Vad en bra samling av digital text bör kunna

Sigfrid Lundberg
Biblioteksdirektionen
Lunds universitet

Inledande tankar

Sverige har — tyvärr — länge varit en efterbliven nation när det gäller bevarande och tillgängliggörande av sitt skrivna kulturarv i digital form. Det är allas vår förhoppning att det kommer till en ändring av detta till en följd av de utredningar och utvärderingar som gjorts, bland annat av Kungl. Bibliotekets verksamhet.

Innan jag kommer till det titeln på detta lilla papper antyder att det skall handla om, skulle jag vilja peka på ett par triviala punkter:

- Det är viktiga skillnader mellan bilder på (i) text, (ii) bilder på föremål och (iii) bilder av bilder. En bild av en text *är* en text, om än i ett olämpligt format — för många, men inte alla ändamål.
- En bild av en text kan inte effektivt återvinnas om den beskrivs med metadatastandard anpassad för bilder på stolar, eller för bilder av bilder på stolar.
- En bok är dels ett intellektuellt innehåll, dels ett objekt. Varje bild av en bok är (i) bild av en unik individ, forskningen *kan* vara betjänt av att få en uppfattning av och kunna beskriva variationen mellan individerna. (ii) Varje del av boken har en relation till helheten och till varje annan del; samlingen av bilder är surrogat för boken och relationerna mellan bilderna har samma relationer till varandra som det de avbildar.
- Varje längre text är strukturerad hierarkiskt. Den är uppdelad i avsnitt eller kapitel, varje kapitel i stycken, vilka i sin tur består av meningar, huvudsats, bisats, och slutligen ord. Denna struktur är till sin natur beroende av det intellektuella innehållet. Är det en tryckt text finns ytterligare en annan hierarki: Sida, rad, tecken.
- Digitala texter är för de flesta svåra och obekväma att läsa, åtminstone svåra att läsa på en datorskärm.

Text har många andra karaktäristika, men dessa är vad jag kommer att tänka på i skrivande stund.

Vad kan vi förvänta oss att användarna vill ha?

En applikation bör kunna det dess användare vill ha. Vad de vill ha beror på hur de kan tänkas vilja använda en elektronisk text, till skillnad från hur de använder en tryckt. Användningsområdena för ett textarkiv går inte att förutse i detalj, och för att få svar på frågan måste man fråga. Vi kan dock utgå ifrån att de *inte* kommer att använda de elektroniska texterna i hängmattan under semestern. Vad vi kan tänka oss för närvarande är minst det följande:

- ☐ Användarna vill ha en så trevlig digital miljö som bara möjligt för att helt enkelt läsa texterna.
- ☐ Men det räcker inte. Det behövs utskriftsvänliga format. Möjlighet att "shoppa" godtyckliga sidor och från dem skall det gå att få en snygg utskrift.
- ☐ Kopplingar till print–on–demand-tryckeri
- ☐ Möjlighet att *söka* efter ord och fraser. "Keyword in Context". Scanlistor över namn på företeelser, personer, platser, karaktärer och dylikt. Allt detta skall kunna göras inom en given bok, och globalt för hela samlingen och sådana sökningar skall kunna filtreras med metadata.
- ☐ Möjlighet att *länka* till godtyckliga positioner i en text; det är en fördel om länkad text blir markerad.
- ☐ Möjlighet till *utgående länkning*: Texter med not-apparat bör kunna förpassa noterna till separat fönster. Textkritiska noter bör ha särskilt användargränssnitt. Stöd för att koppla referenslistor till länkserver (SFX/Metalib-liknande tjänster).
- ☐ Personliga bokmärken och möjlighet till personlig annotation av texterna för inloggade användare.
- ☐ Användarforum för diskussion av digitaliserade texter
- ☐ Möjlighet till att publicera artiklar och annat vetenskapligt material relaterat till texterna, som då bör kunna dra nytta av dem och ingå i en hypertextuell symbios med materialet i arkivet.

Därutöver kommer säkert en massa andra funktioner som jag inte kan komma på.

Hur når man fram till detta?

Det är ett sorgligt faktum att vi förmodligen inte kan producera digitaliserad text som blir så användbar som man skulle vilja; den kodning som det skulle kräva skulle ställa sig dyrbar och tidskrävande. I själva verket föreställer jag mig att en digitaliseringsverksamhet där texter överförs till digital form i flera steg. Det innebär att dokument i olika stadier kommer att kunna behöva samexistera över relativt lång tid.

Jag tänker mig följande stadier:

1. Texter som består av en samling bilder, vars inbördes relationer är fastställda. Dessa verk kan återvinnas vid metadatasökning, navigering inom systemet eller via Libris/OPAC.
2. Texter som består av en samling bilder som för kategori 1, därtill kommer att texterna i denna kategori har genomgått OCR-behandling och finns sökbara. Länkar från ett sökresultat går endast till bilderna.

   Vid OCR-behandlingen sparas information om typografi (t ex kursiv och fet stil, styck- rad- och sidbrytningar etc).
3. Korrekturläst text. Med hjälp av sparad typografisk information åstadkoms halvautomatiskt en grund textmarkering, vilket gör att vissa hyperttextfunktioner kan implementeras. Full sökbarhet uppnås för ord och fraser. Likaså är full ingående länkning möjlig. För inloggade användare: Utgående länkning till personliga annotationer och diskussionsforum. Däremot finns inga markeringar för andra särskilda egenskaper hos texten. Notapparater existerar bara typografiskt, inte funktionellt.
4. Som punkt 3., men nu med djup markering. Texter i detta stadium är dessutom möjliga att förse med noter, textkritiska såväl som ursprungliga, och andra utgående länkar, bibliografiska referenser pekas till andra databaser.

Hur implementerar man detta?

Systemen som skall driva allt detta måste implementeras stegvis. Visioner måste vara högt ställda, och inget som implementeras i dag får vara sådant att det utesluter eller försvårar att man bygger vidare i framtiden.