


Horsk digitalt bibliotek



NDB-Rammeverk


Arbeidspakke 2: Produksjon av og tilgang til dokumenter

Sluttrapport


 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato 27.12.2005		
	Side 2 av 15		
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

INNHold

1	OPPSUMMERING	2
2	INNLEDNING	2
2.1	HENSIKT	2
2.2	ARBEIDSGRUPPENS SAMMENSETNING	2
2.3	MANDAT	2
2.4	MÅLGRUPPE.....	2
3	BAKGRUNN FOR ANBEFALINGENE.....	2
3.1	OMFANG	2
3.2	TILNÆRMINGSMÅTE.....	2
3.3	GRUNNLEGGENDE FORUTSETNINGER	2
3.4	FORHOLD TIL ANDRE ARBEIDSPAKKER.....	2
4	NDB-DOKUMENT	2
4.1	NDB-METADATA	2
4.2	NDB-FORMATER.....	2
4.2.1	<i>Tekst eller tekstdelen av et dokument</i>	2
4.2.2	<i>Multimediedata</i>	2
4.2.2.1	Bilder	2
4.2.2.2	Video.....	2
4.2.2.3	Lyd.....	2
5	PRODUKSJON AV OG TILGANG TIL DOKUMENTER.....	2
5.1	STANDARDFORMAT FOR DIGITAL PUBLISERING.....	2
5.1.1	<i>Dokumenttyper</i>	2
5.1.2	<i>Tegnsett</i>	2
5.1.3	<i>Protokoller for dataoverføring</i>	2
5.1.3.1	Protokoll-typer	2
5.2	VERKTØY	2
5.2.1	<i>Tekstbehandling</i>	2
5.2.1.1	Teksteditor	2
5.2.1.2	Innholdseditor	2
5.2.1.3	Tekstgjenkjenning (OCR)-verktøy.....	2
5.2.1.4	Andre komponenter	2
5.2.2	<i>Bildebehandling</i>	2
5.2.2.1	Andre komponenter	2
5.2.3	<i>Lydredigering</i>	2
5.2.3.1	Andre komponenter	2
5.2.4	<i>Videoredigering</i>	2
5.2.4.1	Andre komponenter	2
5.2.5	<i>Metadata-registrering (katalogiseringsprogram)</i>	2
5.2.5.1	Andre komponenter	2
5.3	TJENESTER	2
5.3.1	<i>Tjenester for gjenfinning og framvisning</i>	2
5.3.2	<i>Tjenester for digital publisering</i>	2
5.3.3	<i>Tjenester for produksjon av digitalt materiale</i>	2
6	VEIEN FREMOVER.....	2

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato		27.12.2005
	Side		3 av 15
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

6.1	TYPOLISERING.....	2
6.2	IMPLEMENTERINGSSTRATEGI.....	2
6.3	FORVALTNING.....	2
6.3.1	<i>Norsk digitalt biblioteks driftsorganisasjon:</i>	2
6.3.2	<i>Kvalifisert utgiver:</i>	2
7	VEDLEGG.....	2
7.1	VEDLEGG 1 - DEFINISJONER OG FORKORTELSER.....	2
7.2	VEDLEGG 2 - REFERANSER.....	2

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato	27.12.2005	
	Side	4 av 15	
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

1 Oppsummering

Arbeidsgruppen for arbeidspakke 2, *Produksjon av og tilgang til dokumenter*, har tatt for seg det som antas å være tilgjengelige digitale dokumenter i Norsk Digitalt Bibliotek. Gruppen mener at et minimumskrav for dokumenter som inngår i NDB er at de er katalogiserte og at de er tilgjengelige på en eller annen måte. Ut over dette har gruppen definert det som kalles et NDB-dokument - et dokument som i henhold til kriterier for dokumentformat og metadata ligger plassert langs det som i virkeligheten er flere forskjellige skalaer fra minimumskravet som er nevnt ovenfor og opp til det som kalles et fullt NDB-dokument. For metadata er dette listet under punkt 4.1, for dataformater er dette listet i en tabell i punkt 5.1.1.

Dokumenter som skal gjøres tilgjengelig gjennom NDB kan ligge lokalt hos innholdsprodusenten eller, hvis man i noen tilfeller finner det hensiktsmessig, på en sentral NDB-database. Definisjonen av et NDB-dokument er gyldig i begge tilfeller.

2 Innledning

2.1 Hensikt

Dette dokumentet representerer NDB-Rammeverk sin anbefaling til retningslinjer og standarder for produksjon av og tilgang til dokumenter i Norsk digitalt bibliotek.

Dokumentet er utarbeidet av arbeidsgruppen for *arbeidspakke 2 – Produksjon av og tilgang til dokumenter* i prosjektet NDB-Rammeverk.

2.2 Arbeidsgruppens sammensetning

Arbeidsgruppen for AP2 har vært satt sammen av personer med relevant erfaring og praksis fra ulike miljø:

- Espen Ore, Nasjonalbiblioteket
Leder for arbeidsgruppen
- Christian Calmeyer, Aschehoug
- Jan Erik Kofoed, Bibsys
- Jens Vindvad, ABM-Utvikling
- Astrid Jenssen, USIT
- Sven Strøm, UbiT

I tillegg har prosjektleder, Petter Rønningsen, deltatt på alle arbeidsgruppemøtene og har fungert som sekretær for arbeidsgruppen.

2.3 Mandat


Arbeidsgruppen har arbeidet ut fra det mandat som ble presentert ved oppstart av arbeidet:

Arbeidsgruppen skal utrede og utarbeide anbefalinger for standarder og tjenester innenfor følgende områder:

- *Standardformater for digital publisering*
- *Verktøy og tjenester for å forenkle distribusjonsprosessen*
- *Verktøy og tjenester for **produksjon** av digitale dokumenter*

2.4 Målgruppe

Vår primære målgruppe er vår oppdragsgiver, styringsgruppen for prosjektet. Vår sekundære målgruppe er sluttbrukerne som skal forholde seg til det vi kommer frem til så langt som våre direkte oppdragsgivere vedtar våre forslag. Sluttbrukere kan være dokumentprodusenter og/eller publiserende institusjoner, biblioteker eller lignende institusjoner som oppbevarer og tilgjengeliggjør dokumenter og det kan være sluttbrukere som søker etter og bruker dokumenter.

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato	27.12.2005	
	Side	5 av 15	
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

3 Bakgrunn for anbefalingene

3.1 Omfang

Arbeidsgruppen for AP2 har arbeidet med utgangspunkt i det mandat som ble framlagt ved oppstarten av arbeidet.

Med utgangspunkt i dette har arbeidsgruppen sett det som sin oppgave å spesifisere hvilke krav som bør stilles til dokumenter som skal gjøres tilgjengelig gjennom Norsk digitalt bibliotek (NDB). Disse kravene danner basis for den anbefaling til spesifikasjoner for NDB-dokumenter som finnes i rapporten. I tillegg omfatter rapporten konkrete anbefalinger til foretrukne formater for ulike typer digitale dokumenter.

3.2 Tilnæringsmåte

Alt arbeidet i arbeidsgruppen har vært organisert slik at gruppens egne medlemmer har bearbeidet de ulike deler av vårt arbeidsfelt, og har skrevet utkast til anbefalinger. Disse anbefalingene har blitt drøftet på det påfølgende møte, innspill har blitt framlagt og man har arbeidet videre med stoffet fram mot neste møte.

I tillegg har arbeidsgruppen fått bidrag fra Nasjonalbiblioteket innenfor området *Langtidslagring av digital lyd og bilder*.

I grove trekk har arbeidsgangen i arbeidsgruppen vært slik:

- Etablere en felles forståelse av hvilke forventninger som eksisterer til NDB
- Etablere en felles forståelse av hvilke typer dokumenter NDB vil måtte forholde seg til
- Utarbeide en definisjon av hva vi oppfatter som "fullverdige" dokumenter i forhold til de standarder og retningslinjer vi utarbeider – NDB-dokument.
- Utarbeide konkrete anbefalinger til foretrukne formater for digitale dokumenter
- Utarbeide konkrete forslag til krav for tjenester og verktøy

3.3 Grunnleggende forutsetninger

De anbefalinger som er gjort i dette dokumentet baserer seg på følgende grunnleggende forutsetninger:

1. Tjenester, formater, protokoller og retningslinjer skal utarbeides i hht. prinsipper som utarbeides i AP7 – *Systemarkitektur og infrastruktur*.
2. Det forutsettes at metadata for NDB-dokumenter også vil inneholde informasjon om filformater etc. som omtalt i punktene 4, 5 og 6.
3. Det forutsettes at det etableres en varig driftsorganisasjon rundt NDB i en eller annen form. En slik organisasjon vil ha ansvar for vedlikehold av standarder ved siden av praktisk drift.
4. Anbefalinger som gis for NDB-dokumenter må samkjøres med de anbefalinger man kommer frem til ved NB, bl.a. for videreføring av Paradigma-prosjektet. Denne samkjøringen forutsettes å kunne virke begge veier.


3.4 Forhold til andre arbeidspakker

Arbeidsgruppen forutsetter resultater fra enkelte andre arbeidsgrupper, og denne gruppens resultater vil igjen være forutsetning for andre grupper. AP2s resultater vil nødvendigvis måtte brukes av AP7, og AP2 forutsetter at det med basis i AP4 og AP5 lages systemer for autentisering/autorisering samt klarering og betaling.

Arbeidet i AP1, Metadata, er spesielt nært knyttet til arbeidet med AP2: enkelte av de aktuelle formatene for digitale dokumenter har metadata direkte i dokumentene (for eksempel TeiHeader for TEI-dokumenter). I noen tilfeller vil også selvstendige metadata lagres som datafiler i XML, for eksempel arkivdata som følger Library of Congress' EAD-standard. Her er det naturlig at AP1 og AP2 harmoniseres når det gjelder krav til formater. Det er også overlappende vurderinger når det gjelder protokoller, selv om protokoller for søk i og overføring av metadata kan være forskjellige fra de som brukes til søk i (inkl. fritekstsøk) og overføring av dokumenter.

4 NDB-Dokument

Et norsk digitalt bibliotek kan tenkes å omfatte alle digitale dokumenter som er tilgjengelige via deltagende institusjoner (hvis det er slik vi ser for oss NDB) uansett. Men i arbeidet i arbeidspakke 2 har vi så langt regnet med at det vil bli stilt visse krav for at dokumenter skal kunne inngå i NDB. Slike krav kan tenkes å være definert på flere nivåer.

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato	27.12.2005	
	Side	6 av 15	
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

MENOTA (Medieval Nordic Text Archive) bruker en slik flernivå-modell for krav til hva som kan inngå i arkivet: det finnes minimumskrav som **må** være oppfylt for at et dokument skal regnes som et MENOTA-dokument, og så er det flere mulige nivåer over dette. Dokumenter i NDB kan plasseres langs flere akser:

- **Rettinghetskse** fra begrenset tilgjengelighet til fritt tilgjengelig.
- **Datatypeakse** fra lineære (e.g. statiske tekstdokumenter) via multimedia, multimedia med tidsbestemte data (lyd, film) til dynamiske som databaser, spill og applikasjoner.
- **Formatakse** fra proprietært format som krever egne visningsprogrammer til generelle åpne formater med all nødvendig informasjon tilgjengelig (for eksempel xml-dokumenter der DTD, XSL-stilark og nødvendige fonter er tilgjengelige)

4.1 NDB-metadata

Arbeidsgruppen anbefaler at det knyttes metadata til dokumentene for å kunne angi format mer. Forslagene er ment som innspill til det som spesifiseres i "AP1 - Metadata". Forslagene i de to arbeidsgruppene bør samkjøres til ett felles metadata sett.

Vi anbefaler følgende metadata:

- Unik identifikator.
- Angivelse av dokumentformat/medietype (video, pdf etc) i henhold til MIME-spesifikasjonen¹
- Angivelse av hvilket nivå i tabellen med oversikt over klasser av NDB-dokumenter som det aktuelle dokumentet tilhører.
- Angivelse av tilgangsrettigheter

4.2 NDB-formater

Det finnes et bortimot uendelig antall åpne og proprietære dataformater og kombinasjoner av dem. Dokumenter som ligger i proprietære ikke-dokumenterte formater er vanskelige å vedlikeholde over tid, og det kan være vanskelig å vise dem eller spille dem av hvis det kreves egen programvare får å få til dette. I en ideell verden ville alle tekstdokumenter vært i XML med en vedlagt DTD eller et schema og med et stilark i XSL eller som en kombinasjon av XSL og CSS. Men vi kan ikke stille slike krav. I stedet har vi i dette notatet laget en inndeling i typer dokumenter basert på hvilke formater data er lagret i. Det er ikke gitt at et dokument bare finnes i ett format, og det er ikke gitt at en bruker vil kunne ha nytte av alle formater - kanskje rett og slett fordi han eller hun ikke har utstyr/programvare som kan håndtere alle mulige formater. For å hjelpe bruker og leverandører har vi derfor, i punkt 5.1.1, listet det vi kaller NDB-dokumenter langs en akse fra ikke-akseptabelt som NDB-dokument til fullt NDB-dokument.

Listen over godkjente formater i NDB (se nivå 2 i tabellen i punkt 5) vil forandres over tid, men dette bør ikke skje like ofte som proprietære dataformater endres (e.g. Word doc). For at et format skal være akseptert i NDB, må det være dokumentert slik at det alltid vil være mulig å lage en applikasjon som kan lese dokumentet. Slike formater er i dag:

4.2.1 Tekst eller tekstdelen av et dokument


Her anbefaler vi ikke-proprietære formater der både data og formateringsinformasjon lagres i det vi (i mangel av noe bedre) kan kalle universelt lesbar tekst.

Følgende tekstformater anbefales:

- Ren tekst i et ISO 646-basert tegnsatt (nasjonal standardiseringsversjon må være kjent), i ISO 8859-x eller UNICODE.
- SGML med publisert DTD
- XML med publisert DTD/Schema og eventuelt stilark
- XML uten publisert DTD/Schema men med stilark
- XML uten publisert DTD/schema og uten stilark
- XHTML

Et eget problem oppstår med dokumenter som bare finnes som HTML-filer med spesialtilpasning for en eller annen nettleser slik at filene ikke lenger er validiserbar HTML. Dette gjelder som oftest sider tilpasset Microsofts Internet Explorer, men problemet er også generelt: det er kanskje færre validerte enn ikke-validerte HTML-sider tilgjengelige på servere rundt i verden.

¹ MIME-spesifikasjonen finnes på <http://www.rfc-editor.org/rfc/rfc2046.txt>

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato	27.12.2005	
	Side	7 av 15	
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

Adobe® Portable Document Format (PDF) er et proprietært format som brukes mye. Ifølge Adobes hjemmesider skal PDF anses som publisert og offentlig tilgjengelig², og PDF er også inne i en standardiseringsprosedyre i ISO TC 130, samtidig som det er et prosjekt som er i ferd med å spesifisere en variant av PDF som heter PDF/A, A for archive: **ISO 19005**³.

Spesielt innen universitetsmiljøer har TeX og påbygg til TeX vært mye brukt i dokumentproduksjon. TeX har i hvert fall ikke opprinnelig vært ment å skulle være et lagringsformat, men et sett av koder for produksjon av trykte dokumenter, så det er vanskelig å se at det passer inn som et anbefalt format.

Microsofts programmer bruker et eget tekstbasert utvekslingsformat, rtf - Rich Text File. Dette formatet er ikke offentlig dokumentert og det ser også ut til å variere over tid når Microsoft lanserer nye programversjoner.

4.2.2 Multimediedata

Følgende bør være retningsgivende for valg av filformater for multimedia:

- Tapsfrie formater
- Unngå proprietære formater
- Ved bruk av ”datareduerte” formater som MP3, bør en velge så åpne lisenser som mulig

4.2.2.1 Bilder

Bilder er i utgangspunktet linjefrafikk (vektorgrafikk) eller rasterbilder. Linjefrafikk kan være en graf tegnet ved en matematisk funksjon - i så fall er funksjonen en kopi av grafen eller omvendt - eller et "manuelt" tegnet bilde. For linjefrafikk generelt finnes nå standarden VSL, som gjør det mulig å lagre linjefrafikk som XML-data, og dette ville være å foretrekke i NDB. Men mye linjefrafikk er utviklet i proprietære formater som på grunn av sitt gjennomslag i markedet er de facto standarder som for eksempel Illustrator-filer.

For rasterbilder er det ideelle for NDB data lagret i en komprimering som ikke fører til informasjonstap. En måte å gjøre det på i dag er med filer i TIFF som kan være komprimert med LZW-algoritmen uten informasjonstap. Man må likevel regne med at mange bildefiler i dag produseres komprimert etter en informasjonsforringende algoritme, JPEG. I NDB-sammenheng er det ønskelig at slike filer finnes i høyest mulig kvalitet.

For 8-bits rastergrafikk har GIF vært bortimot enerådende på Internett. GIF bygger på et patent som er eller blir friggitt, og det har derfor vært arbeidet med alternativer. PNG dekker både 8-bits og 24/48-bit bildefiler, men dette formatet er foreløpig ikke godt implementert i de vanligste nettleserne.

Anbefalte formater for bilder:

Vektorgrafikk:

- VSL
- SVG

Rastergrafikk:

- TIFF med tapsfri komprimering
- JPG
- PNG
- GIF
- Bitmap-filer?
- EPS-filer?


4.2.2.2 Video

Når det gjelder video er det profesjonelle formatet Digital Video, en standard utarbeidet av de store produsentene av videoutstyr. Denne er en variant av MJPEG (Motion JPEG), videoen er altså en rekke av JPEG-bilder etter hverandre. Typisk bygges codec for denne typen video som en hardware-codec direkte i kameraet. Det skal sies at DV er delt i en rekke leverandørspesifikke varianter, men disse skal forholde seg til det standardiserte formatet med hensyn til bitrate, oppløsning og bilder per sekund.

I vår sammenheng må vi forholde oss til andre formater som MPEG og DivX i tillegg til diverse proprietære formater med tilhørende codec'er. Det er særlig det siste som ofte skaper problemer, i den grad en codec er software-basert blir den også sårbar for småfeil i programmeringen og fortolkninger av standarden fra programmerers side.

² Finnes på <http://partners.adobe.com/public/developer/pdf/topic.html>

³ Et siste draft finnes som vedlegg til denne rapporten.

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato	27.12.2005	
	Side	8 av 15	
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

DV er et dominerende format amatør- og halvproff-sammenheng, mens profesjonelt utstyr gjerne har valg mellom DV og MPEG, da med flere variabler i forhold til fargekoding og bitrate.

MPEG er imidlertid klart dominerende i leveringsammenheng, som i DVD og Digital Kringkasting DVB. Her bør man være oppmerksom på at DV og MPEG raskt degraderer kvaliteten, noe som er viktig i lagringssammenheng.

I denne sammenheng bør vi anbefale leverandørene å velge et av to alternativer:

- å velge åpne formater eller
- dersom de leverer i et lukket format, å legge ved den versjonen av codec'en som er bruket.

Dette siste er bare en utsettelse, ved neste større oppgradering av et operativsystem er det sannsynlig at codec'en ikke lenger virker. Det samme problemet vil vi ha med lydformater.

Anbefalte formater for video:

- DV
- MPEG
- Andre formater med åpne algoritmer for komprimering

4.2.2.3 Lyd

Det er viktig å være klar over at en lydfil består av to hoveddeler, et hode som bl.a. spesifiserer hvordan dataene er kodet (PCM, MPEG etc) og selve dataene. Dette betyr at selv om et filformat går ut av bruk, er ikke dataene nødvendigvis tapt. Dersom disse er kodet i et dokumentert format, vil det være mulig å gjenopprette data fra filen. Av denne grunn er det viktig at en velger rett *dataformat*, snarere enn å konsentrere seg om "innpakning".

På dette grunnlaget kan en gå ut fra at de mest robuste lydfilene er filer kodet med PCM siden dette er grunnformatet for digital lyd.

Av samme grunn kan en gå ut fra at de minst robuste formatene er filer der dataene er kodet etter en proprietær algoritme. Disse vil dessuten ofte være "loosy", dvs. at filene komprimeres til mindre størrelse ved å fjerne data.

I praksis må NDB akseptere avlevering også i proprietære formater som RealAudio og QuickTime, spørsmålet blir snarere om hvordan disse skal innpasses i "medlemskapshierarkiet" og hva en kan praktisk kreve av leverandørene i form av avlevering av avspillere etc.

Anbefalte formater for lyd:

- Dokumentert PCM-format⁴ (Eks. BroadcastWAV)
- OGG-Vorbis
- MP3

5 Produksjon av og tilgang til dokumenter

5.1 Standardformat for digital publisering

Dimensjonene som er listet under punkt 4, er her forsøkt integrert til en akse, fra ikke-NDB-dokumenter til komplette NDB-dokumenter.


For at et dokument skal inngå som et NDB-dokument, kreves det at det finnes en katalogpost og at dokumenter presenterbart (minimale NDB-dokument). Herfra går det oppover til et fullt NDB-dokument der dokumentet er i et format som følger en standard (en liste over slike standarder må skrives som vedlegg til dette) opp til et "fullt NDB-dokument" der det kreves kildefil + minst ett presentasjonsformat (man kan tenke seg XML + PDF med inkluderte fonter, eller XML + XSL + fonter). For ikke-tekstlige data gjelder det her at det ikke skal være noe informasjonstap ved komprimering. Det gjenstår altså å liste hvilke formater som gir hvilket nivå som NDB-dokument for tekst, lyd, bilde, video (+ fler?).

5.1.1 Dokumenttyper

I hovedsak har vi sett for oss tre grupper dokumenter:

- Dokumenter som etterlever vår anbefalinger med flere nivåer av struktur og kompatibilitet

⁴ Wav definerer PCM-informasjonen Broadcast WAV, BWF er et wrapperformat, som ikke nødvendigvis inneholder wav-info i det hele tatt, Det kan like gjerne være mpeg audio

 NB	NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter			Dato	27.12.2005
				Side	9 av 15
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon	A	

- Dokumenter som ikke gjør det, men som blir lagret innen rammen av NB eller hvor det finnes formaliserte avtaler om tilgang.
- Dokumenter som er ”der ute” og helt utenfor vår kontroll

Dette kan illustreres slik⁵:

Nivå	NDB-dok	Beskrivelse	Aksepterte filformater			
			Tekst	Bilder	Video	Lyd
10	Ja	Som 2 + Kildefil(er) må være tilgjengelige og minst ett presentasjonsformat. Ikke informasjonstap i komprimering. Dokumentet er direkte tilgjengelig via en URI	Kildefil er ren tekst eventuelt kodet i XML og med stilark i CSS og/eller XSL	Raster: TIFF, GIF Vektorgrafikk: SVG		
9	Ja	Dokumentet kan foreligge i et proprietært format	PDF, MS-Word			
8	Ja	Som 10, men informasjonstapende komprimering aksepteres		JPEG		
7	Ja					
6	Ja					
5	Ja	Dokumentet er ikke direkte tilgjengelig via fast URI (man kan ikke lagre en peker til det), men det må hentes frem via eget grensesnitt				
4	Ja					
3	Ja					
2	Ja	Som 0 + dokumentet finnes i ett eller flere aksepterte formater (se NDBs liste over slike)				
1	Ja					
0	Ja	Katalogpost finnes, dokumentet skal være presenterbart	Hva som helst	Hva som helst	Hva som helst	Hva som helst
	Nei	Oppfyller ikke kravet over				

5.1.2 Tegnssett


NDB vil måtte forholde seg til dokumenter som er en del av en ferdig (oftest kommersiell) pakke, inkl. tegnssett- og kodingsavgjørelser på den ene siden, på den andre siden er det dokumenter som produseres av deltagende biblioteker eller andre organisasjoner der NDB bare kan sett opp ønskede spesifikasjoner for standarder.

Noen dokumenter kan være distribuert i proprietære formater der tegnkoder og fonter også er integrert i dokumentene (e.g. PDF), mens andre kan komme som for eksempel XML eller HTML der det er opp til brukerens visningsverktøy/nettleser + maskinkonfigurasjon å presentere teksten (og tegn/fonter).

Dokumenter i XML (og stort sett i HTML) vil/skal ha høyde for Unicode, men noen kan ha definert sitt innhold som en delmengde eller eventuelt ha kodet det på en annen måte enn UTF8 så som ISO 8859.

Noen dokumenter (kanskje spesielt fra forskningspublikasjoner) vil bruke Unicode til tegnkoding, der det også inkluderes tegnkoder fra PUA (Private User Area). I mange (men ikke alle) tilfeller vil disse tegnkodene brukes sammen med egne fonter. Blant (standard-)organisasjoner som forholder seg til bruk PUA og dokumentasjon/deklarasjon av denne bruken finnes TEI (Text Encoding Initiative) der en nye versjon (TEI P5) skal være klar i løpet av høsten og der kapitlet om "writing systems" blir skrevet om i forhold til tidligere utgaver, og MUFTI (Medieval Unicode Font Initiative). Noen dokumenter kan bruke tegnkoder, for eksempel fra standard Unicode, men kreve spesielle fonter for presentasjon. Dersom disse fontene er proprietære, kan det skape problemer for brukere som ikke har disse fontene. For NDB bør dette tas med i forbindelse med rettigheter og bruk. For dokumenter som krever egne fonter, enten dette skyldes tegn fra PUA eller fra standard Unicode-tegnkoder eller for den saks skyld ISO 8859-x, må dette være dokumentert i en eller annen form for metadata. Det må også dokumenteres hvordan brukerne kan få tilgang til spesialfonter hvis disse ikke er inkludert direkte i dokumentene.

⁵ Dette er en anbefaling, som må bearbeides videre før den kan taes i bruk. Bearbeidingen må bl.a. omfatte spesifisering av minimumskrav til metadata.

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato	27.12.2005
	Side	10 av 15
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen
		Revisjon A

Anbefalte tegnsett:

- Unicode
- ISO 8859-x
- ISO 646: ihvertfall i US og norsk implementasjon

5.1.3 Protokoller for dataoverføring

Publisering, lokalisering og aksess til elektroniske dokumenter vil for det aller meste skje over nettverk. For at dette skal kunne skje på en mest mulig enkel og grei måte er vi avhengig av å bruke standardiserte protokoller.

De protokollene som er aktuelle må være:

- Åpne (ikke proprietære)
- Standardiserte
- Enkle å implementere
- Effektive
- Godt dokumenterte

5.1.3.1 Protokoll-typer

Følgende protokoller anbefales brukt:

- **Transport-protokoller**
 - HTTP
 - FTP
 - SMTP
 - Framtidige protokoller basert på BEEP-rammeverket
- **Tjenesteorienterte protokoller** basert på web services
Disse bygger på underliggende protokoller som kan innbefatte:
 - SOAP
 - WSDL
 - UDDI
- **Protokoller for streaming**
 - RTSP/RTP – Realtime Streaming Protocol / Realtime Transport Protocol
 - SDP Session Description Protocol
 - MPEG-protokoller (kodek og innhold)

5.2 Verktøy

Gode verktøy er viktig ved elektronisk publisering. Det gjelder både for forfattere, de som skal publisere og de som skal søke fram og få tak i dokumentene.

Hvilke konkrete verktøy som er aktuelle vil variere over tid og kan forandre seg raskt. Det har derfor liten hensikt å anbefale konkrete produkter. Men det går an å si en del om hvilke *typer verktøy* som er aktuelle. Det er naturlige å dele verktøyene inn i grupper og spesifisere funksjonene de har som komponenter.

5.2.1 Tekstbehandling.

Verktøy for tekstbehandling vil omfatte tre hovedtyper:


- Teksteditor
- Innholdseditor
- Tekstgjenkjenning (OCR)

5.2.1.1 Teksteditor.

Program for å redigere enkle tekster uten spesiell struktur.

Egenskaper:

- Basis editor-funksjoner som klipp og lim; søk og erstatt etc.
- Støtte for UNICODE i ulik serialisering (UTF-8, UTF-16 mfl.)
- Rense fil for kontroll-tegn og uønsket formatterings-informasjon.

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato		27.12.2005
	Side		11 av 15
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

- Stavekontroll
- Syntaksorientert visning (for filer som inneholder html, xml o.l.).

5.2.1.2 Innholdseditor.

Program for å redigere strukturert tekst.

Egenskaper:

- Alle egenskapene til enkel tekstredigerer pluss:
- Støtte for XML
- Validering i følge DTD eller skjema (W3C og Relax). Mulighet for valgfri parser (plugin).
- Ulike visningsmodi: Tekst, merket, formattert (med stilark)
- Støtte for ulike standard dokumenttyper: TEI, DocBook, Open Ebook mfl.
- Prosessering med stilark: CSS, XSL, FOP (til PDF, PostScript o.a.)

5.2.1.3 Tekstgjenkjenning (OCR)-verktøy.

Program for gjenkjenning og ekstraksjon av tekst fra bilder eller ved skanning av dokumenter.

Egenskaper:

- Støtte kildedokument i ulike bildeformater: TIFF, JPEG, PNG mfl.
- Støtte for ulike skannere (TWAIN-grensesnitt).
- Støtte ulike skriftsnitt (latin, fraktur, gresk, kyrillisk mfl.)
- "Læremodus" (mulighet for å fortelle programmet hvilke former som tilsvarer bestemte tegn).

5.2.1.4 Andre komponenter

Komponenter som kan finnes som enkeltstående verktøy eller inngå som moduler (plugins) i andre verktøy:

- Tegnkonvertering
- Formatkonvertering
- XML-parsing og validering
- Stavekontroll
- Prosessering av stilark

5.2.2 Bildebehandling

Verktøy for bildebehandling bør ha følgende egenskaper:

- Generelle bildebehandlingsegenskaper (endre størrelse, beskjære, zoom, justere farger, rotere etc)
- Støtte konvertering mellom ulike formater
- Støtte både vektor- og rastergrafikk og konvertering mellom disse
- Støtte for ulike fargerom
- Fange bilder fra skjerm (capture)
- Kopling til skanner (TWAIN)
- Validering

5.2.2.1 Andre komponenter


Komponenter som kan finnes som enkeltstående verktøy eller inngå som moduler (plugins) i andre verktøy:

- Formatkonvertering

5.2.3 Lydredigering

Verktøy for lydredigering bør ha følgende egenskaper:

- Innspilling og avspilling
- Oppfangig av lyd fra PC (audio capture)
- Støtte for ulike formater og codecs
- Gode konverteringsmuligheter
- SMPTE synkronisering (tidskode mot video)
- Validering

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato	27.12.2005	
	Side	12 av 15	
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

5.2.3.1 Andre komponenter

Komponenter som kan finnes som enkeltstående verktøy eller inngå som moduler (plugins) i andre verktøy:

- Formatkonvertering

5.2.4 Videoredigering

Verktøy for videoredigering bør ha følgende egenskaper:

- Innspilling og avspilling
- Oppfangning av video fra PC (video capture)
- Støtte for ulike formater og codecs
- Gode konverteringsmuligheter
- Validering

5.2.4.1 Andre komponenter

Komponenter som kan finnes som enkeltstående verktøy eller inngå som moduler (plugins) i andre verktøy:

- Formatkonvertering

5.2.5 Metadata-registrering (katalogiseringsprogram)

Verktøy for metadata-registrering bør ha følgende egenskaper:

- Støtte ulike metadata-formater: MARC, MODS, Dublin Core (DC) mfl.
- Støtte ulike fysiske formater: ISO-2709, MARCXML, DC i XML, RDF, TEI Header, DocBook Info-tags, PDF-metadata mfl.
- Eksport og import av metadata
- Mulighet for innhøstning via OAI-PMH
- Støtte for referansehåndteringsverktøy: Reference Manager og Procite (RIS) samt BibTex.
- Ulike visningsformater: Fortekstformat, ISBD, CIP mfl.

5.2.5.1 Andre komponenter

Komponenter som kan finnes som enkeltstående verktøy eller inngå som moduler (plugins) i andre verktøy:

- Metadata-generator
- Metadata-konverterer

5.3 Tjenester

Alle tjenestene som nevnes her, må ikke nødvendigvis implementeres samtidig fra starten av NDB. Det er heller ikke gitt at en tjeneste bare skal leveres av én institusjon i NDB. Flere deltagende institusjoner - for eksempel lokale bibliotek - bør kunne fungere som tjenesteytere i NDB. Det er videre verd å være oppmerksom på at tjenestene som er nevnt inngår i systemet *som helhet*, uavhengig av hvor distribuert dette er.


Det kan også diskutere hvilket system for mapping mellom unik identifikator (URN, DOI o.a.) og fysisk plassering som velges. For enkelhets skyld er en slik tjeneste, som vi forøvrig mener er nødvendig for å få et robust system, bare omtalt som *URN-tjenester*.

5.3.1 Tjenester for gjenfinning og framvisning

- URN-oppløsning
- Katalogtjeneste og søk
- Levering av dokumenter og andre filer
- Rettighetshåndtering - både i betydningen at klienten får tilgang til relevant rettighetsinformasjon og at systemet gir klienten tilgang til de ressursene denne har rett til å se
- Autentisering og single sign on

5.3.2 Tjenester for digital publisering

- Deponering og oppbevaring av dokumenter og andre filer (mulighet for batch-import må være tilstede)
- Katalogisering og indeksering
- Metadatagenerering
- URN-katalogtjeneste

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato	27.12.2005	
	Side	13 av 15	
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

- Formatkonvertering (eks. konvertering til PDF)

5.3.3 Tjenester for produksjon av digitalt materiale

- Validering av xml; kvalitetskontroll av andre filer (for eksempel PDF og bildeformater)
- Metadatagenerering
- URN-katalogtjeneste

6 Veien fremover

6.1 Typologisering

Under hovedpunkt 4 og punkt 5.1 med underpunkter beskrives NDBs anbefalte standarder og protokoller. Som en del av metadata som lages innen rammen av NDB for dokumenter som inngår i NDB, enten de er produsert/lagret i en NDB-institusjon eller ikke, bør det inngå informasjon om hvor dokumentene passer inn i henhold til disse anbefalingene. Det bør også registreres informasjon om hvor i aksene statiske - dynamiske data befinner seg. Denne typologiseringen bør utføres når dokumentene registreres og/eller når metadata produseres.

Det kan være aktuelt å bruke enkle verifiseringsverktøy: for eksempel programmer som kan sjekke filtyper. Dersom dokumenter er kodet i XML og viser til en DTD eller et XML-schema, kan dokumentene også valideres. Uansett kan man sjekke velformethet i XML-dokumenter. Eventuelle nye former for datakodning kan man regne med at også vil tilby slike funksjoner.

6.2 Implementeringsstrategi

Siden gjennomføring av NDB er et svært omfattende prosjekt, er det naturlig å tenke seg en faseinndelt implementering av retningslinjer og tjenester. Nedenfor trekker vi opp et slikt løp innen AP2s arbeidsområde.

Fase	Omfang
Fase 1	Publisering av retningslinjer for produksjon av og tilgang til digitale dokumenter. Herunder anbefalinger for implementering av lokale deponi. Det forutsettes at det på dette tidspunktet er enighet blant aktuelle leverandører om prinsipper og formater.
Fase 2 (umiddelbart etter fase 1)	Oppbygging av katalog- og indekseringstjeneste / -løsning Oppbygging av database over kvalifiserte (innmeldte) leverandører Tjeneste for validering av XML, eventuelt for andre filformater Tjeneste for metadatagenerering Aktuelle leverandører gjør sitt innhold og/eller metadata tilgjengelig for indeksering og høsting. Tjeneste for konvertering til PDF
Fase 3	Implementering av sentralt deponi Implementering av løsning for autentisering og single sign on

To brikker som er avgjørende for prosjektet suksess er gjennomføring av et system for håndtering av rettigheter til materialet og, ikke minst, en URI-katalog / -resolvertjeneste. Disse bør implementeres så tidlig i prosjektet som mulig, spesielt er URI-tjenesten viktig.


For å få et så robust og standardisert system som mulig, bør en, på alle nivåer, følge gjeldende standarder. Per i dag er det ingen slike for disse to områdene og det er derfor mulig at en må implementere ad hoc-løsninger. I så fall er det viktig at disse konstrueres med blikk for den utviklingen en ser innenfor feltene. Når det gjelder håndtering av rettigheter er arbeidet i forlengelsen av rapporten om enklere tilgang til digitale ressurser viktig.

6.3 Forvaltning

Vi forutsetter her at det opprettes en driftsorganisasjon for Norsk digitalt bibliotek. Driften bør sannsynligvis organiseres likt for hele Norsk digitalt bibliotek.

Ut fra tjenestene som er tenkt tilbudt kan driften organiseres med to typer aktører: Norsk digitalt biblioteks driftsorganisasjon og deltagende institusjoner som er "kvalifiserte utgivere".

Det anbefales at oppgavene fordeles mellom følgende to aktører:

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato	27.12.2005	
	Side	14 av 15	
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

- Norsk digitalt biblioteks driftsorganisasjon
- Kvalifisert utgiver

6.3.1 Norsk digitalt biblioteks driftsorganisasjon:

Det miljø som får rollen som NDBs driftsorganisasjon bør ivareta følgende oppgaver i forhold til forvaltning av løsning og standarder/retningslinjer:

- vedlikeholder kriteriene for at en deltagende institusjon kan godkjennes som kvalifisert utgiver
- fører liste over kvalifiserte utgivere
- godkjenner og anbefaler nye formater
- gir anbefaling om konvertering av avlegse formater
- gir anbefaling om bruk av nye protokoller
- definerer protokoll som avleggs for bruk innen NDB
- delfinansierer videreutvikling og drift av identifikatortjeneste (f.eks. URN-tjeneste)
- fører liste over anbefalte valideringstjenester og valideringsmetoder
- fører liste over godkjente deponitjenester
- vurderer om dokumenter med mindre enn fullt medlemskap i NDB skal katalogiseres og indekseres, og evt. kjøper inn slik katalogisering og indeksering

6.3.2 Kvalifisert utgiver:


Hver enkelt av de kvalifiserte utgivere forplikter seg til å:

- holde juridiske rettigheter på dokumentene oppdaterte og synlige
- gjøre dokumentene teknisk tilgjengelige med akseptabel oppetid
- gjøre metadata tilgjengelige
- dele kunnskap om metoder og drift, angående digital publisering i NDB, med andre kvalifiserte utgivere
- autentisere brukere mot godkjent autentiseringssystem (f.eks. FEIDE)

7 Vedlegg

7.1 Vedlegg 1 - Definisjoner og forkortelser

Begrep	Beskrivelse
ASCII	American Standard Code for Information Interchange. US-implementasjon av ISO 646
CSS	Cascading Style Sheet. Stilarksystem for layout, typografi etc. kan brukes isetdetfor eller sammen med XSL
DTD	Document Type Definition. I SGML og XML definerer dette dokumentstrukturen/dokumenthierarkiet og fungerer da også som en kodebok i og med at alle elementer og attributter er definert her.
FEIDE	Felles elektronisk identitet i utdanningssektoren
GIF	Comuserve's Graphics Interchange Format. Dette var lenge et dominerende 8-bits bildeformat på WWW. På grunn av krav om lisensiering for bruk i henhold til patentrettigheter, ble det lett etter erstatningsformater, men GIF er nå på vei over i det fri uansett.
ISO 10646	Flerbytes tegnssett som ideelt sett skal være universelt. (Se også UNICODE)
ISO 646	7-bits tegnssett som så uforandret eller med modifikasjoner ble antatt av nasjonale standardinstitusjoner. Dette er basistegnssettet for SGML.
ISO 8859	8-bits tegnssett som finnes i flere varianter for forskjellige språk. ISO 8859-1 dekker blant annet engelsk og norsk
ISO 8879	SGML
JPEG	Komprimert billedformat med Joint Photographers' Expert Groups komprimeringsalgoritmer.
MENOTA	Medieval Nordic Text Archive - en organisasjon med deltagere fra Danmark, Island, Norge og Sverige som arbeider bl.a. med kodestandard for koding av middelaldermanuskripter, først og fremst på norrønt. Hjemmeside: www.menota.org .
MPEG	Komprimerte video- eller lydfiler i følge algoritmer fra Moving Picture Experts Group.
MIME	Multipurpose Internet Mail Extensions
MUFTI	Medieval Unicode Font Initiative. Mer info på MENOTAs sider.
PDF	Adobe Portable Document Format
<Protokoller>	
QuickTime	Apple's standarder for komprimering og avspilling av lyd- og videofiler

 NB NDB-Rammeverk, Rapport for AP2 – Produksjon av og tilgang til dokumenter	Dato	27.12.2005	
	Side	15 av 15	
Ansvarlig: Petter Rønningsen	Gransket av:	Godkjennes av: Styringsgruppen	Revisjon A

Begrep	Beskrivelse
RTF	Rich Text Format. Et leselig utvekslingsformat for formatert tekst. Et Microsoft-produkt - det finnes ingen dekkende offentlig publisering av formatet.
SGML	Standard Generalized Markup Language, ISO 8879 . XML er en forenkling og videreføring av SGML
SVG	Scalable Vector Graphics. W3C er ansvarlig for standarden
TEI	Text Encoding Initiative. Lanserte i 1994 TEI P3, Recommendations for text encoding and data interchange. Denne utgaven var basert på SGML. En XML-versjon er publisert i 2002, og en større revisjon er nå i gang med P5 som ventes klart tidlig i 2005. Hjemmeside: www.tei-c.org
TIFF	Tagged Image File Format
UNICODE	Universelt flerbytes tegnssett som harmoniseres fortløpende med ISO 10646 (se dette). The UNICODE Consortium er ansvarlig for standarden.
XML	eXtensible Markup Language - på mange måter en forenkling av SGML, men det inneholder også noen viktige endringer, spesielt nyanseringen mellom "wellformed" og "valid". XML beveger seg nå vekk fra DTDer for å beskrive dokumentstruktur til Schemas. W3C er ansvarlig for standarden.
XSL	eXtensible Stylesheet Language. W3C er ansvarlig for standarden.
XML-skjema	XML-skjema er et XML-basert alternative til DTD og beskriver strukturen av et SML-dokument.

7.2 Vedlegg 2 - Referanser

Referanse	Kommentar
Prosjektmandat	Mandat for NDB-Rammeverk, vedtatt av styringsgruppen 09.12.03
Paradigmanotater [1] <Albertsen, Ketil.> Lasterampe: Dataformat i dokumenter. <Nasjonalbiblioteket, paradigma. 2004>. http://norskdigitalbibliotek.no/NDBRamme/AP2/LasterampeDataformater.doc [2] Stenstad, Asborg. Produkt 1.2.1. Typologi over det digitale dokumentuniverset. Nasjonalbiblioteket, paradigma. 2002. http://norskdigitalbibliotek.no/NDBRamme/AP2/Produkt_1.2.1_vB1_20021002.doc [3] Albertsen, Ketil. Paradigma RFC 05: En taksonomi for "the deep web". Nasjonalbiblioteket, paradigma. 2003. http://norskdigitalbibliotek.no/NDBRamme/AP2/RFC05.1.doc [4] Albertsen, Ketil. Paradigma RFC 09: Langtidsbevaring og dataformater. Nasjonalbiblioteket, paradigma. 2004. http://norskdigitalbibliotek.no/NDBRamme/AP2/RFC09.0.doc	NB! Referansene peker til interne dokumenter utgitt av Paradigma. Lenkene peker til arbeidsversjoner gitt av Paradigma til AP2. De er lagret på NDB-Rammeverks lukkede område.
Oulie, Helge. Digitalisering av fotosamlinger. ABM-skrift #1. ISBN 82-8105-000-4. ABM-utvikling, <Oslo, 2004>. http://www.abm-utvikling.no/publisert/ABM-skrift/2003/Digitalisering.pdf	