Human IT
Tidskrift för studier av IT
ur ett humanvetenskapligt perspektiv

En rapport från EP98, St. Malo, 1-3 april 1998.

av Mikael Gunnarsson

Institutionen Bibliotekshögskolan, Högskolan i Borås

7th International Conference on Electronic Publishing, Document Manipulation and Typography var det sjunde i ett vartannat år återkommande forum för frågeställningar kring elektronisk publicering, vilket den här gången ägde rum i St. Malo, Bretagne, i början av april. En alldeles förträfflig plats att vid denna tidpunkt arrangera en konferens på. Den annars ganska frekventa franska oviljan till att tala något annat än franska gavs det för övrigt endast uttryck för i en enda presentation. [10] Konferensen hölls denna gång delvis samtidigt och i samarbete med 4th *International Conference on Raster Imaging and Digital Typography*. För den som liksom jag är mindre intresserad av tekniska lösningar än av dess implikationer var det kanske mindre lyckat. Det är svårt att bli upphetsad över frågeställningar kring hur mongoliska skriftarters ligaturer skall kunna representeras och återges korrekt [1] eller hur man representerar och förlänger barocknotationens notbalkar [2].

Konferensen var ingalunda endast ägnad frågeställningar kring textproduktion i sig och för sig. Distinktionen mellan konsumtion och produktion av texter, som vilar på antagandet att å ena sidan någonting packas in och å andra sidan att detta någonting packas upp, synes mer och mer som en ofruktbar konstruktion.

> *Electronic publishing is in the process not merely of changing the production of print products (through the submission, editing and printing of books from disk), but also of vastly extending the range of "non-print" publishing, to include hypermedia and database publishing on CD-ROM and disk, and on-line and networked publishing.* (International Encyclopedia of Information and Library Science. - Routledge, 1997.)

"Att publicera" är inte detsamma i dag som för tio år sedan. Givetvis, skall man kanske säga. "Att publicera" hade säkerligen inte heller samma innebörder för 10 år sedan som för 30 år sedan. Men det är ovidkommande. Saker och ting sker, och begreppet "att publicera" är knutet till vissa tekniker som omformar oss och sig självt. Det är så mycket som man kan

säga. Det som följer på omställningen från "traditionell" publicering till *elektronisk* publicering inskränker sig knappast endast till vad IEILS här ovan utmålar under rubriken *publishing*. Man tror sig åtminstone kunna säga så mycket efter ett besök på EP98

Man tror sig åtminstone kunna säga så mycket efter ett besök på LP 98.

Att pappersburna texter inom en rätt snar framtid kommer att kunna hänföras till historiens relikter är det väl inte många som tror nu längre. Det är en tanke som får tillerkännas gångna tiders teknikvisionärer. Papper *är* ett förträffligt *material* som endast i vissa avseenden och i vissa situationer så här långt med teknikens "utveckling" kan sägas vara mindre lämpliga än digitala medier att bära texter. En sådan situation är t ex då man önskar återanvända ett avsnitt text i en annan. *Document reuse* var ett av konferensens teman, möjligen just på grund av att accentuerandet av återanvändningen av textfragment utgör ett av det elektroniska dokumentets stora fördelar. Vidare är det en egenskap som kanske tvingar fram en omvärdering av dokumentbegreppet inom ett större område än det hittills gjort. Så arbetar t ex Ahonen et al [6] och Paradis et al [7] utifrån antagandet att enskilda textfragment kan återanvändas som delar av *virtuella dokument*, där tekniken "allows information to be gathered from multiple sources and combined dynamically to produce a virtual document". Hos de förstnämna är det värt att observera att de textfragment som utgör ett virtuellt dokument inte nödvändigtvis ens behöver bära på spår av dessas uppkomst och ursprungliga kontext.

Dessa virtuella dokument lär kanske inte komma i bruk i sammanhang där dokumentet t ex skall utgöra vittnesbörd om något faktiskt, som vid avtalsskrivning. Papper och digitala lagringsmedier kan förväntas fungera i skilda sammanhang, precis som vi i vissa situationer nödgas ta till särskilt arkivpapper i stället för det vanliga självutplånande. Emellertid är det knappast omöjligt att "familjen" pappersburna texter kommer att flyta samman med sina "elektroniska" motsvarigheter, eller om vi så vill, berikas med en ny teknik. Åtminstone om man får tro Max Copperman och Marc Dymetman på Xerox Research Centre Europe.[3] Papperets fördelar är enligt dessa två att de gör texter beständiga *(permanent)*, bärbara *(portable)* och att de uppträder i ett fysiskt sammanhang *(contextualized)*. Den mnemoniska funktionen hos papperet är därmed en värdefull aspekt. Det elektroniska dokumentet däremot uppvisar inte dessa karaktäristika men är i stället i högre grad reproducerbara och fr a - enligt Copperman och Dymetman - dynamiska och immateriella, samt kapabla att initiera aktivitet *(agentive)*.

Möjligheterna att förena dessa egenskaper i en och samma publiceringsteknik ligger enligt dessa två i s k *intelligent papper*, som blir till genom att ge pappersarket en ytterligare "dimension" (dvs ett lager osynligt bläck) och en motsvarande representation i digital version på nätet. Papperet ges en identitet *(page-id)* som kan omvandlas till en URL och på så sätt knytas till ett motsvarande objekt någonstans på nätet. En liten penna med inbyggd kamera och sändare reproducerar sedan pennspetsens rörelser och sänder dem vidare till den digitala motsvarigheten.

På så sätt skall läsaren t ex kunna utnyttja hyperlänkar mellan papperet och webben eller kopiera text från papper till ett ordbehandlingsprogram. Med en sådan kombination av papper och digitala media menar Copperman och Dymetman att skulle vi kunna uppnå stora fördelar.

Exempelvis kan faktiska länkar som fanns på webbsidan innan det fixerades på papper åter komma till användning. Det blir inte längre nödvändigt att hantera referenser som http://domino.idg.se/natkom/nokart.nsf/b19b268a4e765345c125643a0066f825/a9b3744696ad0f2a412565a70061c62a?OpenDocument i form av manuellt förda anteckningar.

I förlängningen tänker de sig t ex en tryckt pocketutgåva av en stor encyklopedi som förses med faktiska länkar till den kompletta utgåvan i digital form. Det intelligenta papperet rönte givetvis ordentlig uppmärksamhet och man var inte sen att påpeka problematiska situationer. Hur hanterar man en sådan sak som att papper kan delas i flera delar, att papper kan förstöras? Hur representeras sådana handhavanden i den digitala motsvarigheten?

Ansatser mot ett förenande av papper och digitala medier som Coppermans och Dymetmans saknas inte. De båda hänvisar själva till en mängd projekt, bl a det som resulterat i det s k SmartPaper (Johnson, W. Et al, Bridging the paper and electronic worlds // Proceedings of INTERCHI, ACM. - April, 1993). Copperman lämnade plats på scenen för Heather Brown från Univ. of Kent som presenterade resultaten av sin projektgrupps försök med *The DigitalDesk,* "an attempt to produce an intelligent desk that can work with real paper, rather than simulating the paper on a computer screen".[4] Ett projekt som inom den reguljära språkundervisningen funnit en plattform för att berika en tryckt pappersbaserad *Alice i Underlandet* med inbäddade metadata i en digital TEI- representation. På så sätt kan en videokamera placerad över skrivbordet avläsa vad som sker och reproducera didaktiska reaktioner mot skrivbordet, t ex belysa alla ord som är verb eller substantiv.

Gentemot detta perspektiv, där man söker vägar att berika produktionen av texter, står i viss mån ett IR-perspektiv. Här handlar det om att utifrån produkternas *natur* söka metoder att göra det möjligt att finna enskilda textprodukter i en samling av sådana. Målet blir att fastställa kriterier för hur man skall konstruera fungerande index till en sådan samling.

Under decennier har detta IR-perspektiv fokuserat flera frågeställningar av denna art. Bl a har man eftersträvat att, på grundval av lingvistiska teorier om språkets möjligheter att avslöja mening, formulera teorier och konstruera fungerande algoritmer. Dessa skall sedan kunna användas för att skapa program som extraherar och genererar för den enskilda texten signifikanta ord och fraser, - ord och fraser som förväntas avslöja ett dokument *relevans* för en viss bestämd frågeställning. I detta perspektiv framhålles numera inte sällan algoritmernas tillkortakommanden och det pläderas för en återgång till *human intellectual indexing*, där index i stället skall framställas på grundval av professionella indexerares analys av dokumenten och genom en interpolering av denna mot en kontrollerad vokabulär.

> *Intellectually it is possible for a human to establish the relevance of a document to a query. For a computer to do this we need to construct a model within which relevance decisions can be quantified.*(van Rijsbergen, C. J., Information retrieval. - Butterworth, 1975. - S. 5)

I detta sammanhang kan man kanske förstå den indignation som tycktes möta Isabel Cruz när hon presenterade sin grupps experiment med att jämföra HTML-märkningen i den ämnesstrukturerade Yahoos alla kategorier. Att man säger "it looks like a newspaper site" tycks för Cruz och hennes kolleger innebära att det finns en struktur i ett webbobjekt som för tankarna mot en viss genre eller t o m en viss mening. Vidare antar de att den struktur som ett webbobjekt i så fall avslöjar kan avläsas ur HTML-märkningen. Deras idé går därmed ut på att undersöka om likheter i sättet att märka upp webbobjekt kan visa på någon överensstämmelse med de ämnen som Yahoo placerat dem under.[8] Det var tydligt att man var ganska enig om att det i själva verket inte var struktur som de undersökte, utan blott och bart HTML-märkning. "HTML is not structured", var en kommentar.

Forskning kring *information retrieval* har vanligtvis utgått från en situation där ansvaret för indexering och tillgängliggörandet av den samlade mängden publicerat material åvilar särskilda institutioner (bibliotek, databasvärdar eller databasproducenter). Skribenten eller publicisten har endast haft att framställa en för läsaren användbar text, en verksamhet skild ifrån den som så att säga framställer *samlingar av dokument*.

Flera faktorer gör dock denna distinktion överflödig, för att inte säga omöjlig. Det viktigaste är kanske att de automatiserade principerna inte är möjliga att helt och hållet överge med den enorma tillväxt av elektroniska publikationer som föreligger. Den heterogenitet som

"dokuversum" uppvisar utarmar i mångt och mycket möjligheten att använda framtagna modeller för automatisk textanalys. I den stund som en samling av dokument härrör från snart sagt vilken möjlig kontext som helst, och att det språk på vilka de är skrivna kan vara ett eller flera av vilket språk som helst, så blir många av de traditionellt obligatoriska fenomenen i ett index svåra att använda. Stoppordslistor och sammanförande av alla ord med gemensam ordstam är bara två exempel. Det är här frågan om *hur* texter produceras faller inom sfären för vad som kan intressera publiceringssystemet som helhet.

I samband med den teknik för textproduktion som vi hittills har haft att göra med vilar möjligheterna att uttrycka sig på ett skriftspråk. De möjligheter till att uttrycka sig med nyanser som det talade språkets paralingvistiska dimension medger, med tonala nyanseringar och ackompanjerande kroppsspråk, kan i någon mån sägas motsvaras av typografiska konventioner som varierande typsnitt och storlek, samt positionering av vissa textavsnitt i förhållande till andra. Rubriker kan framhävas med större typsnitt, numrering och/eller genom att det omges med större blankutrymmen, dvs i allt genom visuella *konventioner* som kan variera från publikation till publikation. De är mångtydiga och det finns begränsade möjligheter att på automatiserad tillförlitlig väg använda dessa indikationer på vissa textavsnitts vikt i förhållande till andra. Att framställa index på grundval av samspelet mellan språket och en texts lay-out ter sig svårt.

Med dagens tekniker att representera texter med sin struktur indikerad i gömda s k "taggar" eller "märken" (*mark-up*) öppnas emellertid möjligheter att fastställa samband mellan syntax och semantik. HTML-, SGML-, TEI- och XML-objekt lämnar vid sidan av läsarens text också dolda men bearbetningsbara indikationer på vikten av vissa av textens fragment. Det här är ett förhållande som antagligen är välbekant för var och en som arbetat med HTML utan hjälp av Adobe Pagemill, Microsoft Frontpage eller liknande WYSIWYG-editorer, vilka tyvärr fjärmar skribenten från den direkta kontakten med det som hon arbetar med och döljer dess "märken". <H1> och <TITLE> utgör *indikationer* på textfragmentens strukturella innebörd i förhållande till andra i texten förekommande partier och kan givetvis användas för att t ex åstadkomma index som har alla förutsättningar att fungera bättre än de som hittills byggts på automatisk väg med utgångspunkt enbart från statistiska modeller. I synnerhet gäller detta SGML, TEI och XML. HTML har tyvärr utvecklats i en riktning där <H1> inte längre nödvändigtvis betecknar en rubrik, utan i stället används för att åstadkomma en visuell effekt.

Även om den ende representanten för Biblioteks- och Informationsvetenskap Terence Brooks ondgjorde sig över "the effect of textual aesthetics on information retrieval"[5] och menade att t ex det flitiga bruket av vitsiga titlar som *Grant$ for women and girls, Toys 4 Us* eller *.doc* ställer till svåra problem vid den s k återvinningen, så vittnar å andra sidan uppmärkningsspråkens framgångar (XML, SGML och HTML) och den stora uppmärksamhet[1] de ägnas, om motsatsen med tekniska förändringar. Mazundar et al accentuerar detta. [9] Dessa framhåller i sin "adding semantics to SGML databases" den stora

betydelse för tillgänglighet som SGML-relaterade representationsformer för text innebär och har utvecklat en lösning som möjliggör för skribenten att i *detalj* klargöra en texts semantik, vilket sedan utnyttjas som ett alternativ till probabilistiska IR-modeller.

Av detta följer att även om tidigare tekniker för textproduktion inte nödvändiggjort en värdering och analys av olika sätt att framställa texter, så blir nu framställningsmetoderna möjligheter att på automatisk väg också framställa fungerande index.

7th International Conference on Electronic Publishing, Document Manipulation and Typography var således en konferens som vittnade om hur produktionssidans intressen konvergerar med andra aktörers

konvergerar med andra aktörers.

(Åter till början av artikeln)

Fotnot

1. Ungefär hälften av bidragen involverade på ett eller annat sätt strukturella uppmärkningsspråk, med SGML-relaterade sådan som dominerande.

Källor

Flera av konferensens presentationer är publicerade på webben.(http://www.irisa.fr/ep98/) Annars är den dokumenterad i:

Hersch, Roger D., André, Jacques and Brown, Heather (eds.), *Electronic Publishing, Artistic Imaging and Digital Typography, Proceedings of the EP'98 and RIDT'98 Conferences, St Malo: March 30 - April 3, 1998.* - Springer-Verlag: Heidelberg, 1998 - (Lecture Notes in Computer Science Series ; 1375)
ISBN 3-540-64298-6

Vari ingår bl a följande bidrag som kommenterats i det här referatet:

[1] **Kataoka**, Tomoko I. et al, Internationalized Text Manipulations Covering Perso-Arabic Enhanced for Mongolian Scripts. - S. 305-318
[2] **Bouzaiene**, Nabil et al, Un bibliothéque informatiques pour la notation musicale baroques. - S. 319-330.
[3] **Dymetman**, Marc, Copperman, Max, Intelligent Paper. - S. 392-406.
[4] **Brown**, Heather et al, Active Alice: using real paper to interact with electronic text. - S. 407-419.
[5] **Brooks**, Terence, The effect of textual aesthetics on information retrieval. - S. 454-463
[6] **Ahonen**, Helena et al, Design and implementation of a document assembly workbench. - S. 476-486.
[7] **Paradis**, François et al, A virtual document interpreter for reuse of information. - S. 487-498.
[8] **Cruz**, Isabel F. et al, Measuring structural similarity among web documents: preliminary results. - S. 513-524
[9] **Mazumdar**, Subhasish et al, Adding semantics to SGML databases. - S. 563-574.
[10] **Rhissassi**, Habib, Lelu, Alain, Projet Hypermap : pour un environnement complet de génération automatique d'hypertexte. - S. 537-562. Ungefär hälften av bidragen involverade på ett eller annat sätt strukturella uppmärkningsspråk, med SGML-relaterade sådan som dominerande.

Mikael Gunnarsson has been employed at the Swedish Library School of Information Science since 1992, and has been teaching subjects mostly related to networking and electronic documentation.
Beginning as an undergraduate engineer in electronics, MG moved on to studies in theatre, acting, foreign langauges, and history of religion, until finally receiving his diploma in Library and Information Science . His academic interests lie in the hypertextuality of electronic documentation systems.

Åter till Human IT 2/1998