

Digitalisering av bøker i NB
– metodikk og erfaringer

Nasjonalbiblioteket

April 2009

1. Etablering av en produksjonslinje for digitalisering av bøker

1.1 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, etterprosessering av digitalt bilde, 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.

1.2 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.

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. I dag genereres JPEG-bilder som formidlingsformat som en del av digitaliseringsprosessen, mens en PDF-versjon av en bok genereres i det øyeblikket dette formatet etterspørres av en bruker. Med denne siste strategien kan vi lett endre formidlingsformat ved å bytte ut algoritmen som genererer formatet.

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. 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. Bøkene blir da digitalisert i automatiserte skannere som hver klarer ca 100 bøker per dag. Det demonterte eksemplaret blir da kassert etter digitalisering.

Når vi har færre eksemplarer, blir bøkene enten digitalisert i en skanner som automatisk blar gjennom og digitaliserer boka, eller den blir digitalisert manuelt ved at operatører blar gjennom boka og digitaliserer to og to sider. For de mest sårbare bøkene skal skanningen gjøres i samarbeid med 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 5 - 6 personer for å holde i gang to skannere 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.

For bøker som ikke kan demonteres, gir skannere som blar automatisk betydelig høyere produksjon per operatør enn de manuelle skannerne. Dette både fordi skannerne er 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 av OCR-teksten. 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. Tagging av innholdsfortegnelse kvalitetssjekkes, og det gjøres manuell sjekk når systemet varsler inkonsistens i sidenummer. 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 3 000 – 4 000 bøker hver måned i Nasjonalbiblioteket. Med dette volumet er det ikke gjennomførbart å gjøre manuell etterkontroll av OCR-behandling og strukturanalyse.

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.

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. I forbindelse med Bokhyllasatsingen er materiale fra 1690-1699, 1790-1799, 1890-1899 og 1990-1999 gitt spesiell prioritet.

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 eksemplar 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 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. I tillegg har vi en skanner som blir automatisk (4digitalBooks DL3000).

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 permen.

For bla-skanningen skannes permen 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 verifisering av tagging av innholdsfortegnelse og ved avvikshåndtering når programvaren melder om feil i behandlingen av boka (dvs. at behandlingen ikke lyktes innenfor definerte grenseverdier for feiltoleranse, eller at det varsles om inkonsistens i sidenummer).

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

Fargekorrigering

Bøkene som behandles av manuelle skannere eller skannere som blir automatisk, gis et autentisk uttrykk ved at fargene i den digitale boka skal være lik fargene i originalen.

Bøkene som demonteres og skannes i automatiserte skannere fargekorrigeres etter skanning. Målet der er å bringe bøkene nært opp mot bokas tilstand når den ble utgitt, og at det dermed skapes et uniformt uttrykk for disse bøkene. Color Factory brukes til fargekorrigeringen.

Etter OCR, dokumentstrukturanalyse og evt. fargekorrigering, genereres tapsfritt komprimerte JPEG2000-filer for bevaring og JPEG-filer for formidling av alle bildefilene i boka.

Digital bevaring

Når bevaringsformatet er klart 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.