

*Mats Dahlström*

Kulturarvet i Open Source

---

Paper presenterat vid konferensen

# Mötesplats inför framtiden

ARBETSLIV • UTBILDNING • FORSKNING

10-11 oktober 2007 i Borås



# Kulturarvet i Open Source

Mats Dahlström, Bibliotekshögskolan

## **Offentligt finansierade bibliotek**

Vi diskuterar vanligen offentligt finansierade bibliotek (OFB) i dessas förmedlande och dokumentbevarande roll. Samtidigt blir OFB allt flitigare inom digitaliseringen av kulturarvets artefakter. Det digitala resultatet publiceras och tillgängliggörs på nätet eller på ”stand alone”-medier såsom kompaktskivor. På det sättet intar OFB på ett mycket påtagligt sätt en nygammal roll: den som producent och förlag. Den är gammal därför att det finns ett tvåtusenårigt arv att förvalta när det gäller transmission av texter och bilder mellan generationer av bärare och medier. Detta har skett i både ett förmedlings- och bevarandesyfte – även om rollen som text- och bildöverförare, särskilt den textkritiskt kvalitetsgranskande, med seklerna kommit att hamna också på andra institutioners och aktörers bord (Dahlström 2006, s. 55f.). Den är gammal också i meningen att några OFB alltsedan den fotofaksimila teknikens intåg för något sekel sedan fullgör en i sammanhanget begränsad uppgift att i bevarande-, service- och i viss mån marknadsföringssyfte åstadkomma reproduktioner av dokument, framför allt bildmaterial. Men rollen är också ny: det digitala materialet förefaller erbjuda helt andra risker / möjligheter än analogt material när det gäller kopiering, spridning, manipulationer och annan återanvändning. Biblioteken har att fatta viktiga beslut när det gäller hur, i vilken form, i vilken utsträckning, när och för vem materialet kan och skall tillgängliggöras. Med de alltmer omfattande digitaliseringsprojekten och det därpå följande tillgängliggörandet av det digitala materialet finner sig OFB i ett allt skarpare och utmanande läge.

Parallellt med detta finns också andra centrala roller, uppgifter och värden för OFB, t.ex. att säkerställa individers jämlika åtkomst till den kollektiva och den nationella samlade utgivningen, att svara för service, hjälp och lärandestöd till både allmänhet, utbildnings- och forskarsamhälle, att såsom samhällets ”kollektiva minne” bygga upp och bevara öppet tillgängliga samlingar av repeterbara dokument. Det finns också inom OFB en stolt tradition att anamma föreställningar om värdenutralitet, icke-kommersialitet, ”informationens frihet” och de värden som brukar åberopas av de många ideologiska rörelserna med förtecknet ”open”: open access, open content, open source m. fl.

OFB:s digitaliseringsverksamhet av äldre text- och bildverk vars upphovsrätt löpt ut kan i ljuset av sådana värden beskrivas som ett sätt att sakta och med förenade krafter bygga upp ett slags ”digital commons”, en digital allmänning. Det finns sedan en tid tillbaka starka nationella och framför allt internationella politiska signaler till stöd för sådana strävanden, uttryckt exempelvis i Minervaprinциperna (se <<http://www.minervaeurope.org>>). En markering i sådana signaler är att den digitala allmänningen skall kunna stödja inte bara tillgodogörande och användning (use), utan också återanvändning (re-use), studier och forskning (se också se KB-utredningen 2003, s. 202).

## **Digitaliseringsprinciper**

Den digitalisering som utförs vid större OFB såsom national- och universitetsbibliotek är fortfarande i mångt och mycket i en experimentell och teknikutprovande fas. Några principer förefaller dock växa sig allt starkare, helt enkelt för att de svarar mot ett rationellt och ofta ekonomiskt motiverat sätt att arbeta på. För det första den s.k. one input / many outputs-principen. Genom att utnyttja mediets plasticitet och möjligheten att kombinera olika filtyper och programmering kan vi göra en arbetsfördelning mellan filer så att vi matar in exempelvis transkriptioner och andra källdata i en fil, beskrivningar på metanivå i andra filer, instruktioner för

framträdelseform, komposition och gränssnitt i ytterligare andra och så kan dessa olika typer av filer kombineras ihop till olika resultat beroende på vad användare och sammanhang behöver. En därmed besläktad princip är den om arkiv --> leverans. Vi konstruerar till att börja med en "tjock" masterfil med mycket information om objekten. I bilddigitaliseringssammanhang betyder detta högupplösta bildfaksimil i exv TIFF. För textdigitalisering kan det exemplifieras med en höguppmärkt text i en relationsdatabas eller i XML TEI P5, med rika metadata i exv en TEI Header innanför eller utanför textkällfilen. Textfilerna innehåller inte sällan i sin uppmärkningskod rikligt med analysgrundade beskrivningar av både originaldokumenten och de digitala surrogaten. Beskrivningar kan vara av exempelvis bibliografisk, texthistorisk, språkvetenskaplig, litteraturvetenskaplig eller stilistisk natur. Poängen är sedan att vi kan utnyttja principen om "many outputs" till att få denna informationstäta arkivfil att generera många olika typer av mindre informationstäta leveransformat. Från en TIFF kan vi få lättare bildfiler i JPEG eller GIF, ur en tjock TEI XML kan vi producera tunnare textfiler i XHTML eller PDF (liksom vi också kan utnyttja källfilen i TEI XML som master för produktionen av en tryckt bok), från TEI Header kan vi producera DC eller MARC. Och så vidare. En av nycklarna till detta arbetssätt heter transformation (genom t.ex. XSLT).

Principerna är också uttryck för ett långsiktigt tänkande kring digitalisering. Digitaliseringsprojekt är ofta långsiktiga och kan ibland löpa över tiotals år. De är ofta kostsamma saker, och det kan vara resursslöseri att investera i en teknik som, om den sedan går ur mode, tar med sig in i glömskan ett digitaliserat material som är tekniskt knutet till denna. Med andra ord: lägger vi ner stora summor på omfattande digitaliseringsprojekt, vill vi helst inte behöva göra det mer än en gång, allra minst efter bara några år. Den bästa lösningen är rika arkivformat och metadata som gör resurserna "flerspråkliga".

I digitaliseringssammanhang kan det uppstå behov av mängder av olika typer av utprodukter. I stället för att sjösätta projekt och arbetsflöden för alla upptänkliga framträdelse- och visningsformer, kan vi på detta sätt koncentrera oss på att tillverka en master samt ett antal maskininstruktioner, och sedan på automatisk väg och på begäran producera de olika leveransformaten. Digitalisera en gång och använd resultatet många gånger. Den digitala mastern blir en fertil källa som kan användas och framför allt återanvändas om och om igen. Detta är en princip som bl.a. förlagsindustri länge arbetat efter, och den har också vunnit alltmer mark inom bibliotekens digitaliseringsverksamhet, också som ett resultat av att bibliotek börjat fungera som samarbetspartner i stora nationella och internationella digitaliserings- och textutgivningsprojekt. De programmerare och forskargrupper ss textutgivare som biblioteken därmed finner sig samverka med (så exempelvis det intrikata och rika samarbetet mellan bibliotek, arkiv och text- och bildvetenskapliga forskare världen över inom ramen för det väldiga digitala projektet The William Blake Archive <<http://www.blakearchive.org>>) har förmodligen varit en bidragande kraft att införa vetenskapen om sådana här principer genom sitt tidiga intresse för SGML- och sedan XML-tillämpande digitalisering. Men bibliotekens policy vad gäller upphovsrätt, tillgängliggörande och filkvalitet blir i sådana sammanhang en väsentlig faktor som påverkar det digitala slutresultatets potential att möjliggöra inte bara användning utan också *återanvändning*.

## **Problemet**

Nu uppstår således ett problem, och det har att göra med tillgängliggörandet av det tjocka, informationstäta arkivmaterialet. Vem får tillgång till detta fertila material för användning och återanvändning? Med tiden har det alltmer uppstått en policy inom OFB att *inte* tillhandahålla arkivfilerna öppet, "open source" - den policyn är exempelvis helt eller delvis synlig hos Kungliga Biblioteket i Sverige, New York Public Library och The Bodleian Library i Oxford. Med andra ord har det utvecklats en princip att "hålla inne" med masterfiler och bara tillgängliggöra de tunnare och betydligt mindre fertila andrahandsformaten. Genom att behålla exempelvis TIFF-formaten kan

man också hitta en förlängning för en reproverksamhet där beställare åläggs betalning för reproduktionerna. Men det finns förstås mer direkta skäl till varför många OFB glidit in på det här spåret. De är av huvudsakligen tre typer: tekniska, administrativa och juridiska.

### **Det tekniska argumentet**

Länge framstod detta inte som något större problem för digitaliserande bibliotek. De mycket utrymmeskrävande arkivfilerna som framför allt de digitala bildfaksimilerna resulterade i föreföll länge hopplösa att tillgängliggöra via webben: det klena bredbandet gjorde överföring av filer på hundratals MB i praktiken ointressant. Den omständigheten börjar ändras, och argumentet kan om redan några år framstå som antikverat (arkiv- och styrfilerna inom textdigitalisering är förstås såsom mestadels rena textfiler en bokstavligen smal sak att distribuera). Och från producentens perspektiv bör själva tillhandahållandet av arkivfilerna inte erbjuda några särskilt svåra tekniska eller ekonomiska resurskrav utöver de man ändå måste möta genom tillgängliggörandet av leveransfilerna.

### **Det administrativa argumentet**

Ett argument som framförts i sammanhanget är att reproverksamhetens inkomster behövs för att täcka delar av digitaliseringens kostnader. Den undersökning HEDS gjorde 2002 av större europeiska bibliotek ger inget riktigt bra stöd för det argumentet (se Tanner & Deegan 2002). Ett mer övertygande argument handlar om osäkerheten kring hur det fertila källmaterialet kan såväl olovligen manipuleras som missbrukas genom att andra aktörer använder materialet för att åstadkomma derivat (såsom nya reproduktioner) i vinstsyfte. Detta är ett starkare argument. Det hävdas slutligen också att slutanvändare ändå inte är så pass intresserade av arkivfilerna att det skulle motivera kostnaderna att tillgängliggöra dem. Men i gruppen ”slutanvändare” ingår också forskare inom bok-, text-, litteratur- och språkvetenskaperna, och där är det lätt att hitta åtskilliga höjda röster som påkallar behovet av tillgång till informationstäta ”open source” arkivfiler, ett alldeles färskt exempel är Bodard och Garces (u.u.). Användare omfattar också elever, studenter och lärare i utbildningssammanhang. Monica Langerth Zettermans ägnar exempelvis sin pågående doktorsavhandling just åt på vilka sätt digitaliserat material kan fungera som *resurser* genom att kunna integreras respektive *återanvändas* i nya utbildningssammanhang (u.u., se också Broady 2001). Användargruppen kan slutligen också omfatta andra bibliotek, och det finns argument för att tillgängliggöra arkivfiler också bibliotek emellan – en omständighet vi kommer till nedan.

Förmodligen är det samtidigt också så, att många bibliotek ser de digitaliserade verken och deras texter som delar av ”sin samling”, och att bibliotek hämtar mycket av sin kraft och status genom den samling den hyser, dvs det finns en tradition att likställa fysiskt bestånd med institutionens identitet.

### **Det juridiska argumentet**

Slutligen finns det rent juridiska argument för att hålla tillbaka arkivfiler. Även om verken som representeras i de analoga originaldokumenten sedan länge blivit en del av den upphovsrättsfria allmänningen, kan det hävdas att själva det digitaliseringsarbete som gjorts (i form av t.ex. transkription (se Robinson 2006), optimering av bildfiler, fotografering och skanning, TEI-uppmärkning etc) representerar ett sådant mervärde och en så pass hög grad av andligt skapande insats att det digitala materialet erövrar upphovsrättsskydd (man skulle också möjligen kunna tänka sig att bibliotek åberopar mönsterskydd för den digitaliserade samlingen, precis som man gjort när det gäller databaslösningar för opac:er). På sätt och vis är detta korrekt. Derivata och fotografiska verk måste skyddas, och i många fall uppstår en ny upphovsrätt som måste respekteras. Vi återkommer nedan till den problematiken. Men samtidigt är argumenteringen egendomlig. Det

digitala materialet, i synnerhet om vi talar om bildfaksimiler, har som huvuduppgift att efterlikna originaldokumenten så mycket det bara går. Ju mer det lyckas efterlikna (dvs låtas vara) originalet, desto större dess värde som surrogat. Att hävda egenartsskydd för ett verk som gör allt det kan för att undvika egenart förefaller bakvänt.

## Behovet av arkivnivå för återanvändning

Men *varför* är bibliotekens restriktiva policy ett problem? Det är ett problem för bestämda användargrupper av följande anledningar.

Leveransfilerna är derivata inte bara i en teknisk-logisk utan också i en kvalitativ mening. De är komprimerade format där massor av information om såväl originaldokumenten som det digitala slutresultatet har sållats bort. Leveransmaterialet är skraddarsytt för bestämda publiceringsmål och passar därför bara väl för bestämda användargrupper, i bestämda situationer under en begränsad tid – det är själva poängen med leveransformat. De duger således bara i begränsad mening till användning, men de duger väldigt sällan alls till *återanvändning*. Det finns åtskilliga användargrupper, både forskare, textutgivare, studenter/elever och bibliotek, för vilka leveransfilerna därför är otillräckliga, och där det är viktigt med åtkomst till arkivnivån. Vi har redan sett imponerande resultat inom både bild- och textdigitalisering, och OFB börjar på många håll uppvisa en ambitiös och glädjande kompetensnivå när det gäller i första hand bilddigitalisering. Men vi har lärt oss hur teknik kan blända, och det vi idag upplever som sofistikerat tycker vi i morgon är gryntigt och tontigt, när den tekniska utvecklingen lärt oss att se på gränssnittet med nya, skarpare glasögon. För sin samtid framstod tidiga Carusoinspelningar under 1910-talet nästan spöklikt verklighetsåtergivande. Idag hör vi mest bara bruset. Och viktigare: i återanvändningssammanhang förblir hur som helst de derivata leveransfilerna otillräckliga och oundvikligen underkastade alla de tillval och bortval som digitaliseringsaktören gjorde under resans gång. Alla dessa val skriver in derivaten i en snäv marginal när det gäller till vilken återanvändning de kan tjäna.

## Lösningen

Bodard och Garces (u.u.) är två i en lång rad forskare som påkallat behovet av arkivåtkomst. De förklarar "open source" digitaliseringar med "the distribution of raw data, of scholarly tradition, of decision-making processes and of the tools and applications that were used in reaching these conclusions. The protocols and technologies for this manner of publication need to be made available and comprehensible to all textual scholars," liksom "full documentation of sources, references and arguments". Sådana forskare har behov av material som är "free to distribute, learn from, modify and re-distribute." Bodard och Garces kallar digitaliserat material utan open source tillgång för "a dead end; it cannot be built upon" (ibid. s. 3-5).

Bodard och Garces representerar forskarvärlden och dess intressen. Behovet av öppen och djup tillgång till informationstäta masterfiler inom också utbildning har poängterats (Langerth Zetterman u.u.). Men det finns även goda argument att hämta som avser *bibliotekens* vinster med ett open source-förfarande.

David Seaman skrev 2003 ett skarpt inlägg där han påkallade behovet av "deep sharing", ett perspektiv som ligger "open source" nära. Seaman pläderar för ett digitalt sambibliotek

from which libraries can draw files into local collections for innovative reuse and rearticulation as the needs of local users dictate. [I]t would enable librarians and end-users alike to download "digital master" files as malleable objects for local recombinations, to be enriched with context from librarians or teachers, crafted for specific audiences, and unified in appearance and function. A user could download, combine, search, annotate,

and wrap the results into a seamless “digital library mix” for others to experience. [A]t present, all you can do is scrutinize that data where it resides, in formats that the creator of the content determined (...) you can have a passive engagement with the content but not an active one. You cannot combine those scattered objects into something new, improved, and shaped for your local needs. (Seaman 2003, s. 10f.)

OFB är, påpekar Seaman, kollaborativa till sin natur, och pekar på prejudikatet hos samkatalogisering och fjärrlånekedjan. Detta samarbete skulle kunna omfatta också den digitaliserade fulltextnivån. I stället för att olika OFB på egna håll digitaliserar material, inte sällan sådant som redan digitaliserat(s) på annat håll, görs en ansvarsfördelning om vilket material som har och planeras digitaliseras, i vilka format, när, av vem, och under vilka administrativa, tekniska, upphovsrättsliga och bevarandemässiga villkor. Dessa uppgifter samt det digitaliserade materialet läggs sedan i ett kumulativt registerföret arkiv ur vilket de samverkande biblioteken kan hämta respektive lägga in digitaliserat material på arkivnivå, så att det kan återanvändas och anpassas för nya miljöer och användare. Arbetet med registeruppbyggnad har också påbörjats i biblioteksvärlden (se Scherman et al. 2005, s. 17), i stil med exempelvis de mikrofilmsregister som tidigare byggts upp. Den potentiella vinsten kan vara avsevärd. Digitaliserande OFB behöver i allt större utsträckning kunna bekräfta huruvida ett dokument redan har digitaliserats på annat håll, och i så fall huruvida detta skett på en så pass avancerad nivå att en ny digitalisering inte kan anses behövas. ”Imagine”, fortsätter Seaman om själva samdigitaliseringen,

that thirty libraries coordinate to digitize content out of their collections. They individually fund pieces of the endeavour, but all have access to the sum of their activity. For the cost of building one digital object at each institution and depositing it in the [archive], each library would gain thirty downloadable objects built to enduring archival standards. (ibid., s. 11)

Men digitaliseringen i OFB är så här i uppbyggnadsskedet värdefull inte bara för de digitaliserade dokumenten i sig, utan kanske än mer för den kompetens, teknik, lärdomar och erfarenhet som byggs upp vid varje projekt. I bibliotekssammanhang skulle vi därför med ”deep access” och ”deep sharing” se vinster inte bara med tillgång till själva det digitala materialet, utan också till den *kunskap* som genereras av att digitalisera kulturarv och kulturhistoriska allmänningar, dvs *dokumentationen* av projekten. Sådana sker fortfarande ofta i isolation och ”återuppfinner hjulet”. Det kan därför också på så sätt finnas vinster att göra om bibliotek kunde dra nytta av varandras kunskap sådan den är dokumenterad i arkivfilnivån.

Men hur skulle, slutligen, de säkerhetsmässiga och framför allt upphovsrättsliga problemen tacklas med digitaliseringar som i sig gör anspråk på verkshöjd? Det är frestande att peka på en lösning: att ansluta sig till någon av de standarder som innebär att materialet tillgängliggörs ”open access”. Creative Commons (CC) förefaller väl i dagsläget som den mest kraftfulla av sådana standarder, och är också den lösning EU:s Minervaprojekt rekommenderar för digitalisering i OFB-miljö. En lösning som CC innebär att upphovsrätten visserligen sätts ur spel – men bara delvis. CC tillåter specificering i olika nivåer av hur materialet får användas, av vem, i vilka syften och huruvida återanvändandet får ske i kommersiella eller icke-kommersiella sammanhang.

”Open access” är på många sätt ett behjärtansvärt mål också för bibliotek som producenter, men för de syften jag här pratat om skulle en vidgning till open source / deep access / deep sharing filtrerat genom en modererande CC kunna vara det sätt som OFB både kan öppna dörrarna till sin verkstad, och dessutom inte bara upprätthålla utan rentav stärka sina sociala, kulturella och samhälleliga värden i sin nya roll som digital producent.

## **Referenser**

Bodard, Gabriel & Juan Garces (u.u.). ”Open Source Critical Editions: A Rationale.” *Text Editing*,

*Print, and the Digital World*. Eds. Kathryn Sutherland & Marilyn Deegan. Aldershot: Ashgate.

Broady, Donald (2001). "Digitala arkiv och portföljer." *IT i skolan: mirakelmedicin eller sockerpiller?* (Rapport 45/2001). Stockholm: IT-kommissionen. 12-19.  
<<http://www.skeptron.uu.se/broady/dl/p-broady-010916-digitala-arkiv.htm>>

Dahlström, Mats (2006). *Under utgivning: den vetenskapliga utgivningens bibliografiska funktion*. (Ak.avh.). Borås: Valfrid. <<http://hdl.handle.net/2320/1738>>

KB-utredningen (2003). *KB - ett nav i kunskapssamhället*. (SOU 2003:129). Stockholm: Fritzes.

Langerth Zetterman, Monica (u.u.). *Innehållsdesign: verktyg, metoder och tillämpningar inom utbildningshistorisk forskning och undervisning*. (Ak.avh.). Uppsala universitet, Institutionen för utbildning, kultur och medier.

Robinson, Peter (2006). "The Canterbury Tales and other Medieval Texts." *Electronic Textual Editing*. Eds. John Unsworth, Katherine O'Brien O'Keefe & Lou Burnard. New York: Modern Language Association. (preprint: <<http://www.tei-c.org/Activities/ETE/Preview/robinson.html>>)

Seaman, David (2003). "Deep Sharing: A Case for the Federated Digital Library." *Educause Review* 38.4: 10-11. <<http://www.educause.edu/ir/library/pdf/erm0348.pdf>>.

Scherman, Anne et al. (2005). *DIGSAM: Digitalisering och dess samordning inom Kungl. biblioteket*. Stockholm: Kungliga biblioteket. (Rapport nr. 28).  
<<http://www.kb.se/Dokument/Om/projekt/digsam.pdf>>

Tanner, Simon & Marilyn Deegan (2002). *Exploring Charging Models for Digital Cultural Heritage*. Hatfield: University of Hertfordshire (The Higher Education Digitisation Service).  
<[http://heds.herts.ac.uk/mellon/charging\\_models.html](http://heds.herts.ac.uk/mellon/charging_models.html)>