markup Language) är ett språk för beskrivande märkning av maskinläsbara dokument. Jag skall inledningsvis med ett par exempel illustrera skillnaden mellan procedurmärkning och beskrivande märkning.

Procedurmärkning talar om hur den märkta informationen skall hanteras, exempelvis vad datorn och skrivaren eller fotosättningsmaskinen skall göra med en textfil. Var och en som infogat sättningsanvisningar i ett manuskript vet vad det innebär. En viss anvisning kan ange att vissa ord skall tryckas i 12 p kursiv stil. I detta fall innebär procedurmärkningen en instruktion till sätteriet.

En beskrivande märkning ger i stället besked om av vilka element texten består. Några ord med ett ett och samma typografiska utseende, t.ex. ord satta med 12 p kursiv, kan utgöra vitt skilda slag av element, exempelvis 1. en rubrik, 2. något författaren vill framhäva, 3. ett insprängt ord på främmande språk, 4. en boktitel. Om vi tillämpar procedurmärkning skulle i följande fiktiva text samtliga de fyra nämnda slagen av element anges på samma sätt (12 p Courier kursiv):

> *II.3.1. Om Descartes metod*
>
> Descartes har skapat ett mönster för vad filosofisk *metod* vill säga. Men i själva verket är Descartes berömda text ett *préface* till tre naturvetenskapliga och matematiska studier (tillgängliga i René Descartes: *Discours de la méthode*. Texte et commentaire par Étienne Gilson. Paris: Vrin, 6 uppl. 1987).

Med beskrivande märkning, som skiljer på dessa fyra slag av innehållsliga enheter, undviker man att en och samma märkning betyder olika saker. Om vi tillämpar SGML-märkning enligt version 1 av TEIs riktlinjer (mer därom strax) skulle märkningen se ut på följande sätt. Rubriknivå nr 3 markeras med ett inledande märkord `<h3>` och avslutas med märkordet `</h3>`. (Observera att ingen numrering av avsnittsrubrikerna behöver sättas ut; avsnittets plats i dokumentet gör att vi ändå vet att det rör sig om andra kapitlet, tredje avsnittet, första underavsnittet). Ett stycke inleds med `<p>` och avslutas med `</p>`. Framhävningen avgränsas av märkorden `<em>` och `</em>`. Bokstäverna em skall utläsas ”emfas”. (Observera att framhävning ofta representeras typografiskt av kursiv stil, men den kan även representeras på annat sätt, förr i världen med spärrad stil, i dag ibland med rak stil, exempelvis i ett förord där brödtexten är satt med kursiv stil.) Ett insprängt ord på främmande språk, här franska, kan markeras med de inledande märkorden `<gloss><foreign lang=Fr>` och avslutas med märkorden `</foreign></gloss>`. En hänvisning till en boktitel kan ske med korsreferens till ett unikt ställe; om den fullständiga referensen till Descartesutgåvan återfinns i en litteraturlista i slutet av det aktuella dokumentet och där försetts med märkordet `DES87` räcker det med att i slutet av texten ovan infoga `<xref RID=DES87>`. (`xref` uttydes korsreferens, `RID` uttydes referensidentifikation). Så här ser ser texten ut då den på detta sätt försetts med beskrivande märkord:

```
<h3>Om Descartes metod</h3>
<p>Descartes har skapat ett mönster för vad filosofisk
<em>metod</em> vill säga. Men i själva verket är Descartes berömda
text ett <gloss><foreign lang=Fr>préface</foreign></gloss> till tre
```

[5] Avsnittet är i huvudsak hämtat ur D. Broady, a.a., 1993.

```
naturvetenskapliga och matematiska studier (tillgängliga i <xref
RID=DES87>).
```

Detta var ett exempel på att man med hjälp av beskrivande märkning undviker att ett och samma typografiska utseende (kursiv stil) betecknar olika innehållsliga element. Omvänt undviker man också att ett och samma slag av innehållslig enhet märks på olika sätt. Tag citat som exempel. Citat kan utmärkas typografiskt på en rad sätt. Redan citattecknens utseende och placering varierar:

Tyska	»XXXXX«	Franska	« XXXXX »
	›XXXXX‹		
	„XXXXX“	Svenska	”XXXXX”
			»XXXXX»
Engelska	“XXXXX”		
	‘XXXXX’		

Även svenska typografiska konventioner kan variera. Ett citat som utgör ett eget stycke markeras ibland med indragen vänstermarginal, ibland med minskad grad och ibland med citattecken. Tag följande citat ur Viktor Rydbergs *Bibelns lära om Kristus*:

> Det andra af de båda föregifna intygen lemnas af Cypriani skrift “Om kyrkans enhet“ och har följande lydelse:
> *“Herren säger: jag och fadren äro ett. Och återigen är det skrifvet om fadren och sonen och den helige andre: och tre äro ett.“*
> Onekligen har detta intyg ett viss sken för sig. Tvenne olika bibelställen äro här åsyftade. Det ena är, liksom hos Tertullianus, Joh. 10, 30; det andra är utan tvifvel 1 Joh. 5, 8.

som med moderna typografiska konventioner skulle kunna se ut på flera sätt, exempelvis:

> Det andra af de båda föregivna intygen lämnas af Cypriani skrift ”Om kyrkans enhet” och har följande lydelse:
> ”Herren säger: jag och fadren är ett. Och återigen är det skrivet om fadern och sonen och den helige andre: och tre är ett.”
> Onekligen har detta intyg ett viss sken för sig. Tvenne olika bibelställen är här åsyftade. Det ena är, liksom hos Tertullianus, Joh. 10, 30; det andra är utan tvivel 1 Joh. 5, 8.

eller:

> Det andra af de båda föregivna intygen lämnas af Cypriani skrift ”Om kyrkans enhet” och har följande lydelse:
>
> > Herren säger: jag och fadren är ett. Och återigen är det skrivet om fadern och sonen och den helige andre: och tre är ett.
>
> Onekligen har detta intyg ett viss sken för sig. Tvenne olika bibelställen är här åsyftade. Det ena är, liksom hos Tertullianus, Joh. 10, 30; det andra är utan tvivel 1 Joh. 5, 8.

I franskt tryck är sägesatserna inte sällan infogade innanför citattecknen, som här på ett ställe i Claude Lévi Strauss *La pensée Sauvage*:

> « Ce procédé, dit Boas, y est plus fréquent que dans tout autre langage connu de moi. »

och i engelskt tryck förekommer ofta att ett kommatecken omedelbart efter citatet placeras före citattecknet, som här i Paul Feyerabends *Against method*:

> ‘If any metaphysics goes,’ writes Dr Hesse in her review of an earlier essay of mine, ‘then the question arises [...] ’

Vid användning av beskrivande märkning bryr man sig i princip inte om sådana närmast estetiska spörsmål, man nöjer sig med att sätta märkordet <q> före citatet och </q> efter citatet. Sedan ankommer det på hur man styr det typografiska formateringsprogrammet och tillgängliga utskriftsmöjligheter hur den tryckta texten gestaltas.

De exempel på märkning jag här givit ansluter till den internationella standarden för rent beskrivande märkning, SGML (Standard Generalized Markup Language, ISO 8879), som antogs i oktober 1986. Standardverket är Charles F. Goldfarb, *The SGML Handbook*, Clarendon Press, Oxford 1990, som bl.a. innehåller den fullständiga ISO-texten. Arbetet går längre tillbaka i tiden. Ett förstadium till SGML var märkspråket GML, som Charles Goldfarb m.fl. från och med slutet av 1960-talet utvecklade vid IBM. Tanken var då att dokument (det rörde sig om juridiska dokument) skulle kunna märkas på ett enhetligt sätt, så att ett och samma dokument kunde matas in i skilda system för textbehandling, formatering och informationsåtervinning.

Dagens teknik öppnar vissa möjligheter att bearbeta ett och samma dokument eller en och samma dokumentsamling på mångahanda sätt och med många slag av program på många slag av maskiner. En sådan utveckling förutsätter ett generaliserat dokumentbeskrivningsspråk, ett slags esperanto om man så vill, som gör dokumenten oberoende av maskinvara och programvara och nationell teckenuppsättning. Ett och samma dokument kan med andra ord stoppas in i en UNIX-maskin, en Macintosh eller en DOS/Windows-maskin och bearbetas med olika program. På detta område pågår ett livaktigt internationellt arbete med utvecklingen av SGML och med en hel svit av ”dotterstandarder” såsom DSSSL och HyTime.

DSSSL (Document Style Sematics and Specification Language, ISO 10179:1996; akronymen uttalas vanligen ”dissel”) styr den typografiska formgivningen, dvs. sörjer för att layouten blir någorlunda densamma vid utskrift på olika medier och med olika utrustning. DSSSL förutsätter att dokumenten är SGML-märkta. DSSSL antogs som Draft International Standard i augusti 1991 och som internationell standard i januari 1996. Ännu finns inga program som stöder DSSSL, men med tiden kan sådana kan komma att bli viktiga för presentation på skärm eller papper.

HyTime (ISO 10744:1992), som antogs i april 1992 efter en rekordkort förberedelsetid eftersom behovet ansågs akut, är den första internationella standarden för överföring av hypertextdokument och multimediadokument (eller för att vara mer exakt: tidsberoende dokument, innehållande sådant som ljud och rörliga bilder). Även HyTime bygger på SGML-standarden, och även i HYTimes fall finns få program som kan utnyttja standarden (vilket är normalt, det tar alltid tid innan programvaruutvecklingen hunnit i fatt standardiseringen).

Detta standardiseringsarbete öppnar nya möjligheter att överskrida de gränser — nationsgränser, skillnader mellan olika slag av datorer och program, gränser mellan utbildningsväsendets nivåer och ämnen och mellan forskning och undervisning, avstånden mellan datorentusiasterna och de övriga — som i dag förhindrar samlade ansträngningar att skapa och sprida rikhaltiga dokumentbaser av god kvalitet.

Redan möjligheten att komma förbi de till synes triviala men enerverande problemen med olika teckenuppsättningar är ett stort framsteg. I SGML-standarden (dvs ISO 8879) ingår en *reference concrete syntax*, som är utgångspunkten för översättningar mellan olika datormiljöer, nationalspråk, teckenuppsättningar m.m. I denna syntax ingår en teckenuppsättning (kallad *the base character set*) som överenstämmer med standarden ISO 646 (dvs en 7-bit-standard känd som IRV = International Reference Version). Detta är det som i egentlig mening är ASCII-teckenuppsättningen, dvs tecknen 0-127, som ligger till grund för en mängd andra internationella teckenstandarder. Med de 128 tecken som däri ingår och som nästan alla maskiner förstår kan snart sagt varje tecken i de europeiska skriftspråken representeras. Den mest flyttbara flyttbara representationen av å,ä,ö är enligt regelboken (Charles F. Goldfarb, *The SGML Handbook*, 1990, p. 506f):

Å	översätts med `&Aring;`
å	översätts med `&aring;`

Ä översätts med `&Auml;`
ä översätts med `&auml;`
Ö översätts med `&Ouml;`
ö översätts med `&ouml;`

Det pågår arbete med att standardisera diverse mer exotiska teckenuppsättningar, däribland runor. Men i många sammanhang — t.ex. Strindbergs egensinniga tecken och krumelurer, eller utgivarnas typografiska arrangemang i form av stjärnor eller streck som avdelar avsnitt — blir man tvungen att att definiera skräddarsydda entiter som fungerar som platshållare för att markera var en en sträng eller en tecknad bild skall infogas.

Ett SGML-dokument består av tre delar. För det första en SGML-deklaration som talar om på vilket sätt man byggt ut eller modifierat ISO 8879. (Exempelvis anger TEIs riktlinjer att märkorden får vara upp till 128 tecken långa, vilket innebär en avvikelse från ISO 8879 som föreskriver högst 8 tecken). För det andra en DTD (Document Type Definition) som anger hur föreliggande typ av dokument märkes. En DTD kan avse affärsbrev, ytterligare en annan romaner, ytterligare en annan vetenskapliga monografier etc. För det tredje själva innehållet i dokumentet. Det är inte nödvändigt att SGML-deklaration och DTD medföljer varje enskilt dokument, men de måste (som regel som egna filer dit pekare i dokumentet hänvisar) vara tillgängliga någonstans i det system där man handskas med dokumentet.

SGML anger egentligen inte regler för hur man skall koda dokument. I stället är SGML en internationell överenskommelse om en uppsättning regler för hur den som märkt ett dokument skall berätta för andra hur denna märkning gått till, en berättelse som är omedelbart läsbar för mottagaren (antingen denne är en människa eller en maskin).

Att SGML är ett beskrivande språk innebär att SGML-märkningen inte säger något om den typografiska formgivningen. Vid utskrift eller fotosättning eller visning på skärmen vidarebefordras det SGML-märkta dokumentet till ett formateringsprogram som sörjer för att utseendet blir det önskade.

Det är inte meningen att den ordinäre läsaren skall behöva se eller bekymra sig om SGML-märkningen. Han eller hon ser på skärmen eller pappret en version som passerat genom ett formateringsprogram vilket upplöst entiteterna (dvs. ersatt `&Aring;` med bokstaven Å, etc.) och givit rubriker, olika slags stycken, kursiveringar etc en lämplig typografi.

Den personal som genomför märkningen av Strindbergmaterialet måste ha någon hum om SGML, men kraven är måttliga. Till sin hjälp har man en SGML-editor som ur användarens synvinkel fungerar ungefärligen som en vanlig textbehandlare och som dessutom förstår SGML. För att tillföra märkning till en textfil laddar man in denna i SGML-editorn, varefter man till sin hjälp har kontextkänsliga menyer som för varje plats i texten upplyser om de tillåtna märkorden och attributen: här har du lov att stoppa in ett märkord för citat, här har du lov att stoppa in ett märkord för underrubrik på nivå 3, etc. Man väljer oftast märkord eller attribut genom att klicka i menyerna och behöver alltså inte skriva in deras namn. Den som märker behöver inte bekymra sig om de ordinära teckenentiterna, vilka editorn översätter automatiskt vid import till SGML (Å ersättes av `&Aring;`, etc) och åt andra hållet vid presentation och export (`&Aring;` visas som Å på skärmen).

Om i framtiden ordinära textbehandlingsprogram kan användas i stället för de särskilda SGML-editorerna skulle märkningsarbetet bli smidigare. Men det är långt dit. Hittills har textbehandlingsfabrikörerna lovar mer än de kan hålla. De tilläggsprogram som ger SGML-stöd för Microsoft Word och WordPerfect är i sin nuvarande utformning alltför klena föra att duga till märkning av omfattande och komplexa dokument.

6. TEI

Text Encoding Initiative (TEI) är ett internationellt projekt som engagerar humanister och dataloger i många länder i syfte att utarbeta riktlinjer för organisering, märkning och utbyte av all slags humanistisk litteratur: editioner av medeltida handskrifter, diktsamlingar, romaner, pjäser, språkkorpora, lexika etc. 1994 offentliggjordes det första officiella samlade förslaget från TEI under rubriken *Guidlines for Electronic Text Encoding and Interchange*, redigerat av C.M. Sperberg-McQueen vid universitet i Chicago och Lou Burnard vid universitet i Oxford.

Enligt min mening är TEIs riktlinjer en välsignelse. För den som önskar utnyttja sig därav existerar nu en internationell överenskommelse om hur humanister skall koda, arkivera, katalogisera, distribuera och hänvisa till digital litteratur tvärs över gränserna mellan nationalspråk, datorplattformar och tillämpningar. Och övriga, som av någon anledning väljer andra lösningar, kan utnyttja TEIs riktlinjer som en gemensam referenspunkt. TEIs förslag är avsedda att tjäna som inlägg i den bredare diskussionen om hur texters egenskaper skall representeras. Även den som för egen del föredrar andra sätt att representera redaktionella ingrepp, dikters rytm och meter eller lakuner i källor, kan hämta inspiration från TEIs detaljerade behandling av otaliga sådana märkningsproblem. Och till sist, även om man ändå beslutar sig för en mycket egen lösning, kan det tänkas att TEI erbjuder ett utbytesformat som gör materialet flyttbart så att det blir tillgängligt för omvärlden.

6.1 TEIs organisation, tillgängligt material

Det första planeringsmötet ägde rum i november 1987. Projektet stöddes ursrpungligen av Association for Computers and the Humanities (ACH), Association for Computational Linguistics (ACL) och Association for Literary and Linguistic Computing (ALLC). Senare har U.S. National Endowment for the Humanities (NEH), Andrew W. Mellon Foundation samt Europakommissionen (XIIIe direktoratet) tillkommit som fiansiärer. Ett första synnerligen preliminärt utkast (kallat ”P1”) till TEIs riktlinjer publicerades i juli 1990. Ett femtontal kommittéer fortsatte arbetet, och reviderade versioner av de för märkningsarbetet viktigaste kapitlen[6] blev efter hand tillgängliga på internet. Därmed gavs många humanister världen över chansen att genast börja pröva riktlinjerna.

Det fanns således mycken sakkunskap och erfarenhet att bygga på när den nu gällande versionen av riktlinjerna publicerades, den första som inte var försedd med underrubriken ”draft”. Den kallas ”P3”, publicerades i maj 1994 och bestod av dels en omfattande DTD, fördelad på ett stort antal moduler, dels en dokumentation[7] som i utskrivet skick fyller nära 1300 sidor i A4-format.

Verksamheten fortsätter i kommittéer och arbetsgrupper med olika inriktning. Det överordnade organet är en *Steering Committee* med representanter för forskningsfinansiärerna. Styrningen från forskarnas sida sker genom ett *Advisory Board* där 15 akademiska organisationer finns företrädda. De viktigaste samordnande arbetet har hittills utförts av Lou Burnard och C.M. Sperberg-McQueen, vilka fungerar som europeisk resp. amerikansk redaktör för TEIs riktlinjer. En för TEIs fortsatta utveckling central roll har en nyupprättad *TEI Technical Review Committee* (TRC), konstituerad i juni 1996, som från arbetsgrupperna och annat håll tar emot och bedömer förslag till revidering och komplettering av

[6] Kapitel 22, som blev tillgängligt redan i augusti 1992, innehöll de inte minst för bibliografiska syften väsentliga riktlinjerna för skapande av "TEI headers", dvs det avsnitt som skall inleda varje TEI-dokument och som innehåller uppgifter om titel, författare, källa, version, redaktionella principer och allt annat som kan behövas, exempelvis för automatisk katalogisering, för upprättande av index över dokumenten i ett bestånd etc. Kapitel 7 blev tillgängligt i oktober 1992 och innehöll "the bas tag set" för prosa. Kapitel 6, tillgängligt i december 1992, beskrev element som kan finnas i alla slags TEI-dokument.

[7] Burnard, Lou/Sperberg-McQueen, C.M (Eds.): *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Volume I-II. Text Encoding Initiative, Chicago University/Oxford University, 1994. En korrigerad andra upplaga skall utkomma i september 1997.

TEIs riktlinjer. TRC har även i uppgift att samordna verksamheten inom arbetsgrupperna och att vid behov rekommendera upprättandet av nya arbetsgrupper.

Arbetet med digitala utgåvor av Strindbergs verk bör ha kontakt med detta internationella sammanhang.[8]

Aktuell information kan hämtas från TEIs egna officiella WWW-sidor på URL http://www-tei.uic.edu/orgs/tei. Där finns bl.a. instruktioner om hur man laddar ned TEIs DTD:er och dokumentation.

Den synnerligen livaktiga diskussionslista där TEI-relaterade spörsmål debatteras och nyheter presenteras heter TEI-L. Gamla bidrag till denna konferens finns i HTML-format tillgängliga i ”shadow archive” på URL http://www.let.ruu.nl/C+L/loeffen/archive/tei/intro.htm. En mer teknisk diskussion förs på listan TEI-TECH.

6.2 TEIs riktlinjer, några grundläggande principer

Med riktlinjerna vill TEI bidra till gemensamma internationella överenskommelser om hur all slags litteratur av intresse för humanister kan lagras och spridas i digital form. Av ett så generellt syfte följer att riktlinjerna inte får vara alltför styrande.

En viktig princip är att riktlinjerna skall vara ”konkurrensneutrala” i förhållande till olika skolbildningar och forskningsinriktningar. De får med andra ord inte favorisera exempelvis de forskare som vill lyfta fram författarens intentioner, eller de forskare som vill studera historiska och sociala omständigheter, eller de som prioriterar tematisk eller stilistisk analys.

En annan princip är att det finns ytterst få tvingande påbud i TEIs riktlinjer, utöver det tvång som SGML-standarden i sig innebär. För att riktlinjernas krav skall vara tillfredsställda är det framför allt ”meta-informationen” — upplysningar om dokumentets rubrik, upphovsman, redaktionellt ansvar, publiceringsdatum, revisionshistoria etc, dvs. sådant som ingår i den ”TEI header” som skall medfölja varje dokument — som måste ingå och märkas på föreskrivet sätt; detta förstås för att möjliggöra identifiering, automatisk katalogisering, kvalitetsbedömning etc. I övrigt har man stor frihet att i enlighet med egna behov utforma sitt märkningsschema. Man är dock tvungen att i dokumentet precisera principerna för det märkningschema man valt, och TEIs riktlinjer innebär ytterst just ett elaborerat förslag om hur märkningsprinciper kan fastställas så att andra (mänskliga läsare eller datorprogram) begriper dem.

Jag använder här ordet tvång i en teknisk mening. Om man i DTD-konstruktionen och märkningen följer de fåtaliga tvingande påbuden i TEIs riktlinjer (samt SGML-standardens allmänna krav) är man garanterad att slutresultatet blir ” TEI conformant documents”, vilket är en stor fördel. Både man själv och mottagarna vet hur märkningen skall tolkas och vilka program och procedurer som är tillämpliga för den fortsatta bearbetningen. Man kan visserligen ha viss glädje av TEIs riktlinjer beträffande olika märkningsproblem även om syftet inte är att skapa ”TEI conformant documents”, men då går man miste om de fördelar som erbjuds i mittfåran av det internationella standardiseringsarbetet.

Vid sidan av de tvingande inslagen i TEIs riktlinjer finns sådana som är ”rekommenderade” och ”valfria”. Om inga starka skäl talar emot är det klokt - men det är inte nödvändigt för att

[8] Det handlar inte bara om att hämta inspiration från TEIs verksamhet utan även om att från svensk sida efter förmåga bidra till utprovningen och utvecklingen av TEIs riktlinjer. Med tanke på bl.a. kommande behov i samband med märkningen av Strindbergsmaterialet har jag själv varit engagerad i ett förslag till en ny arbetsgrupp som skall syssla med hur manuskript skall märkas och förses med metainformation. (Se Claus Huitfeldt, *Proposal from the Networking of Literary Archives (NOLA) Project for a new TEI Work Group on Encoding and Meta-Description of Manuscripts*, 18 June 1996). Initiativtagare är Claus Huitfeldt, arbetsgruppen är tänkt att ledas av Espen Ore (universitetet i Bergen) och den övr. prel. medlemslistan innefattar f.n. Donald Broady, Richard Gartner (Bodleian Library, Oxford), Richard Giordano (Manchester University, GB), Elli Myllonas (Brown University, USA) och Peter Robinson (Oxford University och De Montfort University, GB).

Denna arbetsgrupp är en utlöpare av konsortiet Networking of Literary Archives (NOLA), i vilket från svensk sida utöver KTH även Kungl Biblioteket varit representerad av Anders Burius. För mer information se URL http://gonzo.hd.uib.no/Nola/Nola.html.

skapa ”TEI conformant documents” — att anpassa märkningen till den praxis som TEI rekommenderar och som dels brukar vara väl motiverad, dels underlättar utbyte och samarbete med andra projekt.

TEIs riktlinjer består av en rikhaltig uppsättning (ca 400) märkord och attribut, samt principer för hur dessa kan utnyttjas för att beskriva olika egenskaper hos skilda slag av texter. Egenskaperna kan vara ”strukturella”, dvs. beskriva textens uppbyggad i termer av titel, författare, rubriknivåer, blockcitat, fotnoter, redaktionella tillägg, scenanvisningar och dialog i en pjäs, etc. Flertalet av de egenskaper som fångas upp av TEIs märkord är dock mer ”innehållsliga”. Med deras hjälp kan man ange om Strindberg skrev med bläck, blyerts eller rödpenna, man kan märka ut personnamn, geografiska lokaliteter eller datum, tematiska och stilistiska egenskaper hos texten, metriska egenskaper hos en dikt etc.

Om de märkord och attribut TEI föreslår inte duger är det jämförelsevis enkelt att modifiera dem eller lägga till egna. Materialet är fortfarande ”TEI conformant” förutsatt att man i eller i anslutning till själva dokumentet på föreskrivet sätt angivit hur ändringarna genomförts. Och i synnerhet åtskilliga av attributen är redan i TEIs riktlinjer definierade så att de kan ges vilka värden som helst.

Det är så riktlinjerna är avsedda att användas, inte som tvångströja utan som flexibelt ramverk som kan anpassas till de egna kraven. Men det visar sig att TEI lyckats förutse många behov. Tag exempelvis den till dags dato troligen mest genomarbetade digitala kritiska utgåvan, den redan nämnda av Peter Robinson redigerade första CD-ROM-produktionen från Canterbury Tales-projektet. Utgåvan, som publicerad i april 1996, ger läsaren avancerade möjligheter att jämföra samtliga från tiden före 1500 bevarade femtioåtta versioner av Chaucers ”Wife of Bath’s Prologue”. Jämförelsemöjligheterna är väsentliga eftersom manuskripten varit svåråtkomliga (de är spridda på 25 arkiv, bibliotek och privatsamlingar) och eftersom Chaucerforskarna är oeniga om attribueringen och i synnerhet dateringen. För första gången finns nu transkriptioner och faksimil av alla manuskripten samlade i flyttbart skick, och på CD-ROM-utgåvan kan manuskripten jämföras med användning av Peter Robinsons kraftfulla versionsjämförelseprogram *Collate* (som många humanister i dag använder även i andra liknande sammanhang). Märkningsarbetet har pågått sedan 1989 och för CD-ROM-produktionen har inte mindre än 10 miljoner element märktes ut och 2 miljoner länkar skapades. Ändå kunde TEIs riktlinjer följas näras nog utan modifieringar (de enda ändringarna var att tre nya element skapades och att tio ”element content groups” modifierades).

7. Ett exempel på märkning

Låt mig till sist illustera hur en sida Strindbergtext kan märkas. Jag väljer början av ett kapitel i Röda Rummet (Saml. Verk, volym 6, 1981, p. 81).

SJUNDE KAPITLET

Jesu Efterföljelse

Följande morgon väcktes han av städerskan som framlämnade ett brev vilket befanns vara av följande lydelse.

Timot. Kap. X, v. 27, 28, 29. Första Korint.
Kap VI, v. 3, 4, 5.

Dyre Br!
Vår H:s J. Kr. Nåd och Frid, Fadrens kärlek och D. H. A:s delaktighet etc. Amen!

Jag såg av Gråkappan i går afton att Du ämnar utgiva Försoningsfacklan. Sök mig i min verksamhet i morgon bittida före 9.

din återlöste
Nathael Skåre

Nu förstod han Lundells gåtor, till en del! Han kände visserligen icke den store gudsmannen Skåre personligen och visste intet om Försoningsfacklan, men han var nyfiken och beslöt att hörsamma den närgångna kallelsen.

Kl. 9 stod han på Regeringsgatan framför det väldiga fyravåningshuset, vars fasad var klädd med skyltar ifrån källarvåningen ända upp till taklisten. *Kristliga Boktryckeri Aktiebolaget Friden* 2 tr. upp. *Redaktionen av Guds barns arvedel* ½ tr. upp. *Expeditionen av Yttersta Domen*, 1 tr. upp. *Expeditionen av Fridsbasunen* 2 tr. upp. *Redaktionen av Barntidningen: Föd mina Lamm* 1 tr. upp. *Direktionen för Kristliga Bönhusaktiebolaget Nådastolen verkställer utbetalningar och beviljar lån mot första inteckning i fastighet* 3 tr. upp. *Kom till Jesus* 3 tr. upp. Obs. ✉ Ordentliga utsäljare som kunna ställa borgen erhålla sysselsättning därstädes. *Mis-*

Det på förra sidan återgivna textpartiet kan förslagsvis förses med följande märkord, hämtade från TEIs riktlinjer.

```
<div1 type="chapter" n='7'>
<pb n="81">
<head type="n">SJUNDE KAPITLET</head>
<head type="main">Jesu Efterföljelse.</head>
<p rend="noindent">Följande morgon väcktes han av städerskan som framlämnade ett brev vilket befanns vara av följande lydelse.</p>
<div2 type="letter">
<opener rend="center">Timot. Kap. X, v. 27, 28, 29. Första Korint.
Kap VI, v. 3, 4, 5.</opener>
<salute> Dyre Br!</salute>
<p>Vår H:s J. Kr. Nåd och Frid, Fadrens kärlek och D. H. A:s delaktighet etc. Amen!</p>
<p>Jag såg av Gråkappan i går afton att Du ämnar utgiva Försoningsfacklan. Sök mig i min verksamhet i morgon bittida före 9.</p>
<close rend="right">
<salute>din återlöste</salute>
<signed>Nathael Skåre</signed>
</close>
</div2>
<div2 type="section">
<p>Nu förstod han Lundells gåtor, till en del! Han kände visserli&shy;gen icke den store gudsmannen Skåre personligen och visste intet om Försoningsfacklan, men han var nyfiken och beslöt att hörsamma den närgångna kallelsen.</p>
<p>Kl. 9 stod han på <placeName><geogName type="gata">Regeringsgatan</geogName></placeName> framför det väldiga fyra&shy;våningshuset, vars fasad var klädd med skyltar ifrån källarvå&shy;ningen ända upp till taklisten. <emph>Kristliga Boktryckeri Aktiebolaget Friden</emph> 2 tr. upp. <emph>Redaktionen av Guds barns arvedel</emph> &slantfrac12; tr. upp. <emph>Expeditionen av Yttersta Domen,</emph> 1 tr. upp.<emph> Expeditionen av Fridsba&shy;sunen</emph> 2 tr. upp. <emph>Redaktionen av Barntidningen: Föd mina Lamm</emph> 1 tr. upp. <emph>Direktionen för Kristliga Bönhusaktiebolaget Nådastolen verk&shy;ställer utbetalningar och beviljar lån mot första inteckning i fastighet</emph> 3 tr. upp. <emph>Kom till Jesus</emph> 3 tr. upp. Obs. <!-- Infoga symbol pekande hand --> Ordentliga utsäljare som kunna ställa borgen erhålla sysselsättning därstädes. <emph>Mis</emph>
<pb rend="shy" n="82"><emph>
```

Låt oss först granska märkorden som beskriver textens strukturella eller formella egenskaper.

`<div1 type ="chapter" n='7'>` anger att här börjar ett avsnitt. TEIs riktlinjer föreslår att märkorden `div0`, `div1`, `div2` etc används för att representera avsnitt på olika nivåer, exempelvis motsvarande en romans delar (div0), kapitel (div1), sektioner inne i kapitel (div2) osv., eller kanske diktsamlingar (div0), däri ingående diktsviter (div1) osv. Avsnitten kan således vara inkapslade i varandra, och det är i linje med TEIs filosofi att inte på förhand föreskriva att de skall kallas kapitel, avdelning, diktsvit, essä, novell eller något annat. Här rör det sig dock om ett kapitel, varför attributet `type` till elementet `div1` tilldelas värdet `chapter`.

Märkordet `<head>` anger rubrik. Attributet `n` anger att kapitlet är numrerat. Attributet `main` anger att det är ett fråga om kapitlets egentliga rubrik. Inget av dessa attribut föreskrives av TEIs riktlinjer utan väljes på det sätt som bäst passar texten i fråga, men principerna måste redovisas i den TEI header som inleder dokumentet.

 och omsluter ordinära stycken (eng. ”paragraph”).

Med hjälp av dessa strukturella märkord har vi beskrivit textavsnittets hierarkiska uppbyggnad, som vi kan föreställa oss som lådor instoppade i varandra (jfr fig. 1 nedan). Det största lådan (div1-nivån) representerar det sjunde kapitlet som i sin tur innehåller dels rubrik, dels olika avsnitt på närmast lägre nivå (div2). Det första elementet på div2-nivån är ett brev, innehållande ett antal märkord som identifierar hälsningsfraser; vissa av dessa element är inkapslade i varandra. Nästföljande element på nivån div2 är ett mer ordinärt avsnitt som innehåller ett antal stycken. Samma struktur kan åskådliggöras med ett träddiagram (jfr fig. 2 nedan).

Som exempel på märkord som beskriver textens innehållsliga aspekter har jag nöjt mig med <geogName>, som försett med lämpliga attribut används för att identifiera namn på länder, städer, gator och andra platser. På liknande sätt skulle andra märkord kunna infogas för att markera t.ex. varje omnämnande av personer i romanen; sådana märkord skulle dessutom kunna förses med ID-attribut som låter läsaren spåra en och samma person trots att denne omtalas på olika sätt (med för- och efternamn, smek- eller öknamn, omskrivningar, eller enbart ett personligt pronomen). Vidare kan man märka ut årtal och datumangivelser och så vidare i all oändlighet. Möjligheterna är obegränsade, men kan förstås inte uttömmas när materialet redigeras i arkivformatet. I stället måste läsarna ges möjlighet att komplettera med egna märkord: somliga Strindbergforskaren kan ha intresse av att tillföra märkning som möjliggör lingvistisk analys, andra önskar skapa sig ett underlag för tematisk eller stilistisk analys, åter andra vill identifiera anspelningar på personer eller händelser i Strindbergs samtid. Det är tänkbart att skilda forskningsprojekt med olika inriktning som ett led i sitt arbete åstadkommer en märkning för eget bruk som kanske senare kan överlämnas som en gåva till forskarsamhället.

Jag har även tagit med några entiteter. Sådana känner man igen på att de alltid inleds med tecknet ”&” och avslutas med tecknet ”;”. De står där för att bli utbytta mot något annat, t.ex. mot ett tecken i en systemberoende teckenuppsättning, innehållet i en teckensträng, eller en fil (t.ex. en liten bildfil). I en fullständig SGML-märkning måste även bokstäverna å, ä, ö bytas mot entiteter för att inte fördärvas vid överföring till system med andra teckenuppsättningar.

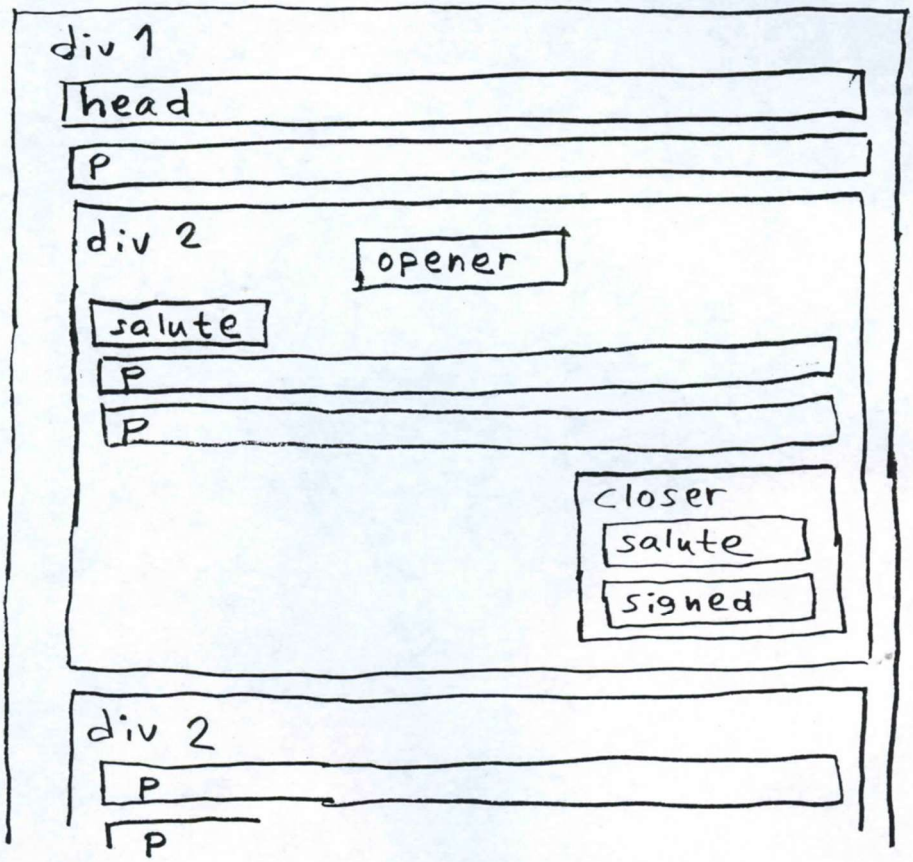
Entiteten uttydes ”soft hyphen” och används här för att representera bindestreck som uppstått i samband med radbrytningen i den tryckta utgåvan av Saml. Verk. Det gäller att hålla reda på dessa, bl.a. eftersom Strindbergs egna bindestreck på ett eller annat sätt måste särskiljas från andra bindestreck (vilket inte sker i Saml. Verk., men detta är lättare att åstadkomma i en digital version).

Entiteten &slantfrac12; används för att representera tecknet ½.

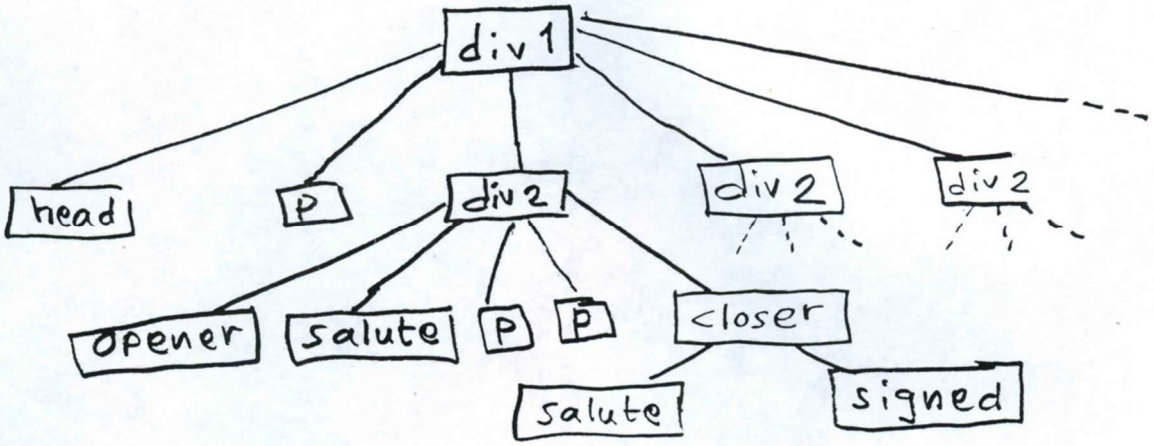
<!-- Infoga symbol pekande hand --> är en SGML-kommentar, dvs en annotering. Dessa inleds med ”<!--” och avslutas med ”-->”, och kan infogas var som helst i texten. Datorprogrammen bryr sig inte om SGML-kommentarer, som fungerar som temporära minnesanteckningar för den som utför märkningen eller som en upplysning till läsaren. Här är kommentaren en påminnelse om att en bild skall infogas. Vid den fortsatta redigeringen kan kommentaren bytas mot en entitet som hämtar bilden från en särskild fil.

Attributet rend uttydes ”rendition” och används bl.a. för att vid behov styra typografin på skärmen eller på papper. Här förekommer attributet rend på två ställen, där det givits värdet noindent som anger att kapitlets första stycke saknar indrag samt värdet center som anger att bibelcitatet i brevets motto skall centreras.

Elementen <pb n="81"> i början och <pb n="82"> i slutet är tomma element, dvs tjänar som platshållare och innesluter ingen text. Förkortningen pb skall utläsas ”page break”. Elementen beskriver sidbrytningen i den tryckta upplagan Saml. Verk.



Figur 1. Sjunde kapitlets struktur åskådliggjord som lådor i lådor



Figur 2. Samma struktur åskådliggjord som träddiagram