

***Digitalisering av bøker i NB***  
***– metodikk og erfaringer***

**Nasjonalbiblioteket**

**september 2007**

# 1. Det digitale nasjonalbiblioteket

Nasjonalbibliotekets visjon er å være en levende hukommelse gjennom å utgjøre et multimedialt kunnskapssenter med fokus ikke bare på bevaring, men også på formidling.

For å lykkes med denne målsettingen er et av våre hovedmål å være et digitalt nasjonalbibliotek, og gjennom det også å være kjernen i Norsk digitalt bibliotek. Et digitalt nasjonalbibliotek er bare en annen måte å være nasjonalbibliotek på. Da er det viktig å ha så mye digitalt materiale som mulig, ikke bare historisk materiale, men også den moderne delen av kulturarven. Dette for å gi tilgang til så mye materiale som mulig for så mange brukere som mulig, når og der brukeren trenger det.

Nasjonalbiblioteket har derfor startet en systematisk digitalisering av hele samlingen. På grunnlag av en moderne pliktavleveringslov avleveres alt som produseres og er av interesse for publikum, enten det er bøker, aviser, tidsskrifter, fotografi, film, musikk eller kringkasting. Alle kringkastere med konsesjon i Norge kan bes om å avlevere opptak av sine programmer for bevaring, og vi har et tett samarbeid med NRK om bevaring og formidling. Vi skal også bevare de digitale signalene som aldri blir konvertert til noe annet før de når brukeren. Vi har siden 2005 høstet inn store deler av det norske nettdomenet .no.

Noe av dette materialet blir avlevert i digitalt format, og vi er i ferd med å få en forholdsvis stor digital samling når det gjelder det audiovisuelle, men mindre når det er snakk om trykt materiale.

Spørsmålet nå er hvordan vi skal dra nytte av å ha bygd opp et digitalt sikringsmagasin til også å gi tilgang til dette materialet for både forskere, studenter og publikum. Dette er en utfordring på mange måter, men først og fremst når det gjelder opphavsrett, teknisk kvalitet og formidlingspedagogikk.

Vårt mål, som er å etablere et enhetlig digitalt bibliotek for alle våre medier med basis i vårt digitale sikringsmagasin, krever følgende strategier:

- Snarest mulig (naturligvis avhengig av bevilgninger) å digitalisere alle våre samlinger. Dette gjøres systematisk.
- Å digitalisere etterspurt materiale
- Å forhandle med utgiverne om å få så mye som mulig avlevert i digitale format
- Å være et digitalt arkiv for andre institusjoner, f.eks. utgivere og aviser
- Å finne strategiske partnere å samarbeide med, økonomisk og faglig
- Å være et sikkert oppbevaringssted for digitalt materiale i det norske samfunnet
- Å gi tilgang til størst mulig del av den norske kulturarven, f.eks. gjennom søkemotorer
- Å forhandle med rettighetshaverne om tilgang til materiale som ennå ikke har falt i det fri, slik at også moderne bøker, filmer og musikk kan søkes i

For å kunne gjøre dette er det viktig å delta i diskusjoner og utvikle formidlingspedagogikk for nettet. Bibliotekenes kunnskap om brukernes behov er et godt grunnlag for å utvikle søkemetodikk i skjæringspunktet mellom metadata og søkemotorenes metoder. Vårt mål er å gi tilgang til informasjon, kunnskap og erfaringer om et gitt emne på tvers av medietyper.

## 2. Digitalisering av den norske kulturarven

### Omfang

Nasjonalbiblioteket har allerede i mer enn 10 år digitalisert et bredt spekter av medietyper. Hovedfokus for denne digitaliseringen har vært fotografier, lydfestinger og aviser (på mikrofilm). Som et resultat av dette inneholder den digitale samlingen i dag mer enn 150 000 timer med radio-sendinger, over 300 000 fotografi og mer enn 1 000 000 avissider. I tillegg har vi i løpet av det siste året digitalisert over 25 000 bøker.

Dette er likevel beskjedent sammenliknet med det vi nå har bestemt oss for å gjennomføre. Inkludert estimert tilvekst til samlingene gjennom pliktavlevering av materiale frem til digital avlevering er på plass, forventes omfanget av det som skal digitaliseres i Nasjonalbibliotekets digitaliseringsprogram å være:

- 450 000 bøker
- 2 000 000 tidsskrift
- 4 700 000 aviser (mer enn 60 000 000 sider)
- 1 300 000 bilder (foto og postkort)
  - 60 000 plakater
  - 200 000 kart
- 4 000 000 håndskrifter/manuskripter
  - 200 000 notetrykk
  - 1 900 000 småtrykk
- 1 000 000 timer radio
  - 80 000 timer musikk
  - 250 000 timer film og fjernsyn

Ev. nye gaver, innkjøp eller deponering av materiale i programperioden kommer i tillegg til dette.

For å klare å gjennomføre et så ambisiøst digitaliseringsprogram i løpet av en overskuelig tidsperiode, er digitaliseringsaktiviteten kraftig oppskalert. Det etableres nå en mest mulig strømlinjeformet prosess fra et objekt velges ut til digitalisering til det ligger digitalt bevart i vårt digitale sikringsmagasin og samtidig er tilgjengelig på nett til autoriserte brukere. Med Nasjonalbibliotekets multimediale samling gir dette spesielle utfordringer, da det må etableres egne produksjonslinjer for digitalisering av hver enkelt materialtype. I tillegg finnes det vanligvis flere varianter innenfor hver materialtype som igjen krever ulike tilpassinger.

Å få dette på plass er meget utfordrende både med hensyn til teknologiske løsninger, logistikk, organisatoriske løsninger, bemanning og økonomi.

### **Rettigheter**

Nasjonalbiblioteket har en lovfestet rett til å digitalisere samlingen for bevaringsformål. For å kunne gjøre den digitale samlingen tilgjengelig i det digitale biblioteket, må vi imidlertid ha avtaler med rettighetshaverne der materialet ikke har falt i det fri.

En del av materialet som digitaliseres, har falt i det fri, og det vil bli gjort tilgjengelig i vårt digitale bibliotek uten begrensninger.

I pliktavleveringsloven heter det at materiale som pliktavleveres, kan gjøres tilgjengelig for forskning og dokumentasjon. Dette gjelder også materiale som pliktavleveres digitalt. I det digitale området er det imidlertid mange uavklarte utfordringer, for eksempel hensynet til personvern i koplingen av ulike omfattende datakilder (inkl. nettsider høstet inn fra Internett), og at risikoen for og konsekvensene av misbruk gjennom kopiering av digitalt innhold, er mye mer omfattende enn for papirbaserte dokumenter. Dette gjør at vi fortsatt ikke har fullgode løsninger på plass for tilgang til digitalt pliktavlevert materiale via Internett.

For materiale som publiseres tradisjonelt, men avleveres digitalt etter avtaler med utgiverne, må vi gjøre avtaler med rettighetshaverne om tilgang. Her tilstreber vi å gjøre avtaler som i det minste gir oss samme rettigheter som for det pliktavleverte materialet.

Nasjonalbiblioteket ønsker også å gjøre avtaler med rettighetshaverne som gjør det mulig å gi bredere tilgang til deler av den moderne delen av kulturarven. Et eksempel på dette er [avtalen](#) mellom Nasjonalbiblioteket og flere rettighetsorganisasjoner om å gi tilgang til bøker og tidsskriftartikler relatert til nordområdene. Det er der gjort avtale om at Nasjonalbiblioteket kan gi fri tilgang via sitt digitale bibliotek til ca. 1400 verk i sin helhet. Disse verkene vil være utgangspunkt for å evaluere bruksmåte og bruksfrekvens for digitale bøker på nett. Resultatet av denne evalueringen vil være grunnlag for å forhandle om mer permanente avtaler med rettighetshaverne.

## **Økonomi**

Foreløpige estimater peker i retning av at Nasjonalbibliotekets digitaliseringsprogram vil koste totalt ca. en milliard kroner. Ca. 60 % av kostnadene vil gå til digitalt lager, innkjøp av digitaliseringsutstyr og programvare, utvikling og integrering av systemer som inngår i digitaliserings- og etterbehandlings-prosessen, samt til lønnsmidler til gjennomføring av selve digitaliseringen og til en del oppdragsmidler. De resterende ca. 40 % av kostnadene vil gå med til registrering av materiale for å etablere nødvendige metadata for gjenfinning, samt til uthenting av materiale fra samlingene, nødvendig konservering og tilbakeføring av materiale til samlingene.

Nasjonalbiblioteket har gjennom en omstilling av aktiviteten i retning av strategiske satsinger prioritert ca. 12 millioner kroner per år til digitaliseringsprogrammet. I tillegg har vi i 2007 fått et tilskudd fra vårt departement på 3 millioner kroner til digitaliseringen. På toppen av dette har vi et budsjett for innkjøp av digital lagringsplass på ca. 13 millioner kroner i 2007, så totalt budsjett for satsingen er ca. 28 millioner kroner i år.

Foreløpige estimater viser at hele digitaliseringsprogrammet bør kunne gjennomføres innenfor en tidsperiode på 15–20 år. Det er naturlig nok en betydelig usikkerhet knyttet til en så omfattende og langsiktig satsing som dette. Det forventes at teknologi vil utvikle seg i retning av billigere digitalt lager, raskere digitaliseringsutstyr og avansert automatisk annotering av materiale som kan gi enklere gjenfinning. Det er imidlertid vanskelig å estimere konsekvensene av en slik utvikling med nøyaktighet, både konsekvenser for totale kostnader og mulige effektiviseringsgevinster.

## **Status**

Store deler av 2006 gikk med til å gjennomføre anbudsprosesser for innkjøp av nødvendig digitaliseringsutstyr, programvare, digitalt lager og andre nødvendige IKT-løsninger. I tillegg ble det ansatt en prosjektleder og noen systemutviklere høsten 2006.

I den utvidede satsingen på digitalisering ble bøker prioritert først. En nærmere gjennomgang av arbeidet som er gjort, er presentert i et eget kapittel.

Vi har også etter en anbudsprosess satt ut oppdrag med digitalisering av aviser fra mikrofilm til et firma i Tyskland (CCS). Foreløpig har vi digitalisert ca. 400 000 sider fra avisene Aftenposten og Adresseavisen, men det vil bli satt ut et nytt oppdrag høsten 2007.

I tillegg til dette er digitaliseringen av fotografi effektivisert gjennom innkjøp av programvare for automatisert etterbehandling av digitalisert materiale, og gjennom oppgradering av de digitale kameraene som brukes til dette arbeidet. Det er også kjøpt inn effektivt utstyr for digitalisering av 35 mm filmnegativer på rull.

Vi jobber også med oppgradering av utstyr for digitalisering av musikk, radio og levende bilder. En viktig del av dette er en mer automatisert kopling mellom digitalisert materiale og metadata som beskriver materialet.

Digitalisering av andre materialtyper gjøres ved behov, men det er foreløpig ikke etablert fullverdige produksjonslinjer som sikrer rask overføring av det digitale materialet til vårt digitale sikringsmagasin, og at det blir gjort tilgjengelig i vårt digitale bibliotek.

Den systematiske digitaliseringen er grunnlaget for digitaliseringsaktiviteten. I tillegg vil vi gjennomføre on-demand-digitalisering som både kan være brukerstyrt og begrunnes i Nasjonalbibliotekets strategiske satsinger. On-demand-digitalisering gis prioritet over det systematiske arbeidet. Dette er gjennomført i alle ledd i digitaliseringsarbeidet. Så langt gjelder imidlertid dette først og fremst de materialtypene vi allerede har etablert produksjonslinjer for (dvs. bøker, fotografi og lyd).

### ***Plan fremover***

Det viktigste på kort sikt er å få alle produksjonslinjene for digitalisering av bøker i operativ drift. Det som gjenstår er først og fremst å få på plass et regime for kvalitetssikring av de digitaliserte bøkene, samt en god nok avvikshåndtering i produksjonsløypa som helhet. Dette inkluderer alle ledd i prosessen fra bestilling av materiale fra magasinene til de digitale utgavene av materialet ligger i det digitale sikringsmagasinet og er tilgjengelig i NBs digitale bibliotek.

Vi har også startet arbeidet med å etablere produksjonslinjer for å sikre både digitaliserte avissider og digitalt avleverte aviser i det digitale sikringsmagasinet, og for å gjøre dem tilgjengelige i vårt digitale bibliotek. I tillegg til avisene vi har fått digitalisert fra mikrofilm, har vi etablert en pilotløsning med daglig avlevering av to av landets største aviser i PDF-format i bevarings-kvalitet. Når denne løsningen går over i ordinær drift, vil vi jobbe for å utvide dette til å omfatte flere av landets aviser.

Vi planlegger videre å starte arbeidet med å etablere en produksjonslinje for digitalisering av manuskripter. Dette er et materiale som både er etterspurt av våre brukere, og som er aktuelt i forbindelse med flere forfatterjubileer i de nærmeste årene. For dette materialet må vi i tillegg til digitaliseringen etablere nødvendige metadata til å få en god nok gjenfinning i det digitale biblioteket.

Etter dette igjen vil det bli etablert produksjonslinjer for digitalisering av tidsskrifter og for digitalisering av plakater. I tillegg planlegger vi å automatisere deler av dagens produksjonslinje for digitalisering av foto for å øke effektiviteten.

Å få på plass avlevering av nyttinger i digitale format som er bevaringsverdige, er også en svært viktig aktivitet. I tillegg til pilotløsningen med daglig avlevering av to aviser, har vi i dag i operativ drift digital avlevering av radio i bevaringskvalitet. Videre avleveres nyproduserte lydbøker i digitalt format fra Norsk lyd- og blindeskriftsbibliotek.

Vi er også i dialog med forlag og fjernsynskringkastere, og vi planlegger å komme i gang med digital avlevering av bøker, tidsskrift og fjernsyn i løpet av det nærmeste året.

### **3. Etablering av en produksjonslinje for digitalisering av bøker**

#### **3.1 Utgangspunktet**

Da vi skulle etablere en helt ny prosess for digitalisering av bøker, var det viktig for oss å få til en mest mulig integrert produksjonslinje som kunne ivareta alle stegene boka skulle gjennom før den var ferdig bevart i NBs digitale sikringsmagasin, og samtidig tilgjengelig for autoriserte brukere i det digitale biblioteket. Her ønsket vi å automatisere så mye som mulig av dataflyt og behandling av den digitale boka, men samtidig å ha fleksibilitet til lett å kunne inkludere nye delprosesser i produksjonslinjen ved behov.

Prosessene vi ønsket å inkludere var: utvalg til digitalisering og bestilling, uthenting av materiale fra magasin, transport til digitalisering, hente metadata fra katalog, digitalisering, OCR-behandling og strukturanalyse, formatkonvertering, generere bevaringsobjekt, legge inn i DSM, varsle katalog om digitalt objekt og indeksere OCR-tekst og metadata i søkemotoren.

Vi så også at ulike typer skannerteknologi ga svært forskjellig effektivitet i digitaliseringen. Hvis bøker kunne demonteres kunne skanningen gjøres minst 10 ganger raskere.

#### **3.2 Overordnede valg**

##### **Bevaring – kvalitet og formatvalg**

Digitaliseringsprogrammet er en del av Nasjonalbibliotekets bevarings- og formidlingsstrategi. Digitaliseringen skal gjøre bevaringen av samlingen mer effektiv og mindre sårbar for fysisk nedbryting. Det betyr at digitaliseringen må gjøres i en kvalitet som er høy nok til at man gjennom digital bevaring på et senere tidspunkt kan gjenskape egenskapene til originalen på en tilfredsstillende måte. Samtidig skal digitaliseringen og valg av format oppfylle krav som stilles i formidlingen av materialet.

Vi har valgt å digitalisere bøker med en oppløsning på 400 dpi og en fargedybde på 24 bit. Vårt bevaringsformat er JPEG2000 med tapsfri kompresjon. Bildet som bevares, er ikke prosessert eller redusert på noen måte etter skanneprosessen.

Ved å velge tapsfritt komprimert JPEG2000 i stedet for ukomprimert TIFF som bevaringsformat, reduserer vi behovet for digital lagringsplass med ca. 50 %. For hele digitaliseringsprogrammet betyr dette en innsparing i størrelsesorden 70 millioner kroner. Vi har gjennom praktiske forsøk vist at vi fra JPEG2000-formatet kan konvertere tilbake til ukomprimert TIFF helt uten tap av informasjon.

Et argument mot å bruke JPEG2000 er at en bitfeil i et JPEG2000-bilde vil kunne ødelegge hele bildet, mens en bitfeil i et ukomprimert TIFF-bilde ikke ødelegger mer enn ett piksel. Med vår lagringspolicy i NBs digitale sikringsmagasin mener vi likevel at risikoen for bitfeil er neglisjerbar.

Kravene som stilles til kvalitet for bevaring, er langt strengere enn de krav man normalt vil stille til kvalitet for formidling. Samtidig har de formatene som brukes til formidling, kortere levetid, delvis fordi det stadig utvikles nye format med avanserte komprimeringsalgoritmer som gir bedre kvalitet med mindre data, og delvis fordi det utvikles nye versjoner av eksisterende format som har bedre algoritmer og gir bedre kvalitet. Vi har derfor valgt å generere formidlingsformatet fra bevaringsformatet i det øyeblikket et bilde etterspørres av en bruker. Med denne strategien kan vi lett endre formidlingsformat ved å bytte ut algoritmen som genererer formidlingsformatet.

I dagens løsning genereres en JPEG-fil i ønsket kvalitet for visning (typisk ca. 200 Kbytes) fra JPEG2000-filen som ligger i DSM (typisk ca. 20 Mbytes).

## **Digitalisere selv eller sette ut oppdrag til andre aktører?**

Flere aktører tilbyr i dag kulturinstitusjoner rimelig eller gratis digitalisering av boksamlingene mot at de får rett til å lagre de digitale bøkene selv, og til å tilby søk i og visning av bøkene. Eksempler på dette er Google og Internet Archive. I Nasjonalbibliotekets boksamling er det forholdsmessig få bøker som er falt i det fri, og som det dermed uten videre kan tilbys åpen tilgang til. Av de 450 000 titlene som skal digitaliseres i løpet av digitaliseringsprogrammet, er per i dag kun ca. 5 000 titler falt i det fri. For Nasjonalbiblioteket er det viktig at bøker som ikke har falt i det fri, kun skal lagres digitalt i Nasjonalbibliotekets digitale sikringsmagasin. Tilgang til slike bøker vil kun gis etter avtale med rettighetshaverne. Videre har det vært et grunnleggende prinsipp for Nasjonalbiblioteket at eksterne aktører skal ha lik tilgang til å tilby tjenester basert på våre samlinger. Summen av dette har gjort at det ikke har vært naturlig å samarbeide med denne typen aktører om digitaliseringen. Vi vil likevel tilby dem på lik linje med andre aktører å formidle det vi har digitalisert.

Det har også vært en del av bildet at Nasjonalbiblioteket har hatt mulighet for å omdisponere egne ansatte til oppgaver knyttet til produksjonsløypen for digitalisering. Dette har gjort at intern digitalisering av bøker har gitt et gunstigere kostnadsbilde enn å sette ut oppgaven til andre.

For andre typer materiale har vi valgt å sette ut digitaliseringsoppdrag. Dette gjelder for eksempel digitalisering, OCR-behandling og strukturanalyse av aviser på mikrofilm.

## **Demontering av bøker**

For å få til en effektiv digitalisering har vi valgt å demontere bøker for digitalisering når vi har minst tre eksemplarer av bøkene i vårt depotbibliotek. Det demonterte eksemplaret blir da kassert etter digitalisering.

Når vi har færre eksemplarer, blir bøkene skannet manuelt ved at operatører blar gjennom bøkene og digitaliserer to og to sider. For de mest sårbare bøkene skal skanningen gjøres under oppsyn av en konservator, og ev. nødvendig konservering utføres i forkant eller som en del av digitaliseringen.

Prosessen med å forberede bøker for demontert skanning er mer arbeidsintensiv enn prosessen med å forberede bøker for manuell skanning. Det kreves egne operatører for demontering av bøkene (skille perm fra resten av boka, samt klippe bort lim med hydraulisk saks), og skanningen av permene er en egen separat prosess. På grunn av dette kreves ca. 4 personer for å holde i gang én skanner for demonterte bøker. Likevel gir totalbildet både lavere kostnader og høyere produksjon enn om de samme ressursene hadde blitt benyttet til manuell skanning.

Per i dag vil ca. en fjerdedel av titlene i Nasjonalbibliotekets boksamling kunne demonteres for digitalisering. For å øke denne andelen planlegger Nasjonalbiblioteket å invitere bibliotekene i

Norge til å sende oss eksemplarer av bøker vi har for få av. På denne måten vil digitaliseringen av bøkene kunne gjøres raskere og mer effektivt i Nasjonalbiblioteket, og bibliotekene får frigitt plass i sine magasiner.

Foreløpig har vi ikke prøvd ut skannere med automatisert blasing i bøker som ikke er demontert. Slike skannere er imidlertid i stadig utvikling, og denne teknologien begynner i dag å bli svært interessant. For bøker som ikke kan demonteres, vil denne teknologien kunne gi betydelig høyere produksjon per operatør, både fordi skannerne er både raske og skånsomme med materialet, og fordi én operatør kan betjene flere slike skannere samtidig. Investeringskostnaden for denne typen utstyr er imidlertid fortsatt høy.

## **OCR og strukturanalyse**

For å gjøre det mulig å søke i fulltekst kjøres alle digitaliserte bøker gjennom en OCR-prosess. I ordinær produksjon gjøres denne prosessen helautomatisk, og det gjøres ingen manuell kvalitetskontroll eller oppretting. Teksten som fremkommer ved OCR-behandlingen, indekseres i vår søkemotor sammen med metadata. Ved søketreff i teksten gis man tilgang til den siden i boka der teksten ble funnet og kan bla videre derfra.

Det gjøres også en automatisert strukturanalyse der ev. innholdsfortegnelse annoteres, og der sidenummer i boka verifiseres slik at vi i visningsgrensesnittet forholder oss til den faktiske pagineringen i boka. Dette gjøres også helautomatisk. Programvaren har muligheter for svært avansert strukturanalyse, men det er foreløpig vanskelig å øke kompleksiteten uten å ha omfattende manuell kvalitetssikring og etterkontroll. For utvalgte deler av samlingen vil det bli gjennomført mer avansert strukturanalyse med annotering av flere deler av dokumentene, som igjen gir mulighet for mer avansert manøvrering i bøkene i visningsgrensesnittet.

Det digitaliseres for tiden 2 000–3 000 bøker hver måned i Nasjonalbiblioteket. Med dette volumet er det ikke gjennomførbart å gjøre manuell etterkontroll av OCR-behandling og strukturbehandling.

Både OCR og strukturanalyse gjøres ved bruk av programvaren docWorks.

## **Nasjonalbibliotekets digitale sikringsmagasin**

Det digitale sikringsmagasinet er en infrastruktur for å bevare digitale objekter over lang tid. Alt som digitaliseres som en del av NBs digitaliseringsprogram, skal bevares som digitale objekter i Nasjonalbibliotekets digitale sikringsmagasin.

Det digitale sikringsmagasinet gjør at bruken av digitalt innhold frikoples fra teknologien som brukes til selve lagringen. Det gjør at vi enkelt kan migrere til nye generasjoner av lagringsteknologi uten at systemene som henter digitalt innhold, berøres. I et tusenårsperspektiv er dette veldig viktig.

Alt digitalt innhold lagres i tre kopier på to ulike lagringsmedier i det digitale sikringsmagasinet. For tiden lagres en kopi på disk mens to kopier lagres på tape.

## **Søkemotor**

For å kunne realisere søk i store datamengder har Nasjonalbiblioteket valgt å basere seg på søkemorteknologi heller enn tradisjonelle databaseløsninger. Både metadata og fulltekst indekseres i søkemotoren, og søk gjøres på tvers av materialtyper. Det er også implementert såkalt drill-ned-søk i metadata. Metadata for objekter som tilfredsstiller søkekriteriene, analyseres i

sanntid ved søk, og det bygges opp alternative manøvreringsveier og ulike måter å avgrense søketreffet på basert på innholdet i metadataene.

Søkemotoren som benyttes, er levert av Fast.

## **Autentisering og autorisering - tilgangskontroll**

Nasjonalbiblioteket har valgt å benytte rollebasert tilgangskontroll. Videre har vi valgt å samarbeide med Feide, som er den nasjonale infrastrukturen for autentisering og autorisering av brukere ved universiteter og høyskoler i Norge. Det betyr at vi kan åpne tilgang for definerte grupper av forskere ved universitetene og høyskolene uten å måtte kjenne til hver enkelt person. Universitetene står selv for autentisering av de som oppfyller kravene til ulike roller definert i systemet.

Hvis vi hadde valgt en tilgangskontroll basert på brukernavn og passord med tilknyttede tilgangsrettigheter for enkeltpersoner, ville Nasjonalbiblioteket måtte brukt mye ressurser på administrasjon av brukere og vedlikehold av tilgangsrettigheter.

Foreløpig har vi ikke implementert løsningen for tilgangskontroll for digitale bøker. Vi har derfor kun indeksert og gjort tilgjengelig bøker som enten har falt i det fri, eller bøker der vi har avtale med rettighetshaverne om å gi åpen tilgang.

Programvaren som benyttes til autorisering og autentisering av brukere, er levert av Sun Microsystems.

## **3.3 Produksjonslinjene**

### **Prioriteringer**

Basis for digitaliseringen er det systematiske uttaket. Vi har valgt å starte med det eldste materialet for raskt å få materialet som har falt i det fri ut i vårt digitale bibliotek. I tillegg til det systematiske uttaket prioriteres materiale spesielt med utgangspunkt i interne behov og eksterne forespørsler. Spesielt prioritert materiale gis prioritet foran det systematiske uttaket.

Et spesielt tilfelle er avtalen Nasjonalbiblioteket har gjort med flere rettighetsorganisasjoner om bøker og tidsskriftartikler relatert til nordområdene. Disse verkene er gitt særskilt prioritet i digitaliseringsarbeidet.

### **Bestilling og uttak fra magasin**

For å effektivisere uttak av materiale til digitalisering, er det utviklet en egen funksjonalitet for dette i Bibsys som er vårt katalogsystem for bøker. Her kan vi bestille ut et gitt antall titler til demontering, der systemet automatisk velger titler vi har mange nok eksemplarer av, og starter med det eldste. I tillegg kan vi bestille ut enkelttitler som skal prioriteres spesielt (både ved uttak fra magasin og gjennom hele produksjonslinjen). Det er også gjort tilpassinger i programvaren som styrer vårt automatlager for bøker, slik at operatørene kan prioritere fjernlån først, og deretter ta ut bøker til digitalisering. Dette systemet er integrert med katalogen, slik at bøkene som bestilles til digitalisering, automatisk dukker opp i grensesnittet til operatørene av automatlagret.

Det er allerede brukt mer enn ett årsverk til systemtilpassinger av katalogen og programvaren for automatlagret.

## **Digitaliseringen**

For bøkene som demonteres, har vi i dag to hydrauliske sakser, tre permskannere (i2s Copibook) og to skannere med automatisert fremtrekk (Agfa S 655). For bla-skanningen brukes i2s Digibook Suprascan. Der har vi fem A2-skannere for normal bla-skanning og en A0-skanner for spesielt materiale. A0-skanneren brukes av konservatorer.

Før permene skannes, hentes alle metadata om boka inn fra katalogen (Bibsys) ved å bruke en strekkode som finnes på alle bøkene som er registrert i Bibsys. Det genereres da en digital id for boken som legges inn i en XML-fil sammen med de metadataene som er hentet fra katalogen.

For autoskanningen skrives det etter permskanningen ut et ark med en ny strekkode som inneholder bokens digitale id. Dette arket legges øverst i bunken med den demonterte boka. Når strekkoden senere kjøres gjennom autoskanneren, identifiseres strekkoden. Dermed koples sidene i boken automatisk til metadatafilen og den innskannede permene.

For bla-skanningen skannes permene og innholdet i boka på samme skannerutstyr. Også i denne prosessen hentes metadata fra katalogen, og det genereres en XML-fil med metadata som følger boka videre i prosessen.

## **OCR/DSA**

Etter digitaliseringen legges den digitale boka med tilhørende metadata i et temporært lager klar for videre prosessering. Bøkene må importeres manuelt inn i programvaren docWorks, men derfra er prosesseringen av de fleste bøkene helautomatisert. Manuelle operatører brukes kun til avvikshåndtering når programvaren melder om feil i behandlingen av boka (dvs. at behandlingen ikke lyktes innenfor definerte grenseverdier for feiltoleranse).

I tillegg brukes operatører til kvalitetskontroll for spesielle deler av samlingen som vi ønsker å behandle utover det normale.

Bøker som har prioritet, legges i egne mapper som importeres før den normale systematiske digitaliseringen.

Etter OCR og dokumentstrukturanalyse genereres tapsfritt komprimerte JPEG2000-filer av alle bildefilene i boka. Dette er formatet som brukes til bevaring.

## **Digital bevaring**

Etter behandlingen i docWorks genereres et METS-objekt med metadata, den digitale boka, den OCR-behandlede teksten og strukturinformasjon. Dette objektet legges inn i NBs digitale sikringsmagasin for bevaring.

Samtidig oppdateres katalogen med bokas digitale id.

## **Indeksering**

Det gjøres jevnlig en OAI-import av data fra katalogen. Hvis denne importen avdekker at en bok er blitt oppdatert med digital id, iverksettes en prosess som henter metadata og teksten til boka fra det digitale sikringsmagasinet og indekserer begge deler slik at boka blir tilgjengelig for søk i NBdigital.

## 4. Viktige erfaringer så langt

### Omfang, kompleksitet og gjennomføring

Å implementere en integrert produksjonslinje med høy grad av automatikk for digitalisering av bøker, viste seg å være langt mer omfattende og komplekst enn først antatt.

For ikke å tape tid ble det etter vedtaket om å etablere en produksjonslinje for digitalisering av bøker, så snart som mulig satt i gang aktiviteter for å realisere den første delen av produksjonskjeden (bestilling, uttak, transport og selve digitaliseringen). For å realisere uttak av materiale på en effektiv måte, måtte det gjøres tilpassinger både i Bibsys, som er Nasjonalbibliotekets katalogsystem for bøker, og i programvaren som styrer automatlagret som bøkene hentes ut fra. Dette ga avhengighet til to eksterne leverandører, noe som ga føringer på fremdriften i utviklingen.

Å gjennomføre en anbudsprosess for innkjøp av skannere er også en lang og tidkrevende prosess. Utvikling av opplegget for selve digitaliseringen kunne ikke gjøres før det var avklart hva slags utstyr som skulle brukes, og videre måtte vi få på plass de første skannerne før vi kunne avslutte implementeringen og starte testing.

Da denne første delen av produksjonslinjen var på plass, startet vi prøveproduksjon. Med den høye produksjonstakten vi la opp til, førte dette raskt til svært store mengder data på midlertidig lager. I påvente av at resten av produksjonsløypen kom på plass måtte det etableres midlertidige rutiner for å sikre det digitale innholdet.

For å få på plass resten av produksjonslinjen og funksjonalitet i vårt digitale bibliotek som gjorde det mulig å søke i og vise frem bøker, måtte et mangfold av utviklingsaktiviteter settes i gang. Eksempler er: Installering og idriftsetting av programvare for OCR og dokumentstrukturanalyse, og integrering av dette systemet i produksjonslinjen, generering av bevaringsobjekt basert på Mets-standarden, prosess som legger Mets-objektene inn i DSM, opplegg for å legge digital id inn i katalogsystemet, OAI-innhøsting av metadata fra katalogsystemet, prosess som iverksetter uthenting av tekst og metadata fra DSM for bøker som har fått digital id, samt indeksering av disse, og utvikling av nødvendig funksjonalitet for søk i og visning av bøker i det digitale biblioteket. Det oppsto samtidig et stort press for raskt å få se resultatene av det digitaliseringsarbeidet som allerede var i gang, noe som førte til pressing av tidsfrister og i en periode svært høyt stressnivå i utviklingsseksjonen som hadde ansvar for utviklingsarbeidet.

Da funksjonaliteten for visning av bøker var på plass i det digitale biblioteket, ble det raskt klart at dette ville bli en meget interessant tjeneste som ville gi et løft til vårt digitale bibliotek. Samtidig hadde vi allerede hatt mye mediedekning på digitaliseringsarbeidet, og forventningen til å få se resultatene var store både fra eksterne brukere og internt i Nasjonalbiblioteket. Vi bestemte oss derfor for å lansere tjenesten, selv om produksjonslinjen som sådan ennå var under utvikling, noe som betydde at det måtte mange manuelle operasjoner til for å få en bok gjennom hele produksjonslinjen. Tjenesten har fungert bra, men forventningene om raskt å få større volum med digitale bøker i tjenesten klarte vi ikke å innfri. Dette skyldtes hovedsakelig at produksjonslinjen ikke var ferdigstilt og følgelig heller ikke satt i ordinær drift.

I ettertid er det lett å se at man fra starten av burde hatt større fokus på helheten i utviklingen av både hele produksjonslinjen og den nødvendige funksjonaliteten i det digitale biblioteket, og at man

allerede i starten burde ha forankret eierskap til tidslinjen i dette utviklingsarbeidet bedre i hele organisasjonen.

### **Faktisk effektivitet**

Med utgangspunkt i digitaliseringsutstyrets spesifikasjoner ble det satt opp produksjonsmål helt fra starten av. Disse tok hensyn til at vi var i en innkjøringsfase. Det viste seg likevel at en del ting vi ikke hadde vært klar over, reduserte den totale effektiviteten. Dette ble mest synlig for de automatiserte skannerne.

Papiret i bøkene var gjennomgående tykkere enn referansepapiret som var brukt i målingene i skanneres spesifikasjoner. Dette førte til at hastigheten på papirfremføringen gikk ned, og det ga store utslag på en dagsproduksjon.

Siden vi begynte med de eldste bøkene, ble vi plaget med at de var veldig støvete. Dette betydde at skannerne måtte ha et betydelig mer omfattende renhold enn det leverandøren hadde forventet. Igjen ga dette seg utslag i en redusert produksjonstid per dag, og dermed også redusert dagsproduksjon i forhold til forventningene.

Vi hadde lagt opp til at skannerne skulle gå mest mulig kontinuerlig hele arbeidsdagen. Dette skulle realiseres ved at man avløste hverandre på denne arbeidsoppgaven og tok pauser på forskjellige tidspunkt. Dette var en ny og uvant måte å jobbe på, og det skapte en viss motstand hos operatørene. I praksis har vi ikke klart å realisere dette godt nok, noe som også har gitt redusert produksjonstid per dag på skannerne i forhold til prognosene.

Faktisk produksjon har vært mellom 60 % og 80 % av det vi hadde som produksjonsmål.

### **Kvalitet**

I den aller første testfasen ble alle sider i de digitaliserte bøkene kvalitetssikret. Etter dette har vi ikke hatt rutiner for kvalitetssikring av digitaliseringsjobben. Etter at funksjonaliteten i det digitale biblioteket ble klar for bokvisning, har det vært gjort usystematisk sjekk av kvaliteten på det som er tilgjengelig i tjenesten.

Vising i det digitale biblioteket avslørte at komprimeringen av de digitale visningskopiene ble gjort med ugunstige parametere. Den visuelle kvaliteten på bildene ble derfor dårligere enn forventet. Dette ble justert, og resultatet ble mye bedre.

Kvaliteten på de digitale bøkene henger naturlig nok tett sammen med kvaliteten på originalen. Vår algoritme for automatisk utplukk av bøker til digitalisering tar ikke hensyn til dette, og vi risikerer følgelig å få ut meget dårlige eksemplarer til digitalisering, selv om vi faktisk har gode eksemplarer i samlingen.

Skannerne som digitaliserer demonterte bøker, skanner begge sidene av hvert ark i én operasjon. Det betyr at det er to forskjellige digitaliseringsenheter som digitaliserer de to sidene på et ark. Her har det vist seg å være meget vanskelig å kalibrere disse to enhetene helt likt, noe som gir seg utslag i form av fargeforskjeller på sidene. Dette har blitt kraftig forbedret siden testfasen, men problemet er fortsatt ikke helt løst. Det har vært gjort forsøk med å skanne inn et referanseark i begynnelsen av hver bok, for dermed å kunne justere ev. fargestikk automatisk i etterkant. Så langt har ikke disse forsøkene gitt ønsket resultat, men vi vil jobbe videre med å få dette til.

Når produksjonslinjen går inn i en ordinær produksjonsfase, ønsker vi å etablere stikkprøvebasert kvalitetskontroll av digitaliserte bøker.

## **Gjennomføring av OCR/DSA**

I utgangspunktet hadde vi planlagt helautomatisk bruk av verktøyet for OCR og dokumentstrukturanalyse. Denne typen arbeidsoppgaver hadde ikke tidligere vært gjennomført i så stor skala, og vi hadde ingen erfaring i bruk av så avanserte verktøy.

Første utfordring var å få på plass et storskala produksjonsopplegg med åtte instanser av programvaren på hver sin blade-server. Dette var nødvendig for å ha tilstrekkelig prosesseringskapasitet, men det viste seg å være mer krevende enn antatt å få dette stabilt og operativt.

Neste utfordring oppsto da det viste seg at det ikke var mulig å kjøre systemet helautomatisk. Det oppsto dermed et uforutsett ressursbehov, og bemanningen av denne arbeidsoppgaven voldte en del hodebry. Vi brukte en del tid på å finne ut av dette, og det førte til at den planlagte opplæringen i systemet ble utsatt. Dette ga igjen et kompetanseproblem på kort sikt siden systemet er komplekst og har svært avansert funksjonalitet. Dette ble til dels løst gjennom nær kontakt mellom oss og leverandøren. Denne utfordringen er nå løst, og opplæring er gjennomført.

Våre opprinnelige forventninger til presisjon i den helautomatiske strukturanalysen har så langt ikke blitt innfridd. Det lar seg gjøre å gjennomføre avansert strukturanalyse, men usikkerheten i denne er så vidt stor at det er helt påkrevd med manuell kvalitetskontroll. Jo mer avansert analyse man ønsker å benytte, jo mer manuell etterkontroll kreves. Dette vil derfor kun bli brukt unntaksvis i spesielle formidlingsprosjekter. Et enkelt regnestykke viser at en etterkontroll som i snitt krever 15 sekunder per side, totalt vil kreve en innsats tilsvarende 18 dagsverk per dag med dagens produksjon. Dette har vi ikke ressurser til. Vi har derfor vært nødt til å legge oss på et absolutt minimum ved kun å kreve riktig paginering i tjenesten i det digitale biblioteket samt at en ev. innholdsfortegnelse skal kunne lenkes direkte.

Så langt har vi hatt fokus på OCR og DSA av utgivelser med latinske bokstaver. Her har vi en akseptabel presisjon i bokstavgjenkjenningen. For gotiske bokstaver er resultatet dårligere, men også der opplever vi en gjenkjenning som gir interessante muligheter til fritekstsøk. Vi kjører separate konfigureringer av systemet for latinske og gotiske bokstaver. Bøkene kategoriseres når permene skannes inn, og rutes deretter inn til riktig konfigurasjon. Her antas det å ligge et forbedringspotensial i ytterligere opplæring av programvaren og mer avansert konfigurering av systemet.

## **Det digitale sikringsmagasinet – skalering og ytelse**

Så langt har det vært kjørt en egen instans av DSM (digitalt sikringsmagasin) for bokdigitaliseringen. Foreløpig bruk tyder ikke på ytelsesproblemer knyttet til DSM, men bruken må forventes å øke mye i forhold til dagens trafikk etter hvert som volumet av digitale bøker i tjenesten øker.

Siden vi ikke har på plass en tilgangsløsning for bøker (autorisering og autentisering av brukere), har vi valgt foreløpig kun å legge inn i DSM bøker som vi kan gi tilgang til i det digitale biblioteket. En del av logikken i dag er at man ved innlegging i DSM også legger bokas digitale id inn i katalogen. Dette fører i neste omgang automatisk til at boka blir hentet ut fra DSM for indeksering til søkeindeksen til vårt digitale bibliotek.

Denne strategien betyr at brorparten av de digitaliserte bøkene fortsatt ligger på mellomlager, dog sikret med samme lagringspolicy som de som er lagret i DSM.

Når en bruker forespør en gitt bokside i en gitt kvalitet (p.t. kan man velge mellom tre kvalitetsnivå i grensesnittet), så genereres det automatisk en JPEG-fil i ønsket kvalitet ut fra JPEG2000-filen i DSM. Det er foreløpig ikke lagt opp til intelligent forhåndscaching eller buffering av sider ut fra DSM for å øke ytelsen sett fra brukersiden, men dette kan være en interessant videreutvikling.

Så langt har teknologiutviklingen gitt stor nok fleksibilitet til at vi kan operere med en en-maskin-løsning for DSM. Dette gir likevel sårbarhet i tilfelle feilsituasjoner. Mer robuste løsninger for DSM vurderes derfor kontinuerlig.

### **Statistikkmuligheter – produksjonsoppfølgingsverktøy**

Så langt har vi brukt enkle Unix-verktøy for å generere nødvendig statistikk til å kunne følge med produksjonen. Vi vurderer nå å utvikle et mer avansert generelt produksjonsoppfølgingsverktøy som kan gi oss til enhver tid oppdatert informasjon om hvor i prosessen et gitt objekt befinner seg, og som samtidig kan brukes til å generere produksjonsstatistikk for alle ledd i produksjonskjeden.

### **Avvikshåndtering**

Med få unntak er nå alt i produksjonslinjen som kan automatiseres, automatisert. Det oppstår likevel avvik i alle ledd av produksjonslinjen, og disse avvikene må håndteres og følges opp manuelt. Dette har vært en av de største utfordringene i produksjonen så langt. Her jobbes det nå med å utvikle rutiner og presisere ansvar for slik oppfølging i linjeorganisasjonen.

## **5. Oppsummering**

Implementeringen av en strømlinjeformet produksjonslinje for bøker har ikke vært uten problemer. Vi har imidlertid lært av de feilene vi har gjort underveis, og vi har i dag en avansert produksjonslinje for bøker på plass. I løpet av året vil produksjonslinjen bli satt i ordinær drift, og de siste brikkene med plassering av ansvar og avvikshåndtering vil komme på plass.

Til tross for utfordringene vi har hatt underveis, har vi hatt en betydelig produksjon i løpet av det året vi har gått i prøvedrift. Nesten 26 000 bøker, med et snitt på over 200 sider per bok, er nå ferdig digitalisert, og de fleste av disse har også vært gjennom OCR og strukturanalyse. Litt over 1 500 bøker er fritt tilgjengelig i sin helhet i vårt digitale bibliotek, der de også er søkbare i fulltekst.

Utfordringen fremover er å etablere produksjonslinjer for alle materialtypene som skal digitaliseres, slik at vi for alvor kan få på plass det multimediale digitale nasjonalbiblioteket.