2. Juni 2000

Anita Iren Oppedal

Institutt for datateknikk og informasjonsvitenskap
NTNU

"Alt er metadata"

Bruk av metadata i et integrert brukersystem

Hovedoppgave i informatikk

INFORMASJONSFORVALTNING

FORORD

"Jeg ville takka livet, som har gitt meg så mykket" synger Arja Saijonmaa. Jeg vil også synge ut en stor takk til alle de som har bidratt til fullførelsen av min hovedfagsavhandling. Arbeidet med oppgaven har vært en lang og lærerik prosess, med mange "riv meg i håret" frustrasjoner. Uten alle dere som sammen har gitt "liv" til meg gjennom oppmuntring og støtte gjennom hele denne prosessen, hadde min hovedfagsavhandling vært en "saga blott".

En takk går til alle de som har deltatt i brukerundersøkelsen min ved Adresseavisen, og mine "hjelpere" ved Arkivet og andre nøkkelpersoner der som har bidratt med masse viktig informasjon gjennom løpende kontakt underveis. Deres hjelp har vært uunnværlig.

En takk går også til medstudenter og venner som har oppmuntret meg underveis i arbeidet med oppgaven. En spesiell takk her går til medstudent Nina Michalsen for korrekturlesing og mange fruktbare diskusjoner.

Jeg takker også min veileder Ingeborg Sølvberg for god veiledning og konstruktive og inspirerende tilbakemeldinger.

En siste takk går til Kjersti og Elin for hjelp til korrekturlesing i sluttfasens kritiske minutter. Uten dere.....

SAMMENDRAG

Denne hovedfagsavhandlingen setter fokus på hvordan metadata brukes i et integrert brukersystem i en bedrift.

I et informasjonsrom er informasjonsressurser fra ulike medier intergrert, og en trenger et felles "bindeledd" for å støtte bedre gjenfinning og tilgang til informasjon i informasjonsrommet. Problemet er ofte at de ulike medier bruker ulike format for beskrivelse av sine informasjonsressurser, noe som vanskeliggjør interoperabilitet mellom de ulike medier. Dersom de ulike medier kan bruke samme metadataformat til å beskrive sine informasjonsressurser, vil det bedre interoperabiliteten.

Dublin Core Metadata Element Set (DC) er et format utviklet med tanke på publisering av informasjonsressuser via Intranett og Internett. Det er DC som er bindeleddet i det virtuelle informasjonsrommet som denne avhandlingen tar utgangspunkt i.

Sentralt i denne avhandlingen står vurderingen av hvordan Adresseavisens indekseringsbehov kan tilfredsstilles i DC for informasjonsressurser som artikler, bilder/illustrasjoner og film. Forslag til et kjerneformat for Adresseavisens informasjonsressurser, med medieavhengige variasjoner legges frem. Dette er informasjonsressurser hvor avis er brukskontekst.

Forslaget som fremlegges imøtekommer resultater fra brukerundersøkelsen, og opplysninger og observasjon av hvordan indekseringsformatene allerede benytter brukes.

Undersøkelsen har resultert i følgende funn :

- De fleste brukere velger fritekstsøk fremfor metadatasøk
- Opplæring virker inn på bruk av metadata
- Arbeidsoppgaver/Informasjonsbehov påvirker bruk av metadata
- Erfaring med databasesystemet og hyppighet i søk i databasen kan påvirke bruk av metadata
- Noen metadataelement er mer bedre egnet for søk enn andre.

Undersøkelsene gir også anbefalinger som kan være nyttige ved navngivning av metadata. Følgende fremgår av undersøkelsen:

- Forkortelser i navngivning av metadata bør unngås for å gjøre dem mer selvforklarende
- Tvetydige begreper i navngivning av metadata gjør dem mindre intuitive i forhold til forståelse for innhold

Undersøkelsen er presentert med stolpediagram og tabeller, som er metoder som kan brukes til kvalitative analyser.

Kapittel 4:

Kapittel 5:

Kapittel 6:

Kapittel 7:

Kapittel 8:

Vedlegg A:

Vedlegg B:

Kapittel 1:

Kapittel 2:

Kapittel 3:

Kapittel 4:

Kapittel 6:

Kapittel 7:

Kapittel 8:

Kapittel 1

Introduksjon

"Alt er metadata,
metadata er alt"

1.1. Problembeskrivelse

Ny informasjonsteknologi har gjort det mulig å lagre informasjon elektronisk, for raskere og mer effektiv tilgang til informasjon. Det har ført til at vi alle er blitt brukere og produsenter av informasjon gjennom det konseptet vi kjenner som Internett, WWW eller fornorsket til Verdensveven, hvor begreper som tid og sted nærmest har opphørt å eksistere.

Denne unike kommunikasjonsmuligheten hvor vi alle kan dele informa sjon uavhengig av tid og sted, har ført til at vi blir oversvømt av informasjon fra alle kanter fra ulike media daglig. Hele hensikten med å dele informasjon ut fra et vitenskapelig perspektiv er at den kan gjen-brukes, og ettersom mengden av informasjon øker, ser en viktigheten av å finne rett informasjon til rett tid. Vi lider av en form for "informasjons-bulimi", hvor det er umulig å fordøye all den informasjonen vi blir servert. Vi må derfor være kritisk til informasjonen, hvor vi velger hvilken informasjon vi vil ta til oss, og hva vi må forkaste. Vi lærer oss å skille mellom relevant og ikke relevant informasjon, men til hjelp i denne prossessen trenger vi nye metoder som skal bedre på gjenfinning prossessen og relvansvurderingen.

Et forholdsvis nytt begrep er informasjonsrom eller arbeidsarenaer der du har tilgang til all den informasjon du trenger til støtte i de arbeids-prosesser du skal utføre.

Innenfor fagområdet informasjonsforvaltning og digitale bibliotek omtales dette som corporate memory.

Det er ønskelig at mest mulig informasjon som lagres skal kunne gjenfinnes, og et alternativ til dagens klikk og les metode på Internett er å koble metadata til informasjonsobjektene og bruke metadataene ved utsøking og navigering i informasjonsrommet.

Metadata er et sett med strukturerte elementer som beskriver dokumentets innhold for å unikt identifisere det og blir brukt til å bedre gjenfinning av dokumentet ved at det søkes direkte på de metadataelement du har tilgjengelig. Det benyttes i dag ulike metadataformat alt etter hvilken informasjon som skal representeres, det være seg tekst, bilder, lyd m.m.

Bibliotek bruker standardiserte formater som MARC formatet som utgangspunkt, mens de fleste bedrifter i dag bruker ad-hoc formater. Et ad-hoc format er mer eller mindre tilfeldig laget av en utvalgt gruppe i bedriften og skreddersydd til bedriftens behov og følger ikke nødvendigvis noen standardisert oppskrift.

Dublin Core (DC) er et metadataformat for kategorisering av dokumenter som publiseres over Internett. DC har status som de-facto standard, men er et format som er forholdsvis nytt og som det fremdeles forsket på med hensyn til endringer og utvidelser.

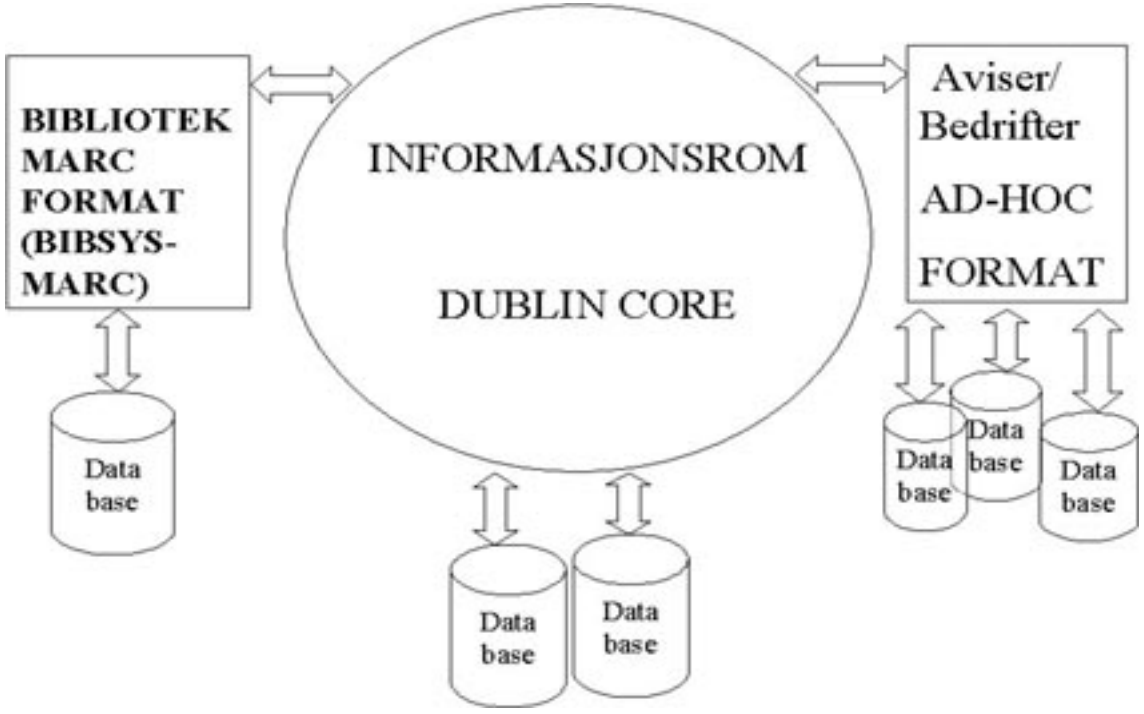
Bibliotekene har ulike tilpasninger av MARC formatet, og i Norge bruker vi NORMARC, mens andre varianter er DANMARC, USMARC.

Problemet med ulike format er at de ikke samsvarer med brukere og arbeidsoppgaver, fordi det er mange forskjellige måter å beskrive dokumenter på, alt etter hvilken type informasjon vi har med å gjøre. Det er ikke alltid like lett å tilby et felles grensesnitt mot alle medier, da medier og bruker har ulike behov som skal ivaretas, og det er ulike typer informasjonsobjekt som skal beskrives. Det er ikke alltid at de ulike informasjonsobjektenes og bedriftenes behov kan tilfredsstilles ved å velge et felles format, og da må de metadataformater som eksisterer harmoniseres. Vi ser at harmonisering av metadataformater danner grunnlaget for kommunikasjon og gjenfinning mellom bruker og informasjonsobjektene, nettopp fordi det er ønskelig å tilby et vindu mot verden fra flere informasjonskilder til deg som bruker for tilgang til disse informasjons-objektene. Uavhengig av om du søker etter bilder eller tekst, lydfiler eller annet skal det kunne gjøres via et web-vindu.

I dette prosjektet tenker jeg meg et fremtidig informasjonsrom som skissert i figur 1.1. hvor informasjonsressurser fra bibliotek og universitet, nyhetsmedia og Internett er integrert og fungerer sammen med et felles grensesnitt mot bruker. Siden disse ulike informasjonsmediene bruker ulike metadata formater, er det behov for å kunne utvikle en felles metadataontologi slik at disse formater skal kunne kommunisere med hverandre i et informasjonsromperspektiv.

Prosjektet er tverrfaglig og vil se på aspekt fra fagfeltene IT, bibliotek og kommunikasjons-teknologi. Som case for ad-hoc formatet i dette prosjektet har jeg valgt Adresseavisen sitt format. Siden nettopp et ad-hoc format er laget ut fra et brukerperspektiv hvor du ikke trenger spesielle kunnskaper om katalogisering for å bruke det, kan det være interessant å se hvordan dette formatet kan integreres sammen med DC formatet og BIBSYS - MARC formatet.

BIBSYS MARC er de metadata elementer universitetsbibliotekene i Norge har valgt å bruke for sitt elektroniske søkesystem for alle de dokumenter som finnes i Norske universitetsbibliotek. Prosjektet skal støtte de arbeidsprosesser som er gjeldene når du trenger å søke etter relevant informasjon fra mediene Internett, bibliotek og aviser. Det er nemlig brukerens interesser som skal ivaretaes og hvordan bruker best mulig kan finne frem til den informasjon vedkommende søker etter.



FIGUR 1.1. Virtuelt Informasjonsrom [1]

1.2. Avgrensninger

Det er mange problemstillinger å begrave seg i når det gjelder å bedre informasjonsgjenfinning i Internett og Intranett sammheng rundt om i bedrifter og organisasjoner.

1. Pilene viser informasjonsflyten mellom mediene

Det er mange interesser som skal ivaretas, faktisk dreier deg seg om en helt ny infrastruktur for det Internettsamfunnet vi kjenner i dag. Hver av problemområdene vil være en doktorgrad i seg selv, og dette er bare en hovedoppgave. Jeg har valgt å se på beskrivelsesmetadata og i hvilken grad metadata kan brukes i et integrert brukersystem for å være til støtte i informasjonsgjenfinningen ut ifra et brukerperspektiv og med hensyn på publisering av informasjonen via Internett og Intranett.

I mine vurderinger forutsetter jeg en aktiv bruker i informasjonssøkeprosessen som er i stand til å tolke og ta avgjørelser basert på den tilbakemelding bruker får fra systemet det skal benytte.

Med utgangspunkt i dette ser jeg på de ulike metadataformater som brukes til å beskrive informasjonsobjekter og hvordan disse kan beskrives på en bedre måte ut fra brukerbehov og med hensyn på en metadataontologi. Adresseavisen har ikke i dag en metadataontologi for alle informasjonsressurser. Med metadataontologi mener jeg en felles forståelse for indeksering av informasjonsressurser. Informasjonsressursene som jeg ser på beskrivelsen av er objekter i SIFT, Fotostation og Internettsystemet og består av fulltekst artikler, metadatadokument for fysisk film og illustrasjoner, digitale bilder og mulitmediadokumenter.

I den grad jeg bruker modeller i prosjektet, er dette for å vise informasjonsflyt og kommunikasjon mellom de ulike systemer, heller en tekniske spesifikasjoner og detaljer.

1.3. Mål for oppgaven

Hovedmål for hovedfags prosjektet er:

Fremlegge løsninger for hvordan DC eventuelt kan fungere som en metadataontologi for informasjonsrommet som integrerer mediene BIBSYS, Adresseavisen og Internett med tanke på enhetlig tilgang til og bedre gjenfinning av informasjonsobjektene.

1.3.1. Delmål:

1. Case: Se på hvilke spesialbehov Adresseavisen har og hvordan deres metadataformat presenteres, og om de er i samsvar med arbeidsprosesser de er ment å støtte. Metode for å få innsyn i brukerbehov er observasjon av hvordan metadataformat brukes og gjennomføring av spørreundersøkelse med fokus på bruk av metadata i informasjonssøkingsprosessen.
2. Kartlegge egenskaper og anvendelsesområder for DC, MARC og Ad-hoc formatene.
3. Undersøke om Adresseavisens beskrivelsesbehov for informasjonsobjekter kan tilfredsstilles i DC ved å foreta en mapping mellom de ulike Ad-hoc formatene mot DC.

4. Undersøke hvilken informasjon som tapes ved “mapping” av ad-hoc formatene til DC, og vurdere om "tapet" er viktig med utgangspunkt i brukerbehov som er kommet frem i observasjonsstudie og case studiet.

1.3.2. Mål for brukerundersøkelsen:

1. Finne ut hvilke metadataelementer som faktisk blir brukt ved søk når de daglige rutiner og arbeid skal utføres. Er det visse metadataelementer som er mer brukt enn andre av brukerne i informasjonssøkingsprosessen.
2. Er det forskjell på brukergrupper? Her ser en om det er ulike grupper som skiller seg ut sett i forhold til bakgrunnsinformasjon, opplæring m.m. Spørsmål å besvare:
    - Er det flere som bruker fritekstsøk enn metadatasøk?
    - Er det noen metadatasøkere som er mer hyppige brukere enn andre?
    - Er det noen grupper som bruker mer metadata ved søk enn andre grupper?
    - Vet alle hvor mange metadataelement som er søkbare?
    - Har alle de ansatte fått tilstrekkelig opplæring i bruk av SIFT databasen?
3. Hva er gode og dårlige metadata til bruk i søkeprosessen med tanke på effektiv informasjonsgjenfinning.?

1.4. Bakgrunn / Motivasjon for oppgaven

Aspekter og metoder rundt det å bedre gjenfinningen av informasjon er et område som har interessert meg før og under studietiden. Å lagre informasjon digitalt har interesse for å gi rask og effektiv tilgang til relevant informasjon. Gjennom min tid som student har jeg ofte vært frustert over og ikke finne den riktige informasjonen i den store informasjonsstrømmen som finnes på Internett i dag. Etterhvert som jeg lærte de ulike web-søkemotorene å kjenne og hvordan disse fungerte fant jeg ut at avansert søk var veien å gå får å få flere relevante treff, men selv ikke dette var nok til å møte mitt behov tilfredsstillende.

Internett er langt fra bra når det gjelder å finne frem i den informasjonen som publiseres der, og det er et stort behov for en felles infrastruktur for organisering og beskrivelse av informasjonsressurser. Det jeg har behov for er å finne relevant informasjon til rett tid som kan gi meg svar på de spørsmål jeg trenger til å utføre en arbeidsprosess eller de faktaopplysninger jeg søker. Flere metoder og løsninger trengs for å bedre informasjonsgjenfinningen på Internett i dag og akkurat dette er motivasjonen min for oppgaven.

Uansett hvordan informasjonen organiseres for bedre gjenfinning av riktig informasjon, er jeg overbevist om at riktig bruk av metadata vil spille en vesentlig rolle i denne sammenheng. Det er ikke snakk om at målsetningen er å organisere hele Internett, men åpne opp for web-arenaer for både strukturert og ustrukturert informasjon. Jeg ser for meg at det vil komme mange mulige tilnærmingsmåter for å organisere den informasjon som publiseres på Internett i dag. Organisering av informasjonsressurser i digitale bibliotek og i informasjonsrom er arenaer for strukturert informasjon som jeg tror vil ha stort vekstpotensial.

For interoperabilitet mellom ulike databaser, utveksling av informasjon fra ulike medier og bedre gjenfinning ved søk i informasjonsrom er det ønskelig med felles metadataontologier for å beskrive informasjonsressurser.Vi trenger ny kunnskap om hvordan en skal kunne bruke metadata tilfredsstillende for å støtte ulike brukerbehov både ved indeksering av informasjon og søk etter informasjon for å bedre kvaliteten på indeksering og gjenfinning. Jeg tenker meg at noen metadata er mer intuitive å bruke ved søk og indeksering enn andre, og akkurat dette ønsket jeg å finne ut mer om. Bruk av egnete metadata ut fra kontekst den brukes i vil kunne effektivisere indeksering og gjenfinning av informasjon. Akkurat dette avhenger både av hvilken informasjon som beskrives, brukerne som søker etter informasjonen, læringsteorier, språket oppbygning m.m. For å få ny kunnskap om hvordan bruke metadata riktig kreves det omfattende brukerstudier på området.

Nettopp fordi metadata skal være mest mulige tilpasset den naturlige måten å lære på, må ikke metadata som benyttes være for kompliserte. De må allikevel være effektiv for bedre kvalitet på gjenfinning og indeksering av informasjon.

Gjennom arbeid på arkivet på Adresseavisen og universitetsbiblioteket har jeg også sett at det ikke alltid er en selvfølge at du finner det du vet ligger lagret i databasen. I tillegg erfarte jeg at de aller fleste brukte fritekstsøk og ikke metadatasøk, noe jeg tenkte kunne være årsaken til at de ikke fant det de ønsket. Siden jeg selv kjente til bruk av metadata visste jeg at dette hjelp meg i søkeprosessen, og jeg ønsket å finne ut om det virkelig var så få som brukte metadata og hvorfor. I tillegg så jeg at det var lite samsvar mellom de ulike metadataformater som ble benyttet for de ulike databasesystem. De ulike metadataformatene hadde heller ikke hadde noe utspring i en felles metadataontologi, hvor samsvar mellom de ulike arbeidsprosesser som var i interaktivitet med databasesystemene var koordinert. Informasjonsobjekter var også lagret i flere databaser uten at metadata var brukt til å vise relasjoner mellom disse.

Ved å kartlegge brukerbehov tenkte jeg at det måtte være mulig å finne frem til metadata som var tilnærmet intuitive og tilpasset naturlige læringsprosesser for å effektivisere både indeksering og gjenfinning av informasjon. Dette til tross for at jeg viste at dette var et omfattende område som en brukerundersøkelse ikke nødvendigvis trengte å gi klare svar på.

Med tanke på å formidle informasjonen i databasesystemene gjennom Internett og Intranett, fikk jeg interesse for om DC kunne være tilfredsstillende nok for Adresseavisen med tanke på de brukerbehov de hadde.

Min hovedinteresse for informasjonsgjenfinning og oppdaging av metadata sine muligheter fikk meg interessert i å finne ut mer om hvordan riktig bruk av metadata kan bedre informasjonssøkingssprosessen og andre brukerbehov ved en bedrift. Dette er drivkraften bak arbeidet med denne oppgaven.

1.5. Metode / Planlegging av oppgaven

En metode er i følge Tranøy [34] en fremgangsmåte for å frembringe kunnskap eller etterprøve påstander som fremsettes med krav om å være sanne, gyldige eller holdbare. Vi ønsker altså å bruke en fremgangsmåte for å analysere datamaterialet vårt, og ut i fra dette danne oss et resultat som kan gi oss svar på det vi ønsker å løse, og dermed gi oss ny kunnskap om et problemområde. Det er viktig å være kritisk til bruk av metode, alt ut i fra hvilke svar du ønsker å få. Som oftest velger en enten kvantitativ eller kvalitativ metode som utgangspunkt for analyse, men det er også fullt mulig å benytte "metode-triangulering" som er en kombinasjon av disse.

I avhandlingen min benytter jeg spørreskjema, observasjoner, samtaler med ulike brukere av SIFT, dataansvarlige og andre nøkkelpersoner ved Adresseavisen og Statens Data Sentra. I tillegg kommer egne erfaringer med SIFT som ansatt ved Arkivet ved Adresseavisen.

En rekke brukerundersøkelser er gjort når det gjelder søkeoppførsel mot databaser, men ingen av disse har fokus direkte på hvilke metadata som brukes i søkeforespørsler etter informasjon, hvordan du forstår metadata, navngivning av metadata etc. Dette er et område som har bred interesse i forskning innen digitale bibliotek, men er likevel et område vi vet for lite om i dag. Det har ikke lyktes meg å finne noe publisert arbeid som viser til ny kunnskap om bruk av metadata i informasjonssøkingsprosessen.

Forhåpentligvis vil min forskning gi resultater på akkurat dette, som vekker interesse for videre undersøkelse og forskning på bruk av metadata i integrerte brukersystem med hensyn på de ulike brukernes informasjonsbehov.

1.5.1. Datamateriell

Datamateriellet som her foreligger er :

- resultater fra brukerundersøkelsen
- informasjon om de ulike formaters brukskontekst basert på observasjon, samtaler og egne erfaringer
- informasjon om de ulike databaser sine kommunikasjonslinjer basert på samtaler med data ansavarlige i Adresseavisen og andre nøkkelpersoner.

I avhandlingen benyttes kvalitative metoder. Spørreskjemaet var i utgangspunktet ment å være kvantitativt, men p.g.a den lave svarprosenten er det ikke mulig å analysere dataene statistisk.

For å se om Adresseavisens behov kan tilfredsstilles i Dublin Core (DC) formatet bruker jeg en fremgangsmåte som kalles mapping. Å mappe vil si å ta hvert enkelt metadatalement de ulike ad-hoc formatene og ut fra brukskontektsen se om DC har et tilsvarende element som dekker dette behovet. For de element som eventuell ikke mapper vil det vurderes om DC har element som kan tilfredsstille dette behovet på en annen måte. Dersom det etter dette fremdeles står ad-hoc element igjen som enda ikke er mappet, vil disse elementenes relevans bli drøftet og vurdert ut fra arbeidsprosesser og resultater fra undersøkelsen. Resultatet av mappingen er et felles format for indeksering av informasjonsobjekt ved Adresseavisen.

Dersom bare en bruker har et behov for et metadataelement ved søk, eller til bruk i sine arbeidsprosesser bør dette bli tatt i betraktning. En vanlig oppfatning er også at selv om få brukere er representert er det ofte slik at dersom en bruker opplever noe som et problem, er det også sannsynlighet for at flere har opplevd det samme. Dersom de behov som her er representert da blir tatt i betraktning vil det tilføre en forbedring til disse brukernes arbeidsprosesser.

Det som kjennetegner den kvantitative metoden er den fordel at informasjonen kan formes inn til målbare enheter, som vi igjen kan dele inn i ulike brukergrupper for sammenligning mellom disse opp mot ulike faktorer som kan påvirke brukeroppførselen og brukerbehov. Når svarprosenten blir såpass lav som i dette tilfellet må resultater analyseres kvalitativt. Når variasjonene mellom brukerne er på 1 eller 2 i overkant eller underkant sier det seg selv at det ikke er noe grunnlag å utføre statistiske beregninger på.

Det som kjennetegner den kvalitative metoden er at den fanger opp meninger og opplevelser hos den enkelte bruker som ikke lar seg tallfeste eller måle, men som kan drøftes opp mot resultater fra kvantitativ metoder, for bedre forståelse av brukerens behov. Den kvalitative innsamlingen min som beskrevet i Case-studiet vil dermed være farget av min forståelse for situasjonen og avhengig av min tolkning av det hele. I observasjon og samtale vil en da legge merke til det jeg

synes er viktig, noe som kan føre til at jeg ikke har fanget opp alle viktige detaljer. Dette er feilmarginer som må tas i betraktning.

I presentasjon av resultater fra undersøkelsen benytter jeg stolpediagram og tabeller. Dette er metoder som også kan brukes i kvalitative analyser av et spørreskjema i følge Robert Johnson [23]. Johnson sier at stolpediagram og kakediagram er metoder som kan brukes til kvalitative analyser.

Ut i fra at jeg selv har arbeidet ved arkivet mener jeg å kunne trekke inn egne erfaringer i drøftinger av resultater fra undersøkelsen. Egne erfaringer vil være subjektive, men siden jeg har god kjennskap til bruk av SIFT er jeg en bruker av systemet på lik linje med øvrige deltakere i undersøkelsen.

1.6. Kapittelinndeling

Oppgaven har fått totalt 9 kapittel. Disponeringen av disse er som følger:

Kapittel 2. Fagfeltet informasjonsforvaltning

Dette kapittelet gir deg innsyn i informasjonsforvaltning og digitale bibliotek.

Kapittel 3. Metadata

Introduserer deg for begrepet metadata, dets muligheter og bruksområder i informasjonsrom og ellers. Dette er et viktig kapittel for forståelse av videre tema i prosjektet.

Kapittel 4. Informasjonssøkingsprosessen

Her får du vite hvilke faktorer som påvirker informasjonssøkingsprosessen, samt tradisjonelle gjenfinningsmetoder i denne prosessen. Du får også vite litt om informasjonssøking via søke-motorer, forskjeller mellom fritekstsøk og metadatasøk og litt om journalistisk søkeoppførsel som er den største brukergruppen i undersøkelsen.

Kapittel 5. Metadataformater

Her får du innsikt en kort innsikt i MARC og BIBSYS-MARC og dets historikk og anvendelser. Deretter følger en grundig redegjørelse for Dublin Core formatet (DC), dets sub-element, historikk, målsetninger, utviklingsstatus og lagringsmuligheter.

Kapittel 6. Case-studiet.

Dette er en beskrivelse av situasjonen ved Adresseavisen, datanettverket, SIFT databasesystem, og deres arbeidsprosesser. En sentral del i dettet kapittelet er også presentasjonen av de ulike ad-hoc formatene som benyttes til å beskrive de ulike informasjonsressursene og de ulike metadataenes brukskontekst.

Kapittel 7. Resultater

Resultater fra undersøkelsen presenteres og resultater og konklusjoner oppsummers. Kapittelet inneholder flere tabeller og stolpediagram for å visualisere dette.

Kapittel 8. Mapping

Her får du se om Dublin Core kan tilfredsstille Adresseavisen sitt behov for å beskrivelse av informasjonsobjekter. Forslag til et kjerneformat legges frem, med medieavhengige variasjoner.

Kapittel 9. Evaluering av oppgaven

Her drøftes kort de erfaringer jeg har gjort meg i prosjektet, eventuelle feilvurderinger og forslag til videre arbeid.

Kapittel 2

Fagfeltet Informasjons-forvaltning

2.1. Informasjonsforvaltning

Informasjonsforvaltning (IF) er en norsk oversettelse av begrepet Information management, ofte også kalt information resource management (IRM). Dette er et tverrfaglig fagfelt som spenner vidt, men som har sterke elementer fra kunnskap og informasjonforvaltning i seg.

I hovedsak omfatter det funksjoner som innsamling, behandling, styring og anvendelse av informasjon og kunnskap. Dette innebærer:

- å samle inn informasjonen,
- behandle og analysere datamaterialet
- sørge for forsvarlig og hensiktmessig lagring
- gjøre informasjonen tilgjengelig på en adekvat måte med hensyn på gjenfinning.

Men IF er også mye mer, som skaping av ny informasjon og kunnskap gjennom læring i organsisasjonen, f.eks ved å kombinere eksisterende kunnskap på en ny måte eller å lage systemer for å dele informasjonen til støtte i arbeidsprosesser. Kunnskapen og informasjonen har til hensikt på anvendes på en produktiv måte, og være et verdiskapende element i produksjonen. I slike systemer er også brukervennlige grensesnitt, konstruksjon av effektive gjenfinningsrutiner og modellering av informasjons-og kunnskapssystemer også viktige sider når en skal vurdere hvordan informasjonen og kunnskapen skal forvaltes i ulike kontekster.

2.2. Digitale bibliotek

Av det som er nevnt av funksjoner som inngår i IF, ser vi at det meste av dette også berører digitale bibliotek i stor grad. Flere og flere bibliotek verden over velger å digitalisere samlingene sine, så dette er absolutt en fremtidens arena for IF.

Digitale bibliotek er komplekse informasjonssystemer hvor store datamengder er automatisering, og hvor informasjonsressursene er strukturert beskrevet. Til sammenligning er ikke Internett er ikke noe digitalt bibliotek, da det inneholder store mengder utstrukturert informasjon.

Michaeal Lesk definerer digitale bibliotek slik: "*Digital libraries are organized collections of digital information. They combine the structure and gathering of information, which libraries and archives have always done, with the digital representation that computers have made possible.* "

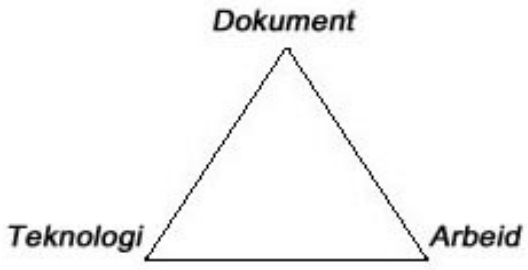
En annen dekkende definisjon fra "Digital Libraries Federation (DFL) (1999), USA"

"*Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by defined community or set of communities.*"

Ut i fra disse definisjonene ser vi at digitale bibliotek favner et vidt område, og omfatter alle områder av informasjonsforvaltnings. Siden digitale bibliotek er et forholdsvis nytt forskningsområde gjenstår det mye forskning på området, siden det en enorm informasjonsforvalter og mange hensyn som må vurderes med hensyn på gjenfinning, katalogisering, copyright m.m for vi kan snakke om digitale bibliotek på lik linje med hva tradisjonelle bibliotek kan tilby. Alikevel er det en målsetning at digitale bibliotek skal fungere på samme måte som tradisjonelle bibliotek og oftest ønsker en å tilby tilgang til samlingene via Internett. Flere universitets og forskningsbibliotek verden over har allerede digitalisert noen av sine samlinger som er tilgjengelig via Internett, som f.eks Alexandria Digital Library, Standford m.m.

Viktige områder innen forskning er metoder, algoritmer og programvare for bedre tilgang til informasjonen gjennom søk, innhenting, manipulering m.m av informasjon og hvordan en skal presentere informasjonsobjekter. I tilegg forskes det på utvikling av intelligente og brukervennlige brukergrensesnitt, brukeroppførsel, hvordan digitale bibliotek skal brukes i utdanning og læringsprosesser og også økonomiske og sosiale faktorer knyttet til dette. Dette er så stort område at det er umulig å ramse opp alt. Bruk av metadata og organisering av informasjon gjennom arkitekturbygging er sentrale pukt innen disse forskningsområdene.

Marshall og Levy [25] viser følgende perspektiv på bibliotek som beskrives av figur 2.1.



FIGUR 2.1. Perspektiver på bibliotek

Dokument representerer her de dokumentsamlinger som biblioteket forvalter. Teknologi er det vi trenger for å gjøre dokumentene distribuert tilgjengelig for brukeren, for å frembringe flere eksemplarer av dem, trykke dem osv. Vi kan derfor si at teknologi gir oss mulighet til å massedistriubere samlingene og forvalte informasjonen i større skala. Arbeid er det arbeid som utføres av bibliotekarer og andre som arbeider med dokumentene, f.eks beskrivelse og klassifisering av dokumenter kommer inn under arbeid. Altså trenger vi både dokumenter og teknologi for å arbeide med dokumenter.

Disse tre perspektiver kan videreføres til digitale bibliotek, som ifølge Marshall og Levy får denne betydning:

- Dokument: Digitale bibliotek sine samliner består av dokumenter som er dynamiske og statiske. Noen dokumenter endrer seg aldri, mens andre typer dokumenter som f.eks togtabeller og busstabeller endres stadig når rutetider må legges om. De fleste dokumenter ligger omtrent midt i mellom de dokumenter som aldri endrer seg og de som hele tiden endres. I tillegg kan en også ha dokumenter som inneholder statiske deler, og som også består av deler som er under endring.
- Teknologi: Digitale bibliotek er basert på digital teknologi, dvs datateknikk og informasjonsvitenskap.
- Arbeid: Digtale bibliotek er tenkt brukt av individer og grupper som arbeider alene og sammen. De skal også støtte ulike arbeid- og samarbeidsprosesser.

2.3. Terminologi

Ikke all terminologi har en felles forståelse, og desto mer det er "i tiden" å bruke ord og uttrykk desto mer brukes de, med det resulat at meningen risikerer å blir “vannet ut”. Siden mange forfattere lager sine egne definisjoner og vinklingsmåter, vil jeg her gi en kort presentasjon av generelle begreper og uttrykk som blir brukt igjennom denne oppgaven, slik at du som leser skal vite hvordan de skal forståes i denne sammenhengen.

2.3.1. Repository

Dette er et annet navn for “lager”, dvs at en database hvor du lagrer digitale objekt er et repository. Et repository har mekanismer for å legge til nye digitale objekt i samlingen, og for å gjøre dem tilgjengelig for andre, via f.eks nettressurser og andre distribuerte systemer.

2.3.2. Digitalt objekt/Informasjonsobjekt

Et informasjonsobjekt er et objekt som er bærer av en eller annen form for informasjon. Dette informasjonsobjektet kan være digitalt lagret, men kan også eksistere på mange andre måter. Eneste kriterium for et informasjonsobjekt er at det inneholder informasjon. Et informasjonsobjekt som er digital lagret, kalles da et digitalt objekt. Et informasjonsobjekt som er digitalt lagret, kan være alt fra et rent tekstdokument, til bilde og multimediapresentasjoner. Før brukte en begrepet digitalt objekt om informasjonsobjekt som var digitalt lagret, men i dagens postmoderne samfunn hvor ALT dreier seg om informasjon og kunnskap har altså begrepet blitt utvannet til fordel for informasjonsobjekt. Begrepet digitalt objekt brukes alikevel fremdeles, men jeg synes det er mer dekkende for mer "teknisk ladede" vinklingsmåter, som når det er snakk om objekter som er bærere av informasjon på maskinvarenivå .

Jeg ønsker at leserne skal assosiere informasjon eksplisitt til det objektet som er bærer av informasjon, noe de får ved å benytte begrepet informasjonsobjekt i stedet for digitalt objekt. I og med at jeg i oppgaven min ikke kommer inn på tekniske detaljer og løsninger knyttet til overføringer av informasjon, men ser på informasjon ut fra et mer overordnet abstrakt nivå, velger jeg å bruke begrepet informasjonsobjekt.

2.3.3. Dokument /Digitalt dokument/ Dokumentasjon

Buckland [6] sier at den ordninære betydningen av dokument visert til en tekstlig bibliografisk post/enhet. Tidlig på bgynnelsen av dette århundret ble dokumentbegrepets betydning tatt opp til revurdering. Årsaken til dette var den stadig større fremveksten av dokumenter og etterspørselen etter dette, og om det var på sin plass med en utvidelse av begrepets betydning. Buckland sier

blant annet at Paul Otlet var en av dem som reiste spørsmål om f.eks en skulptur eller et objekt i et museum kunne betraktes som et dokument.

Ordet dokument er nært relatert til ordet dokumentasjon. Den økende etterspørselen etter tilgang til flere dokumenter økte behovet for flere eksemplarer til utlån og distribusjon. Aktivitetene med å samle inn, preservere, organisere, beskrive/indeksere, velge ut, reprodusere og distribuere dokument ble tidligere kalt for en bibliografi (samleverk), men dette begrepet måtte vike for begrepet dokumentasjon som ble innført i rundt år 1920. Senere er ordet dokumentasjon blitt mindre brukt til fordel for nye begrepet som informasjonsvitenskap, informasjonslagring, informasjonsgjenfinning, og informasjonsforvaltning.

Alikevel forekommer begrepene side om side, og må vel heller ses på som en utvidelse av ordforrådet, heller enn at det er store forskjeller mellom begrepene. Ordet dokument brukt igjennom denne oppgaven skal forstås i betydningen "Alt som er bærer av informasjon", hvor både en artikkel, et bilde, en metadatapost m.m er å betrakte som et dokument.

Et digitalt dokument er bare en detaljering av begrepet dokument hvor eneste forskjeller er at dokumentet er lagret digitalt.

En vanlig oppfatning av bruken av ordet dokument har vært at det er noe som omfatter en skrevet tekst, f.eks et brev, en rapport, en bok osv, som hovedsaklig har vært informasjon i papirform. Etterhvert har begrepet fått en videre tolkning og bruk enn bare det å inneholde tekst, nettopp fordi digitalisering av dokumenter har gjort et mye mer sammensatt av mer enn bare tekst. Ordet dokument rett oversatt fra fremmedordboken er definert som et skriftstykke, en skriftlig redegjørelse, aktstykke, eksemplar av et medium som lagrer informasjon for senere overføring, lytting eller lesing, så det er absolutt rom for en videre tolkning her. Dokumentbegrepet brukes i dag om noe som er bærer av informasjon, det være seg en rapport, et bilde, et mulitmediadok-ument, en video m.m. Ordet dokument gir oss også assosisasjoner til ordet dokumentalist som er veldig likt og er knyttet til det å klassifisere og gjenfinne informasjon og vi ser at dokument begrepet trolig har oppstått og vært mest brukt omkring informasjon som har vært tilgjengelig via bibliotek. Ved fremveksten av digitale bibliotek har vi også fått begrepet digitale dokument, som ikke har annet annerledes fra ovennenvte beskrivelse annet enn at informasjonen er lagret elek-tronisk.

2.4. Forkortelser brukt gjennom avhandlingen

Disse forkortelser brukes gjennom oppgaven:

ID	= Identifikator
DC	= Dublin Core Metadata Element Set
MARC	= Machine Readable Catalogue Format
RDF	= Resource Description Framework
/ill-id/	= Direkte referanse til et bestemt metadataelement i et format, hvor **/ill-id/** er metadataelementet. Dette er for å skille metadataelementene fra annen tekst, f.eks når du har metadataelementet **/produkt/** og andre som kan forveksles med vanlig tekst.
IM-bruker	= Ikke metadata bruker (fritekstsøker)
M-bruker	= Metadatasøker

Kapittel 3

Metadata

"Metadata sounds sexy, but it really stands for cataloging" - David Seaman [9]

3.1. Hva er metadata?

Metadata er kort definert "data om data". Du bruker et sett av elementer til å beskrive et dokument eller et objekt med den hensikt at du skal kunne finne tilbake til dokumentet. Elementene er attributtene til et informasjonsobjekt. Det finnes flere definisjoner på metadata. Noen er videre enn andre, alt etter hva de skal beskrive og i hvilken kontekst de skal brukes. I BIBLINK prosjektet [7] beskrives metadata som:

> *"data which assists in the identification, description, evaluation and selection of an information object"*

Metadata kan være integrert i et dokument, men kan også eksistere som egne metadataobjekt som er knyttet til dokumentet.

Metadata er et begrep som brukes i forbindelse med klassifisering av elektronisk lagrede informasjonsobjekter/informasjonsressurser, og sammenlignes ofte med tradisjonell katalogisering der du setter data som tittel, forfatter, utgivelsesår m.m i tilknytning til en bok eller et tidskrift for å kunne identifisere disse. De opplysninger som ligger på katalogiseringskortet og som gjør oss i stand til å finne boken på hyllen er eksempel på metadata. Du bruker ulike sett av metadata helt avhengig av hvilken type brukere du har, hvilke prosesser som skal utføres og hvilken type informasjonsressurs du skal representere. Dette kaller vi for ulike metadataformater. *F.eks så er MARC formatet utviklet med tanke på å tilfredsstille de behov et bibliotek har for beskrivelse av informasjonsobjekter og er da tilpasset de arbeidsoppgaver som bibliotekarer, katalogisatorer og andre i biblioteket utfører.*

I databaseforskning kan metadata tolkes som:

- enheten som beskrives er et dataelement
- fokuseringen er på integrering av ulike databaser

I en digital biblioteks verden kan metadata tolkes som:

- enheten som beskrives er en informasjons ressurs
- fokuseringen er på informasjonsoppdagelse og administrasjon.

Metadataformater har er et avtalt sett av dataelement med

- avtalt/enighet om semantikk
- avtalt syntaks
- avtalte regler

for å beskrive innholdet i informasjonselementene (IE)

3.2. Hva brukes metadata til?

Metadata brukes til:

- søking
- lokalisering
- utvelgelse
- semantisk interoperabilitet
- ressurs administrasjon

Du trenger metadata til:

- Oppdagelse: Hva eksisterer? Hvor kan jeg få tilgang til det?
- Grensesetting og betingelser: Hvor mye vil det koste?
- Kontekst: Hvem skapte det? Hvorfor? Som ledd i hvilken prosess?
- Struktur: fil format, hvordan det skal representeres

- Innhold: hva er i objektet? Hva handler det om?
- History of use: Hvem er ansvarlig for det? Hvordan har det blitt brukt tidligere?
- Lenking, påvise relasjoner/sammenheng: lenker til annet arbeid?

3.3. Ulike typer metadata.

Vi har tre hovedtyper av metadata. Disse er:

- Beskrivelses metadata (tittel, forfatter m.m)
- Tilgangsmetadata (IP-nr, servernavn)
- Administrative data.

3.4. Ulike metadataformater.

Vi har ulike typer metadataformater, og vi deler dem inn i to kategorier :

- enkle metadataformat (Altavista, hotbot)
- rike metadataformat (MARC, RDF)

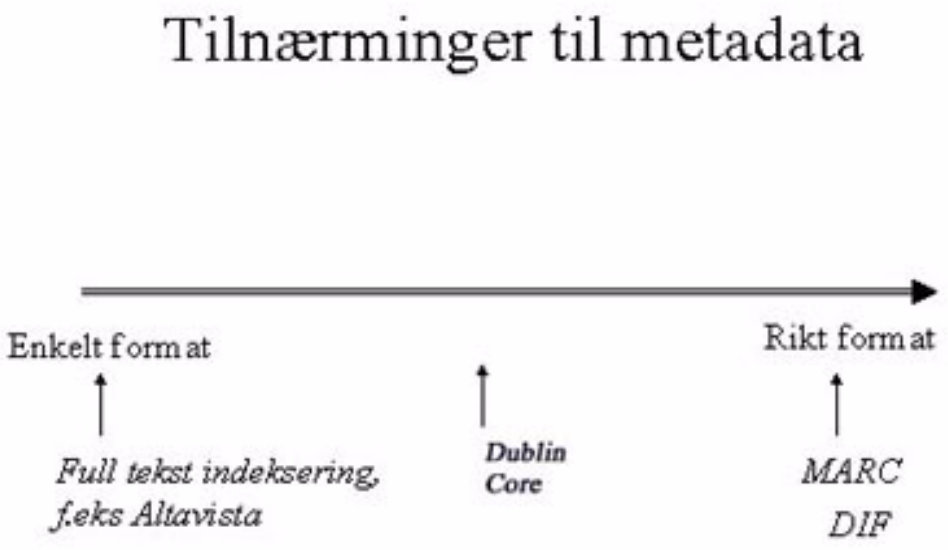
Enkle formater:

- kan bli brukt i mange kontekster
- dokumenter er lette å finne
- generell eiendomsrett
- er hovedsaklig brukt for å lokalisere ressurser

Rike formater:

- Brukes innenfor avgrensete arbeidsområder (Bibliotek hovedsaklig)
- Kompilering av dokumenter krever folk som er opplært i formatet
- Tillater å definere sub-felter
- Lav interoperabilitet (dokumenter er ofte skapt av eksperter og brukt av datamaskiner)

- Er i samsvar med internasjonale standarder
- Er brukt ikke bare til å lokalisere objekt, men er også brukt som basis for mer sofistikerte analyser og navigeringsverktøy



FIGUR 3.1. Tilnærminger til metadataformater

3.5. Metadata i forhold til tradisjonell katalogisering

I bibliotekskatalogisering beskrives metadata som:

- Det å organisere en samling med bibliografiske element med den hensikt å skulle lette arbeidet med å identifisere, lokalisere, gi tilgang til og bruke samlingen.
- En katalog er i seg selv en samling, og en samling er et surrogat for en post i den opprinnelige samlingen.

Jeg skal ikke her invitere til "begrepsslåsskamp" om hvilket fagmiljø som skal ha rettigheter til å bruke begrepet metadata, selv om det ved første øyekast kan virke slik. Ut fra beskrivelsen av metadata i biblioteksammenheng ovenfor og den mer generelle "data om data" definisjonen for begrepet er:

- et katalogkort i et bibliotek

- attributter i en relasjonsdatabase
- tag’er i et HTML dokument

dekkende for begrepet metadata.

Metadata er et begrep for katalogisering som er blitt innført i forbindelse med fremveksten av digitale bibliotek, og er ikke et begrep som primært har vært brukt i forbindelse med tradisjonell katalogisering tidligere, selv om det ikke ville vært feil å bruke det i forhold til begrepets definisjon. Metadata begrepet ble derfor innført for å markere et paradigmeskille mellom tradisjonell katalogisering og digital katalogisering. Det er derfor ulikheter mellom tradisjonell katalogisering og digital katalogisering vi vil se på her. Metadata begrepet skal her forstås i sammenheng med digital katalogisering.

Stefan Gradmann sier at bruk av metadata gjør at en må tenke igjennom relasjoner om beskrivende data og referanse element som er spesielt for elektroniske dokumentlignende objekter, såkalte Digital Like Objects (DLO’s).

Gradmann [6] mener at det er fundamentale forskjeller mellom metadata og katalogisering både når det gjelder metadata produksjon, konteksten den brukes i og relasjonen mellom metadata og de objekter som refereres til av metadata og katalogiseringsposter. Gradmann sier også at det er nødvendig å skille ut disse forskjellene for å redefinere bibliotekarenes rolle i det raskt voksende informasjonsparadigmet.

Artikkelen tar utgangspunkt i de facto standarden Dublin Core (DC), for å belyse forskjeller mellom katalogisering og metadata. Gradmann [6] går såpass langt som å påstå at metadata og katalogisering er ikke bare fundementalt forskjellig, men kan også stå i konflikt til hverandre når det gjelder arbeidskonsepter som ligger til grunn for begge modellene.

Metadata har som hensikt å beskrive et informasjonsobjekt, og DC formatet er et veldig forenklet katalogiseringsformat sett i fra bibliotekets rolle som har stor grad av detaljering i sitt katalogiseringsarbeid. Alle katalogiseringsformater laget for den digitale verden er enkel, da hensikten er at alle brukere av Internett til en viss grad skal kunne katalogisere det de publiserer på nettet og ikke trenge kunnskapen til en bibliotekar for å kunne gjøre dette. Metadata skal i tillegg forstås av maskiner noe som ikke har vært nødvendig ved tradisjonell katalogisering som kun har vært ment for det blotte øyet til en bibliotekar. Et annet viktig aspekt ved DC metadata og andre metadata er at det skulle lette arbeidet med å oppdage informasjonsressurser i et digitalt nettverksmiljø, og ikke nødvendigvis skille ut ressursbeskrivelser. Det å oppdage ressursene har sterkere fokus i metadata sammenheng.

Relasjonen mellom metadata og ressursreferanser skiller seg substansielt fra relasjonen mellom en katalogiseringspost og en bok som finnes i et bibliotek. Katalogisering er ofte veldig detaljert og ikke så lett å forstå for andre enn bibliotekarer. f.eks MARC formatet som har mange hundre ulike felter for beskrivelse. Katalogisering er dermed ikke rettet spesielt mot slutt bruker, noe som er tilfelle med metadataformater som har en klar sluttbrukerprofil. Dette er det fordeler og ulemper med fra begge sider. Bibliotekene har blitt mer og mer oppmerksom på at katalogisering ikke er særlig bruker orienterte og er villig til å se på hva som kan gjøres bedre. En sterk grad av brukerorientering gjør også formatet veldig sårbart mot endringer i sluttbruker oppførsel og brukskontekster, og kan dermed risikere å ikke tilfredsstille kontinuitet. Tradisjonell katalogisering har også vært basert på fysisk levering av dokumentet, med en peker fra katalogiseringskortet til plasseringen på hyllen. Men en katalogiseringspost kan også ha mer detaljert tilgangsinformasjon, *f.eks hvor på nettet man eventuelt finner dokumentet i fulltekst.* Problemet med metadata som beskriver metadata ressurser er at linker kan endres og man har plutselig *"broken links"*, noe som enda ikke er et løst problem. En kan vel si at en metadata post med en brutt link, er like verdiløs som ingen metadata post, forutsatt at denne linken viser vei til dokumentet i fulltekst. Men det kan oppstå såkalte "broken links" i et bibliotek med fysisk plassering av bøker også. Dersom du tar en bok ut av hyllen og setter den inn på feil plass i hyllen, er det tilnærmet umulig å finne den igjen før det skjer ved en tilfeldighet eller oppryddingskontroll, og boken er dermed å regne som *"tapt"*.

En metadata post og en katalogiseringspost har hovedsaklig forskjellig produksjonsparadigme og brukskontekst, og begge tilhører forskjellige infrastrukturer for informasjon.

Til slutt har du konsistens av autoritet og gyldighetsproblemet som har vært til stor bekymring for bibliotekene og hvordan dette skal ivaretaes på digitale samlinger med publisering mot Internett. *"Internett slik vi kjenner det i dag ble ikke konstruert for å støtte organisert gjenfinning slik som bibliotekene har som oppgave"* påpeker Ole Husby [20] og det er derfor en utfordring å se hvordan dette kan ivaretas ved hjelp av metadata.

Ann Chapmann m.fl.[8] summerer opp disse forskjellene mellom tradisjonell katalogisering og metadata:

- beskrivelse (bibliotekene mye mer detaljerte enn metadataformater)
- headings (I tradisjonell katalogisering søker du på overskrifter som forfatter, tittel osv, mens du på Internett har mulighet til å søke i alle felter samtidig og også i abstract feltet hvor du kun har tekstlig beskrivelse)
- autoritetskontroll (Bibliotekene bruker systemer som BLNAL for autoritetskontroll, noe som gjør at det er stor grad av konsistens innenfor fellesskapet, dvs at navn skrives på en bestemt måte. Ved bruk av metadata kan en rekke ulike skrivemåter av termen benyttes, for å få treff).

- utvelging (Biblioteksansatte med kunnskap innenfor et bestemt område og noen ganger også akademiske personalet står for utvelgelse av informasjonen, og dette er en separat oppgave fra katalogisering. Noen ganger kan også brukere foreslå hvilke dokumenter som skal kjøpes inn. Når det gjelder Internett ressurser involverer dette i tillegg oppdagelse av ressurser og kvalitetssikring, fordi en ikke vet hva som ligger der i utgangspunktet. Bibliotekene har dermed utvelging, mens Internett har oppdagelse av ressurser)
- Besittelse av informasjon og tilgangen til informasjon. (Bibliotekene opererer med fysiske pekere til dokumentet på hyllen som f.eks et plasseringsnummer. Du får også opplysninger om dokumentet befinner seg ute på lån og hvilket format dokumentet er i. På internett snakker vi om tilgang til ressurser istedenfor beholdning av ressurser. Det er bare de katalogiseringspostene som beskriver ressursen som befinner seg på serveren, resten er en virtuell samling, hvor du blir *"linket"* videre til den aktuelle informasjonsressursen via en gateway til den aktuelle serveren som informasjonen ligger på. Det virtuelle bibliotek her består av ressurser som ligger på servere over hele verden. Katalogiseringspostene på Internett må hele tiden sjekkes opp for å se at "linkene" til den faktiske ressursen til enhver tid er operativ. I SOSIG prosjektet i England [73] har de en automatisk link sjekker som kontinuerlig sjekker for *"broken links"*, slik at disse kan oppdateres eller slettes fra databasen.)

3.6. Metadata i forhold til informasjonsrom /informasjonsdeling og corporate memory arbeidsmiljø?

Kenneth A. Megill [27] definerer Corporate Memory til å være all den historiske informasjonen i en bedrift som er:

- verdt å dele
- administrere og
- ta vare på

med den hensikt at den skal gjenbrukes.

En ønsker å samle hele bedriftens kunnskap slik at den kan være en ressurs for gjenbruk og skapelse av ny kunnskap. Informasjon som inngår i et corporate memory er alt fra bedriftens metoder, analyser, korrespondanse, ideer, markedsføringsinformasjon, tekniske rapporter, bibliotekskataloger m.m. Her er det viktig at ikke all informasjon i en bedrift skal lagres, som

f.eks personlig informasjon, ulike versjoner av dokumenter osv, men at det faktisk er informasjon som kan gjenbrukes. I AIS [1] prosjektet [29] brukes begrepet informasjonsrom på lik linje med beskrivelsen av et Corporate Memory som gjøres tilgjengelig via bedriftens Intranett. Når en bruker metaforen informasjonsrom fokuserer en på den informasjonen i Intranettet som er felles tilgjengelig. Dette informasjonsrommet strukturerer informasjonen slik at den støtter publisering, lagring, gjenfinning og forvaltning av informasjon, og er dermed å regne som et digitalt bibliotek. Et intranett kan bestå av flere informasjonsrom hvor strategier for lagring og gjenfinning av informasjon kan variere.

Det å bygge et informasjonsrom som skal fungere effektivt er en komplisert sak, og det finnes ingen standard oppskrift som passer til den enkelte bedrift. Det er i midlertid en del grunnleggende prinsipper som er viktig å holde orden på når et informasjonsrom skal genereres.

De første stikkordene er byggeklosser og relasjoner, som er beskrevet av en modell. Denne modellen omtales som en metamodell som beskriver alle de ulike komponentene som inngår i et informasjonsrom. De vesentligste byggeklossene å starte med her er:

- De ulike informasjonsressurser, det være rapporter, brev, saksdokumenter m.m.
- Begrepsmodeller, såkalte ontologier som definerer strukturer
- Perspektiver som tar utgangspunkt i brukerbehov
- Presentasjoner av ulike brukerbehov

Et informasjonsrom er altså en integrering av felles tjenester og ressurser fra ulike medier, som skal støtte en eller annen arbeidsprosess og ulike brukerkontekster og hvor kunnskapen til den enkelte skal ivaretas i systemet og kunne gjenbrukes.

Vi snakker om informasjon som en felles ressurs, der alle kan benytte seg av og gjenbruke den informasjon som eksisterer. Tenker vi i arbeidssammenheng er den informasjonen og kunnskapen som alle de ansatte sitter inne med, bedriftens “Corporate memory”. Men det er først når denne “Corporate memory” digitaliseres at alle ansatte kan dra nytte av hele bedriftens samlede kunnskap i sine daglige arbeidsprosesser, fordi den da er lett å akksessere via Intranettet i bedriften.

1. AIS=Avansert Intranett Samarbeid

3.6.1. Hva er en Metamodell

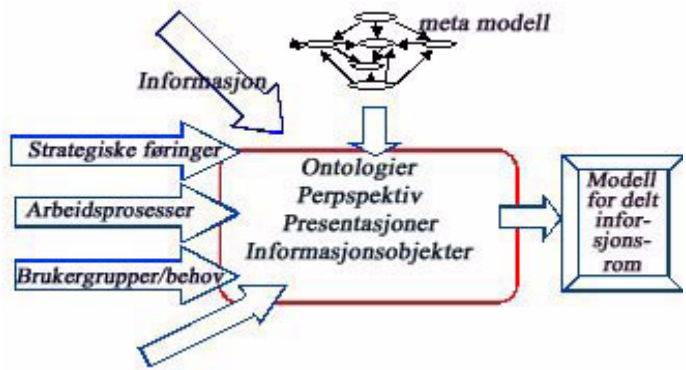
Dette begrepet benyttes i forbindelse med Informasjonsrom, og er en modell over strukturen i et informasjonsrom, og relasjoner mellom de ulike "byggeklossene" som et informasjonsrom består av.

I AIS prosjektet til Sintef [29 , s 23] er en metamodell illustrert som i figur 3.2.

En instans av modellen i figur 3.2 er en modell av et informasjonsrom, hvor egnede begrepsmodeller angir hvordan informasjonsrommet skal bygges opp, er modellert. Vi ser her at en modell for et informasjonsrom skal inneholde informasjonsobjekter, ontologier, perspektiv, presentasjoner og informasjonsobjekter. En metamodell skal forutenom å brukes til å generere selve informasjonsrommet også tjene som kommunikasjonsverktøy for struktur, informasjonskategorisering og brukerperspektivtilpasning. Videre skal den også definere terminologi som brukes til å beskrive domener og informasjon, og sist men ikke minst skal det være mulig å utvide og endre på modellen når det er behov for det.

3.6.2. Arbeidsprosesser og informasjonsprosesser

En arbeidsprosess er de oppgaver som inngår i rollen i kraft av å være f.eks fotograf eller journalist, mens en informasjonsprosess er et behov for informasjon til å utføre en arbeidsprosess. En rolle kan ha mange arbeidsprossesser, og mange brukerbehov. En rolle inngår ikke direkte i en stillingsbetegnelse, for selv om man er journalist kan man også spille rollen som fotograf. En fotograf har andre brukerbehov enn en journalist, dermed vil rollen din spesifisere brukerbehovet ditt og informasjonsbehovet ditt som skissert i figur 3.3.



FIGUR 3.2. AIS sin metamodell [29 , s23]

I forbindelse med en arbeidsoppgave kan en spille flere roller, *f.eks man får i oppdrag å dekke en sak, og i tillegg til å skrive artikkel om saken har man også med seg et kamera for å ta bilder som illustrerer saken. Du spiller da rollen som både fotograf og journalist.* Dermed illustrerer at en en og samme arbeidsoppgave utarte seg ulikt for en fotograf, arkivar eller en journalist, og avhenger av den kunnskap og erfaring det enkelte individ som utspiller rollen innehar.

Videre skal den enkeltes behov kunne tilfredsstilles i metamodellen for et informasjonsrom noe som er definert i figur 3.3. som bygger på metamodellen i AIS prosjektet [29]. Vi ser her at metadatasett og metadatakomponenter er byttet ut for begrepsmodell og begrepsmodell komponenter. Det er nødvendig med en forklaring til modellen. En begrepsmodell er det samme som en ontologi. DC er en metadataontologi. I figur 3.4. vil DC Element Set settes inn i boblen for metadatasett og de ulike DC elementene i boblen for metadatakomponenter. DC er grunnlaget jeg tar utgangspunkt i for å se om DC som metadataontologi kan tilfredsstille de ulike brukerbehov og roller som Adresseavisen har. Dette gjøres i mappingen av formater i kap 8.

Videre forklaring på de ulike bestanddeler av metamodellen er nødvendig.

3.6.3. Hva er en ontologi.

Ontologibegrepet tillegges ulik betydning alt etter hvilket fagfelt det brukes innen. Innen filosofien betyr begrepet "eksistensens natur" eller "det å eksistere". I datavitenskapen er det innen retningene Artificial Intelligence (AI) og kunnskap og informasjonsforvaltning at begrepet er mest benyttet. Her tillegges det betydningen av "en felles forståelse innen et interesseområde" og "enhetlig ramme for å løse problemer".

Ifølge Uschold m.fl.[34] snakker vi om flere typer ontologier som kan karakteriseres ut fra tre dimensjoner:

1. *grad av formalisme*
2. *hva ontologien brukes til*
3. *hva ontologien handler om*

Grad av formalisme går på om ontologiene er svært uformell, strukturert uformell, semiformelle eller strengt formelle i henhold til språk og definisjoner.

Når det gjelder hva en ontologi kan brukes til omfatter det:

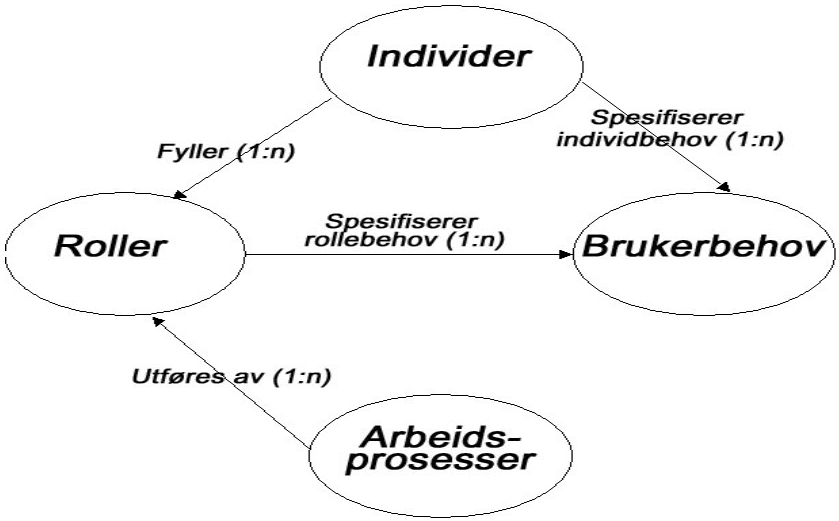
- kommunikasjon mellom mennesker
- støtte interoperabilitet mellom datasystemer, dvs ontologien brukes som utvekslingsformat

- støtte systemutviklingsprosessen, ved at den viser til formelle måter å representere prosesser og informasjonsflyt mellom prosesser på

Den sistnevnte typen av ontologier er knyttet til tema/domene, og hva ontologien handler om. Dette kan innebære hva som helst innen et emne, men det er noen domener som er mer brukt enn andre innen for denne typen ontologi. Disse domenene er kunnskapsdomener innen medisin, geografi m.m, innen emneområdet problemløsning hvor ontologi er orientert mot å løse oppgaver/problemer generelt, eller innen emneområdet språk for å representere kunnskap.

Ofte kan en ontologi være en kombinasjon av alle disse dimensjonene som nevnt ovenfor. For dette prosjektet er det DC som er metadataontologien, hvor DC blir vurdert som mulig ontologi for å beskrive Adresseavisens informasjonsressurser ut fra de brukerbehov som eksisterer i de ulike arbeidsprosesser.

Det er også ment at DC skal fungere som utvekslingsformat og interoperabilitet mellom arbeidsproseser. Jeg skal i dette prosjektet ikke implementere modellen ved Adresseavisen. Metamodellen er tatt med her for å eksemplifisere hvordan en ontologi som DC kan implementeres med utgangspunkt i en metamodell som presentert i figur 3.4., og dermed tilpasses ulike brukerbehov.



FIGUR 3.3. Brukerbehov

I tillegg vil de brukerbehov som ikke kan tilfredsstilles i DC bli tilfredsstilt i en annen beskrivelsesmodell eller ontologi. Metamodellen i figur 3.4. består av komponentene begrepsmodell, perspektiv, begrepsmodellkomponenter, informasjonsressurs og presentasjon.

3.6.4. Begrepsmodell og begrepsmodellkomponenter

En begrepsmodell betyr det samme som en ontologi som forklares i avsnitt 3.6.3. Et informasjonsrom kan inneholde flere begrepsmodeller. Begrepmodellkomponentene er byggeklosser som utgjør begrepsmodellen. Et eksempel her er DC som begrepsmodell hvor hver enkelt begrepsmodelkomponent viser til alle DC sine 15 metadataelementer, alle DC sine sub-elementer og deres definisjoner og relasjoner.

3.6.5. Perspektiv

Et perspektiv er hvordan du velger å betrakte "verden" eller "noe" fra. Perspektivet avbilder en brukermodell som tar utgangspunkt i relasjonen mellom brukeren som individ, den rolle brukeren innehar, og de interesseprofiler og brukerbehov som brukerrollen spesifiserer. Et eksempel på et perspektiv i denne sammenheng er dersom en har en arbeidsprosess hvor en skal indeksere bilder i SIFT filmarkiv. En spiller da rollen som arkivar. Da vil f.eks perspektivet være et skjermbilde som dukker opp ut fra den arbeidsoppgave og rolle du har som er spesialtilpasset nettopp for de informasjonsbehov du vil ha i denne sammenheng. Dersom du som journalist skal utføre en indekseringsoppgave vil perspektivet tilpasse seg din rolle som journalist i denne sammenheng. Et perspektiv gir derfor et individ eller en gruppe av individ sitt "syn" eller tilpasning inn til informasjonsrommet, såkalte "views" som dette omtales som. Et perspektiv kan forholde seg til en eller flere begrepsmodeller, og et perspektiv definerer et navigeringshierarki i informasjonsrommet hvor konsepter og begrepskomponenter inngår for å vise frem til informasjonsressursene. Med utgangspunkt figur 3.4. blir verdiene med DC som utgangspunkt:

Metadatasett	= DC
Metadatakomponenter	= DC komponenter

I de tilfeller hvor DC elementer ikke kan brukes i indeksering forholder perspektivet seg til et annet sett av metadata når indekseringsoppgaven skal gjennomføres. Det er altså mulig å operere med flere metadataontologier. Begrepsmodellen, f.eks DC er den konseptuelle modellen som definerer de ulike DC elementene og relasjonene mellom dem, mens et perspektiv spesifiserer hvilken informasjon fra den eller de konseptuelle modellene bruker har behov for når han eller hun skal utføre sine arbeidsprosesser.

3.6.6. Presentasjon

En presentasjon er en avbildning av et perspektiv. Presentasjonen skjer igjennom en presentasjonskanal som ikke er annet enn det vi kjenner som et dokumentvindu i Internettloser som f.eks Netscape eller Internett Explorer.

3.6.7. Informasjonsressurs

En informasjonsressurs er alt som er bærer av informasjon og som er tilgjengelig i informasjonsrommet. I et informasjonsrom hvor informasjonsressursene er kategorisert er de kategorisert i henhold til en eller flere begrepsmodeller. I Addresseavisen sitt tilfelle ville det si at dersom det var noen informasjonsressurser som ikke kunne kategoriseres i henhold til DC formatet ville de måtte forholde seg til bli kategorisert i henhold til en annen begrepsmodell. I et Intranett er en informasjonsressurs alt fra en rapport, et skjema, en tegning, en powerpoint presentasjon, videopptak, bøker, artikler osv.

3.6.8. Instansiering av metamodellen

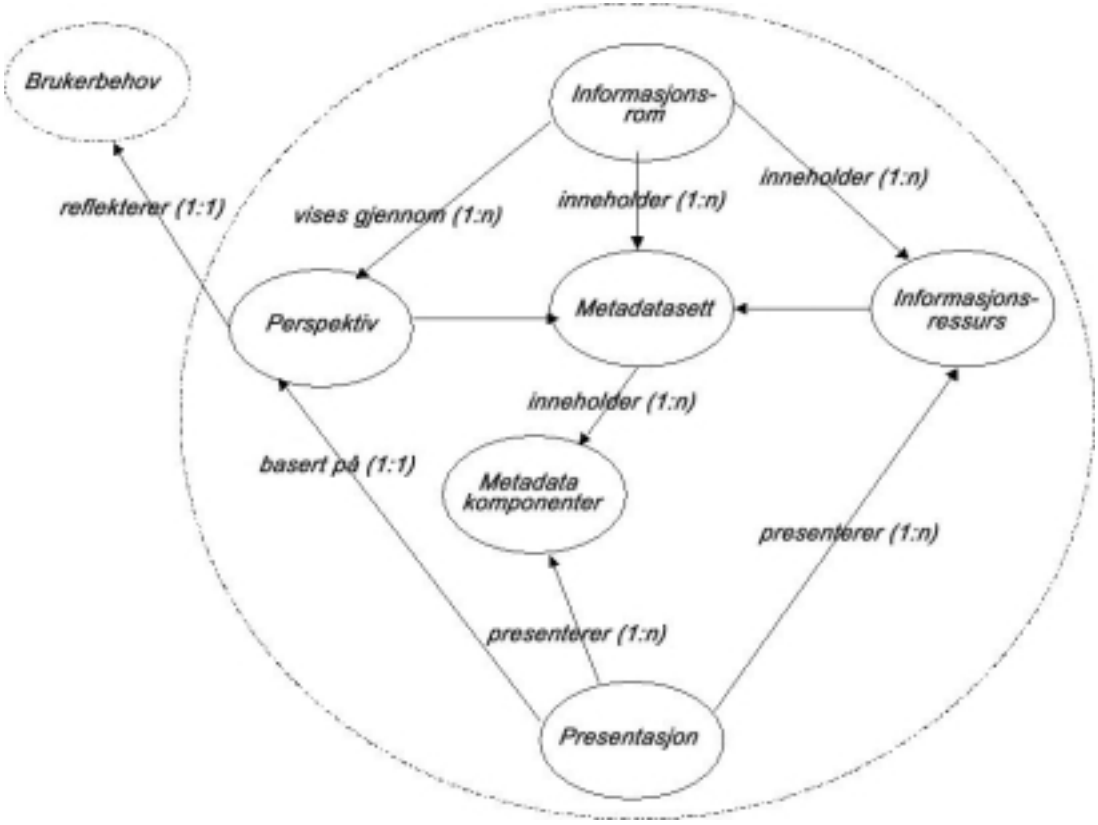
Her tar jeg utgangspunkt i metamodellen for informasjonsrommet i figur 3.4. og forklarer hvordan denne metamodellen kan instansieres ut fra en arbeisoppgave. Å instansiere en arbeidsoppgave sin "levesyklus" fra A til Å i modellen, vil være for tidkrevende her, og er heller ikke min hensikt.

Dersom man tenker seg at man har en oppgave å utføre, *f.eks at du har behov for å hente ut en del opplysninger fra SIFT databasen som du trenger til å utføre en bestilling av bilder for en kunde.* Du må da søke opp filmmappen som bildene befinner seg på i SIFT. Kunden ønsker et portrettbilde av Ole Gunnar Solskjær som var tatt i et portrettintervju i Adresseavisen for omtrent en uke tilbake. Kunden vet ikke hvem som er fotograf for bildet.

De opplysninger du har som utgangspunkt for å finne akkurat det bildet kunden vil ha, er den metadatainformasjonen du har som utgangspunkt og som vil utforme det perspektiv du får inn i informasjonsrommet. Den rolle du innehar vil også ha noe å si for utforming av perspektivet. Arbeider man ved arkivet, har man mye erfaring i bruk av SIFT og man utfører bildebestillinger ofte. Man vet derfor hvordan man skal navigere seg frem til infomasjonsressursen, og besitter annen metadatainformasjon enn f.eks enn journalist ville ha dersom vedkommende skulle utføre samme søkejobben. *F.eks her så visste ikke kunden hvem som var fotograf, men av erfaring vet f.eks arkivaren at det er "Per Hansen" som ofte har slike saker.* En vil da da ha denne metadainformasjonen å søke på som første utgangspunkt:

- Journalist: Per Hansen (DC.Creator.Agentname)
- Informasjon om at det er portrettbilde (DC.Description.Genre)
- Emneord: 1.0 Solskjær, Ole Gunnar - Fotballspiller (DC.Subject.Classification)
- Dato: Bildet er tatt ganske nylig, innenfor en uk
- Portrettintervju (som oftes er lagt i sak).

Dersom en da tenker at DC er metadatasettet som danner ontologien her vil disse 5 punkter med metadatainformasjon gjenspeiles i perspektivet ditt. Denne metadatainformasjonen kan ses på som et "metadata surrogat" til metadatasettet som en her tenker er DC. Overfører en informasjonen i punktene over til DC element som en vet dokumentet er beskrevet med, kan en søke på DC sine metadataelement for denne informasjonen som nevnt over i parantes bak punktene. Du navigerer deg frem til ressursen ut fra det utgangspunkt du her har.



FIGUR 3.4. Metamodell for informasjonsrommet

3.6.9. Kvaliteten på metadata

Vi snakker om gode metadata, dersom de er god nok til det formålet de skal tjene. Et annet begrep å bruke her er "egnethet", dvs hva er metadataene sin egnethet i forhold til det formål det skal tjene? Kvaliteten på metadata avhenger av situasjonen eller konteksten den brukes i, hvem som er brukere av metadata og hvilke type informasjonsobjekt som skal beskrives. En felles metadataontologi som er spesialtilpasset for beskrivelse av ulike informasjonsressurser basert nettopp på ulike brukerbehov og konteksten det brukes i vil bedre kvaliteten på metadata.

Kort sagt kan en si at gode metadata dekker et brukerbehov tilfredsstillende, mens dårlige metadata ikke gjør det. Det er ikke alltid at bruk av metadata er sett i sammenheng med rolle, brukerbehov og den arbeidsprosess som skal utføres.

3.7. Inkorporerte metadata

Inkorporerte metadata er metadata som er integrert i selve dokuementet ditt, sammen med koden og råteksten. *F.eks når vi har et HTML dokument kan du sette metadata sett som Dublin Core inn mellom <HEAD> ...</HEAD> tag'ene i HTML koden. Dublin Core metadataene er da inkorporert i selv dokumentet, og vil da hele tiden følge dokumentet.*

3.8. Interoperabilitet mellom metadata

De fleste bedrifter /organisasjoner bruker Ad-hoc formater eller proprietære formater som de også kalles til å beskrive sine informasjonsobjekt, noe som gjør det vanskelig å skape interoperabilitet ved integrering av heterogene kilder. Et felles kommunikasjonsformat ville kunne støtte interoperabilitet, noe som nettopp Dublin Core (se kapittel 6) formatet er ment å tjene som.

Interoperabilitet er i dag blitt mer og mer viktig ved fremveksten av Internett og Intranett. I Intranett sammenheng er det blitt utbredt og integrere informasjon fra kilder som støtter visse arbeidsprosesser, gjennom såkalte informasjonsrom, noe som krever høy grad av interoperabilitet mellom de ulike formater som benyttes. For å sette interoperabilitetsproblemet i perspektiv kan en sammenligne dette med språk og kommunikasjon, hvor engelsk har fungert

som et kjernespråk som åpner dører for samarbeid og kommunikasjon verden over, fordi man har en felles basis forståelse gjennom et verdensspråk. Har en da mennesker fra mange ulike kulturer som alle behersker engelsk går integreringen nærmest av seg selv. Likestilt med at vi trenger et felles språk for å kommunisere med hverandre på tvers av kulturer, trenger vi også et felles språk for utveksling og gjenfinning av digital informasjon i distribuerte Nettverk som Internett.

Ulike beskrivelsesmodeller som benyttes på Internett fører også til konflikter i søkeprosessen, fordi de ulike kildene man søker i ikke har semantisk interoperabilitet. Ideen her er at en skal utvikle et felles format som skal fungere som en kjerne for katalogisering av informasjon, noe som ville støtte utveksling og gjenfinning av informasjon uavhengig av medium. Dette er Dublin Core formatet ment å tjene som og dette blir nærmere omtalt i kapittel 6. Ved å utvikle et felles format her vil en oppnå semantisk interoperabilitet som gjør det mulig å søke etter informasjon hvor som helst på tvers av fagdisipliner.

Vi har to nivåer av Interoperabilitet:

1. utveksling
2. søking

For at to applikasjoner skal utveksle metadata effektivt, må metadataene ha den samme semantikken, og dele en felles struktur og syntaks. Et eksempel på effektiv utveksling her er ved bruk av MARC fomatet som nettopp er brukt for utveksling av bibliografiske poster mellom bibliotek. MARC formatet presenteres i kapittel 6.

Ved søking kreves det et minimum av ordinær semantikk og i tillegg en protokoll for å ta imot og gi respons til forespørsler i form av queries. f.eks dersom en klientapplikasjon søker i en DC repository applikasjon etter ett nøkkelord er det tilstrekkelig at de to applikasjonene deler semantikken for DC.Subject, og har som har måte å overføre dette på gjennom en passende protokoll.

Dersom en benytter DC som "minste felles multiplum" for indeksering av sine informasjonsobjekt og informasjonsressurser, kan en lettere integrere sitt lokale "digitale bibliotek" med eksterne "digitale bibliotek" eller informasjonsrom. Dette åpner nye dører for støtte i arbeidsprosesser og gjenfinning og gjenbruk av informasjon.

Dersom et repository bruker Dewey Decimal Classification (DDC) for å klassifisere et dokument, og klientapplikasjonen "forstår" DDC så er muligheten for å få til en nyttig transaksjon betydelig økt.

Applikasjonen vil være istand til å dra fordel av DDC som en browsing struktur i tillegg, og dermed øke effektiviteten ved søk. Dersom en klient applikasjon er ignorerende til DDC kan et

nøkkelord søk fremdeles lykkes, selv om bruker ikke har noen spesiell kunnskap eller ekspertise på DDC, men vil være mye mer tilfeldig styrt /sukksessraten her vil ikke kunne forutsies.

Applikasjoner som bruker ulike skjema for å kode DC.Subject kan fremdeles operere sammen på søkenivå, men det er lite trolig at de kan utveksle metadata på en effektiv måte.

3.9. Konflikter mellom metadata

Det er ulike konflikter som kan oppstå i forhold til metadata. Vi har:

- Navnekonflikter
- Granularitetskonflikter
- Syntakskonflikter
- Strukturelle konflikter

3.9.1. Navnekonflikter

Her kan ulike metadataelementer ha samme type data men ikke ha samme navn, og omvendt. Vi har konflikter på synonymer og homonymer. *F.eks på synonymer når du har metadataelement som heter Creator som i DC, eller forfatter og journalist er de begge metadataelement for opphavspersonen av dokumentet.* Når vi har konflikter på homonymer, har vi samme metadataelement men ulikt innhold. *F.eks er datoelementet et homonym, det inneholder en dato, men en kan bruke feltet forskjellig dersom det ikke er gitt hvordan det skal tolkes. Noen bruker dato feltet til utgivelsesdato, andre til publiseringsdato osv.* Konflikt på homonymer kan også oppstå når du har emneord, emneord kan være fire emneord tatt vilkårlig ut i luften, men de kan også være fra et kontrollert vokabular som en Thesaurus eller en egendefinert liste som følges. Dublin Core har løst noen konflikter her ved å bruke sub-element og attributtet SCHEME for å spesifisere innholdet ytterligere i metadataelementet, f.eks har dato elementet fått sub-element som DC.Date.Issued, DC.Date.Available osv.

3.9.2. Granularitetskonflikter

Konflikter som oppstår her skjer når metadataformatene har ulik detaljeringsgrad, dvs at ikke bare har et element til å representere et innhold, men ulike felt som tar seg av det samme. MARC formatet er mye mer "finkornet" på dette enn f.eks Dublin Core formatet. Eksemepl på detaljeringsgrad er at du kan ha ulike attributter for primærforfattere og biforfattere.

3.9.3. Syntakskonflikter:

Dette er konflikter som oppstår når de dataene som oppbevares i et metadataelement kan representeres på ulike måter, f.eks i forhold til personnavn og dato. Vi kan representere et navn på formen:

- Per Hansen
- Hansen, Per
- P. Hansen

osv. I noen kulturer er det også helt vanlig at navnet leses fra høyre til venstre og dermed vil etternavnet stå først uten komma, f.eks Hansen Per. På samme måte kan du representere dato ved å skrive

1. *dag - måned - år*
2. *år - måned - dag*
3. osv.....

Her bruker man også ulike operatorer, *f.eks enten punktum, komma, tekst eller tegnene / og - for å skrive datoen.* Ulike varianter:

- 18.04.2000
- 2000.04.18.
- 18/04/2000
- 2000/04/18
- 18-04-2000
- 2000-04-18
- Fredag, 18 April 2000
- 18-04-00 og 3-12-99, hvor år er skrevet med to siffer.

Så langt det er mulig bør en her bruke standarder for å representere innholdet, men det er allikevel ingen garanti for at andre bruker samme syntaks. Spesielt når det gjelder Ad-hoc formater trenger det ikke alltid være samsvar på dette innenfor de ulike metadataformater som brukes i bedriften heller.

3.9.4. Strukturelle konflikter:

Strukturelle konflikter oppstår ved utstrakt bruk av sub-element. I DC formatet er det foreslått et utvalg av sub-element pr. april i år [58], men dette er bare et forslag fra DCMI og er ikke ment å passe for alle applikasjoner. Dette har ført til at de ulike prosjekt tilknyttet DCMI har brukt ulike varianter for sub-element for elementet DC.Relation tilpasset de ulike behov. Disse sub-elementene er ikke direkte kompatible og de ulike metadataverktøy som er utviklet for DC tar liten hensyn til sub-elementer på nåværende tidspunkt.

Kapittel 4

Informasjonssøkingsprosessen

"Any piece of knowledge I acquire today has a value at thsi moment exatly proportional to my skill to deal with it"
- Mark Van Doren, *Liberal Education.*

4.1. Introduksjon

Dette kapittelet er ment som en innføring i hvordan informasjonsgjenfinning og informassjonssøkingsprosessen henger sammen. I den sammeheng videreføre kunnskap om metadata som en har tilegnet seg fra kapittel 3 til å forstå hvordan metadata kan effektivisere informasjonssøkingsprosessen og bedre gjenfinning av relevant informasjon.

4.2. Informasjonsgjenfinning?

Informasjonsgjenfinning har eksistert nærmest siden tidenes morgen. Før informasjon ble nedskrevet og trykt, var det den verbale kommunikasjonsmåten som var gjeldende, og måten folk fikk tilgang til informasjon på. Da papiret og boktrykkerkunsten var en realitet åpnet muligheten seg for å massepublisere informasjon og kunnskap. For å ta være på denne økende mengden informasjon som vokste frem, kom bibliotekene på banen. De fleste av oss forbinder bibliotek med "tilgang til informasjon", og der vi kunne få hjelp til å finne den informasjonen vi søkte. Men etterhvert som mengden av informasjon har økt, har det blitt stadig vanskeligere å finne frem i denne mengden. Da det ble mulig å lagre informasjon digitalt åpnet det seg nye muligheter for raskere tilgang og bedre gjenfinning av informasjon. Etter Internett revulosjonen er mendgden informasjon blitt så stor, at utfordringer når det gjelder å gjenfinne og sikre informasjon på Internett står i kø.

I dag stiller vi høyere krav til effektivitet enn tidligere, og det er ikke lenger nok å finne relevant informasjon, vi må også finne den til rett tid,

ellers trenger vi den ikke. Derfor er det også begrenset hvor lang tid vi ønsker å bruke på å finne informasjonen.

Biblioteket sitter med den eldste kunnskap for gjenfinning av informasjon, da det eksisterte lenge før mye av teknologien som benyttes i dag var påtenkt. Biblioteket hadde som hensikt å støtte organisert gjenfinning av informasjon. Hjelpemidler de brukte for å katalogisere informasjon, var å dele dem opp i indeksgrupper etter forfatter, type informasjon og emneord slik at en kunne søke etter informasjonen på flere måter. De indekser vi benytter i dag er veldig lik de som ble brukt på 1800 tallet. Gamle gjenfinningsmetoder er fremdeles levedyktige i kombinasjon med moderne teknologi, og er fremdeles en av de viktigste redskapene som brukes for å gjenfinne informasjon tiltross for over 200 år med avansert teknologi utvikling. Ny teknologi har imidlertid gjort det mulig å få raskere tilgang til informasjon, fordi indeksene er blitt digitalisert, og en kan søke i flere indekser samtidig, noe som er tidsbesparende. I tillegg har en også mulighet til å søke i fulltekst der hele dokumenter er tilgjengelig via datamaskinen og en ikke trenger å lete seg frem til en bok på hyllen. I det stadige voksende informasjonsparadigmet hvor vi “alle” er lesere og utgivere av informasjon via Intranett og Internett, stiller dette også større krav til informasjonens kvalitet for gjenbruk og pålitelighet og at det faktisk er sannhet i den informasjonen vi leser. Dette er et området som blir stadig viktigere å ivareta. Ved fremveksten av flere arenaer som "digitale bibliotek" og informasjonsrom tilgjengelig via Intranett og Internett vil dette åpne for mer strukturert og bedre gjenfinning av informasjon. Informasjon og kunnskap er noe mennesker alltid vil søke, og vi trenger hjelp til å sortere den ut når "informasjons-bulimien" herjer som verst Det er faktisk grenser for hva vi klarer å fordøye av informasjon. Fagfeltet Informasjonsgjenfinning vil hele tiden sette fokus på dette, og forske på nye løsninger for å bedre informasjonsgjenfinningen i takt med de behov som oppstår og utvikling av ny teknologi. Teknologiutviklingen vil følge vårt informasjonsbehov som "hånd i hanske" i tiden fremover, for spørsmål om HVOR og NÅR en kan finne kunnskap blir stadig viktigere i vårt postmoderne[1] samfunn.

Med utviklingen av Internett, og stadig flere digitaliseringer av bibliotek er det behov for å finne nye løsninger og metoder for hvordan informasjonen bedre kan struktureres med den hensikt at man nettopp skal finne den igjen. I og med at vi alle er blitt brukere og forfattere av informasjon via Internett, er ikke all informasjon som vi finner på nettet kvalitetssikret. Vi trenger noen som kan hjelpe oss til å sortere ut den informasjon som er uviktig for oss.

Løsningene for gjenfinning er mange og avhenger av ulike faktorer, som i hvilken kontekst søk foretas, størrelsen på samlingene og ønsket mål ved søk.

.

1. Postmoderne = Et samfunn som preges av utstrakt bruk av informasjonsteknologi

Når en lever i et samfunn der endringer skjer så raskt at en føler at "alt flyter", har gamle gjenfinningsmetoder som bruk av indeks og søking på nøkkelord fortsatt "livets rett". Informasjonsgjenfinning benytter seg altså av gammel og ny teknologi. Tiltross for over 200 års utvikling innen teknologi trenger vi stadig kraftigere verktøy for å behandle de store mengdene med informasjon i dagens Informasjonssamfunn.

Når det finnes så mye informasjon tilgjengelig oppstår forvirringen og usikkerheten om påliteligheten i dokumentenes innhold. En vet ikke helt om en skal stole på informasjonen som strømmer over en. Vi trenger metoder som kan kvalitetssikre kildene og gi oss bedre tilgang til informasjon i distribuerte miljø som Intranett og Internett.

Jeg skal her gi en presentasjon av de gjenfinningsalgoritmer og aspekter ved informasjonsgjennfinning som ses som relevant i forhold det som er fokus for denne avhandlingen.

4.2.1. Informasjonsgjenfinnings metoder

Information retrieval (IR) systemer er utviklet for å støtte gjenfinning av informasjon. Mange universitet, bedrifter og bibliotek bruer IR systemer for å gi deg som bruker distribuert tilgang til bøker, tidsskrifter og andre dokumenter raskere og mer effektivt. Dette skal gi en anledning til selv å kunne låne dokumenter når en selv har tid og anledning. Et IR system fungerer på den måten at en gir inn en forespørsel til systemet i form av et query, som gjenspeiler den informasjons man søker. Deretter behandles forespørselen din mot det som finnes av dokumenter i databasesystemet. Deretter gis man en tilbakemelding i form av en treffliste av dokumenter som "matcher" den forespørsel man har gitt. Alle disse prosesser er en del av det som kalles en infomrasjonssøkingsprosess. Det finnes flere ulike måter å utføre query’s på alt etter hvilke gjenfinningsmetoder som benyttes, men når man gir inn en forespørsel til et system bruker man enten fritekstsøk eller metadatasøk i kombinasjon med boolske operatorer. Den gjenfinningsmetode som brukes i IR systemet vil vise seg i tilbakemeldingen man får fra systemet, hvor noen vurderer relevans i dokumentene etter finnes eller ikke finnes, hyppighet av termer i dokumentet, vektlegging av ulike termer, bruk av metadata m.m. Ofte bruker en også en kombinasjon av ulike gjenfinningsmetoder.

4.2.1.1. Boolske System

Et boolsk IR system kombinerer termer i query ved hjelp av ulike operatorer. Veldig mange søkesystemer benytter seg av boolsk logikk. Vi har følgende boolske operatorer:

- AND / OG
- OR / ELLER
- NOT / IKKE

Dersom man har mengdene A og B, og kombinere disse med OG slik:

A AND B / A OG B = Det mengden i A som er felles med mengden i B.

Dersom man har mengden A og B, og kombinerer dette med ELLER slik:

A ELLER B/ A OR B = Det mengden A og det mengde B utgjør til sammen.

Dersom man har mengden A og mengden B og kombinerer disse med operatoren NOT slik:

A NOT B / A IKKE B = Det som mengden A har etter at mengen B er trukket fra A.

I et boolsk query vil en sjekke om søketermene stemmer overens med de termer som finnes i dokumentet, og ser ikke på hvor ofte de søketermene forekommer i dokumentet, men om det finnes eller ikke, og så må man selv vurdere relevansen av dokumentenes innhold. For boolsk logikk er spørsmålet "TO BE in the document, or NOT TO BE in the document".

4.2.1.2. Vektor Systemer

Her er dokumenter representert som en vektor før det lagres i et repository. For hvert nytt dokument som legges til databasen blir det laget en vektor for dokumentet som bestemmer dets relevans i forhold til de andre dokumentene. Dokumentene blir her klassifisert i klynger, dvs clustering i IR, hvor hver klynge har en grenseverdi som representerer likhet mellom dokumenter. Fordelen her er at dokumenter kan være klassifisert i flere klynger, noe som kan gjøre det lettere å finne igjen dokumenter som ligger i gråsonen mellom flere kategorier. Vektoren utformes etter en liste av termer, hvor listen blir sammenlignet med om disse termene finnes i dokumentet eller ikke, og relevans blir bestemt ut fra hyppighet eller spesiell vektlegging av disse termene. Hensikten med vektormodellen er å bedre forholdet mellom presisjon og relevant tilbakemelding, omtalt som precision og recall som presentert i avsnitt 4.2.2.1. Når man da søker på en søketerm, vil man få opp de dokumentene med høyest relevans i løpende rekkefølge. Søkemotorer som Altavista benytter ikke vektor system, siden å utforme vektorer som representant for dokumenter blir vanskeligere desto større dokumentmengden er,

og man skal bestemme relevansen i forhold til alle andre dokumenter i basen. Der er også derfor at Altavista ikke er særlig pålitelig med hensyn til relevans i tilbakemeldingen en som søker får.

4.2.2. Relevance feedback

Når vi snakker om former for tilbakemelding snakker vi om Relevance Feedback, og et begrep som brukes her er Relevant Feedback. Annen feedback er uinteressant. Disse former for tilbakemeldinger er nødvendig for å bestemme hvor at et informasjonsobjekt gjenfinningssystem vil være effektiv. Hovedsaklig går relevance feedback ut på at vektlegging av ulike query-termer

4.2.2.1. Recall / Precision

Recall = Dette er antallet av relevante informasjonsobjekt som er gjenfunnet for et gitt query, i forhold til det totale antallet av relevante dokument som finnes i databasen for det query som faktisk er gitt. Mer spesifikt vil det si:

Precision = Dette er forholdet mellom det antallet av relevante informasjonsobjekt og antallet av informasjonsobjekt totalt i trefflisten din etter et søk.

Forholdet mellom recall og precision kan vi illustrere ved en matematisk formel. Vi setter opp følgende verdier:

N = Totalt antall av informasjonsobjekt i et system, dvs antall enheter lagret i databasen.

T = Antall informasjonsobjekt som er gjenfunnet, dvs de dokumenter man får opp i en treffliste etter at et query er gitt.

D = Antall relevante dokumenter i databasen. Dvs de informasjonsobjekt som er relevant for akkurat det query man har gitt inn.

R = Antall dokumenter som er gjenfunnet og som er relevante for det query som er gitt.

Vi får da formlene:

Recall = R/D

Presisjon= R/T

Dersom man har et gitt query som skal gi treff på alle informasjonsobjekt innen et spesielt emne, og vi har en database som består av 10 dokumenter. Siden man har lagt inn alle dokumentene selv vet en at 3 av 10 dokumenter er relevante for queryet som er gitt. Har man en database på flere tusen eller millioner informasjonsobjekt, vil det være vanskelig å ha den totale oversikten over om det faktisk ligger noen relevante dokumenter igjen i databasen som en ikke får opp i treff listen. Recall kan derfor være vanskelig å bestemme, da det er en urealistisk å gjennomløpe databaser av denne størrelsesorden for å finne ut hvor mange informasjonsobjekt man faktisk skulle ha fått som recall.

F.eks: En har en database som består av 20 informasjonsobjekt totalt. Man gjør så et søk mot Databasen ved ett gitt query, og D for dette query er 5. Når en gjør ett søk mot basen for dette query får man opp 15 informasjonsobjekt, som da blir T. Av disse informasjonsobjekt i trefflisten din er bare 4 relevante for queryet.(ett relevant dokument ligger da igjen i databasen).

Recall= 4/5=0,8

Precision= 4/15= 0, 26

4.2.3. Tesaurus

En tesaurus er et verktøy man bruker for vokabular kontroll. Her kan man velge mellom en fast liste av termer som en kan bruke ved indeksering, som kan hjelpe til å bedre kvaliteten på gjenfinning. Det er vanlig å ha tesauruser som dekker ulike fagfelt , *f.eks egen tesaurus for arkitektur, realfag som matematikk, fysikk og informatikk eller tesaurus tilpasset indeksering av grafisk materiell.* Når man har vokabular kontroll unngår man at noen informasjonsobjekt blir indeksert ved hjelp av termer som ikke er ønskelig, og en får mindre sprik i indekseringen. De som indekserer må forholde seg til et fast sett av emneord og kan ikke finne på egne emneord etter behov. Å bruke en tesaurus vil bedre gjenfinning ved søk, (dersom en kjenner emneordene som brukes) fordi individuell "språkrikdom" ikke tillates. I prinsippet vil dette *f.eks si at dersom man søker på "innbrudd", ville en ikke få treff dersom alle dokumenter i basen som omhandlet slike forbrytelser var indeksert etter emneordet "ran".*

En tesurus består av hovedsaklig subjekt og subjektfraser. Det er ikke alltid like lett og finne gode emneord som beskriver et informasjonsobjekt godt nok og da kan en tesaurus fungerer som en "oppskrift" på god indekseringsskikk. Når hvert fagfelt har sin tesaurus vet en også hvilke termer en skal søke på ved utforming av et query.

En tesaurus kan genereres automatisk enten ved å benytte samlinger av dokumenter, eller ved å browse allerede eksisterende tesauruser.

Fordelene med en tesaurus er at en har en fast liste av emneord som kan brukes som utgangspunkt både for indeksering og søk, slik at indeksering og søk blir mer konsistent. Hele hensikten med en tesaurus er å oppnå kontroll over bruk av emneord, dvs f.eks for å beskrive en hendelse måtte jeg velge fra en liste med emneord, og akkurat disse samme emneordene ville jeg hatt tilgang til ved søk. En tesaurus inneholder ikke alle mulige synonymer for emneord for et fagfelt men velger ut de termer for emnet som er foretrukket å bruke. Dette kan eksemplifiseres med at *f.eks emneordet "Informasjonsvitenskap" benyttes istedet for termen "Datavitenskap"*.

4.2.4. Automatisk indeksering

Automatisk indeksering er prosessen hvor algoritmer undersøker informasjonsobjekt for å generere en liste av indekstermer. Første steg i denne prosessen er leksikalsk analyse, som vil si å ta utgangspunkt i queryprosessen, dele queryet fra hverandre og sette det sammen til en strøm at tegn og deretter ord som kan si noe om hva vedkommende søker. De termene som kommer frem i leksikalsk analyse er kandidater til en indeksliste, og kan også bli lagt til indeksen dersom den ikke finnes fra før. Query prosessering vil si å analysere queryet og sammenligne termer i query med termer i den genererte indeksen for å finne relevante informasjonsobjekt. I automatisk indeksering må man ha en form for siling av termer som ikke er relevante, og det gjøres ved å gjøre de utvalgte termer fra queryanalysen gjennom en stoppordliste (som man også kan generere automatisk ut fra bestemte regler). I stoppordlisten ligger termer som ikke er egnet for indeksering.

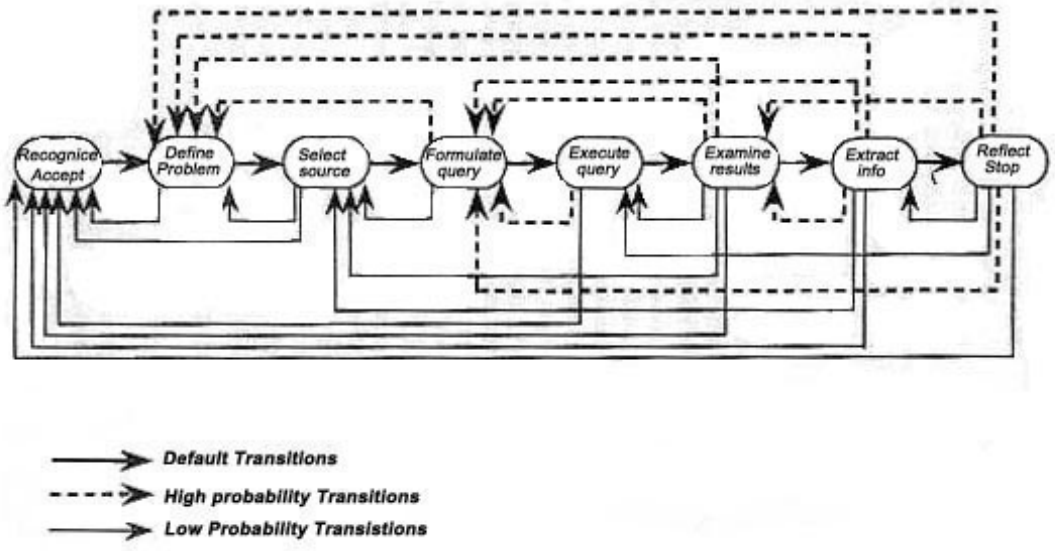
Svakheten med automatisk indeksering er at det er vanskelig å vurdere hva som skal oppfattes som en term i et query og hvilke indekstermer som skal være tilgjengelig for de informasjonsobjekt som allerede finnes i databasen. Manuell indeksering for hvert dokument som legges inn vil gi et mye bedre bilde av dokumentets relevans enn hva som er tilfelle med automatisk indeksering. Etter min mening kan ikke en indeksering bli optimal uten at human resources er innblandet i indekseringsprosesen. Selv om man velger ut alle subjekter i et dokument for å generere en indeks for databasen, har en ingen garanti for at disse subjektivene er

representativt for innholdet i dokumentet. Språket er en kompleks helhet og med mye bruk av metaforer og anne språkrikdom, kan ikke en maksin vurdere dets fullstendige innhold.

4.3. Informassjonssøkingsprosessen

Informasjonssøkingsprosessen er en komplisert sak, og avhenger av hvor godt bruker er i stand til å definere hva vedkommende ønsker å finne, og formulere dette i et søkeuttrykk som databasesystemet forstår. Informasjonssøkeprosessen din er avhengig av flere faktorer noe jeg vil komme mer innpå i dette avsnittet. Marchionini har en modell [26] for informasjonssøkeprosessen som er representativt for hvilke faser som inngår i søkeprosessen og hvordan dette utarter seg som representert i figur 4.1 nedenfor. En søkeprosess består av både tankearbeid og det er ikke alle fasene som er direkte innvolvert i interaksjon med en PC. Modellen til Marchionini er ment å være generell og passer også inn i et mønster for informasjonssøkeprosesser som ikke involverer bruk av datamaskin. Dette eksemplifiseres i de ulike fasene som blir nærmere presentert nedenfor.

Vi ser videre av modellen at en beveger seg mellom fasene i modellen, og at informasjonsprosessen kan være en uendelig og komplisert prosess. De stiplete linjene viser veiene som det er høy sannsynlighet for at en må bevege seg etter i de fleste informasjonssøkingsprosesser, mens de uthevete linjene viser veier en må gå i noen tilfeller.



FIGUR 4.1. Information Seeking Process [26 , s50]

Marchioninis modell består av 8 faser. Disse blir delt inn i 3 grupper:

Tabell 4.1. Kategori inndeling av Marchioninis 8 faser i informasjonssøkingsprosessen

Forståelse av Problemet	Planlegging og utførelse	Evaluering og bruk
Kjenkjenne problemet (fase 1)	Velge kilde/søkesystem (fase 3)	Undersøke resultaet (fase 6)
Akseptere problemet	Formulere forespørsel/ Bestemme start punkt (fase 4)	Trekke ut informasjon (fase 7)
Definere problemet (fase 2)	Utføre forespørsel (fase 5)	Reflektere over /Gjenta eller stoppe søk (fase 8)
	Undersøke resultatet (fase 6)	

Her blir de ulike fasene nærmere presentert:

Fase 1: Gjenkjenne/Akseptere et problem/behov.

I denne første fasen blir en gjort oppmerksom på hva problemet er, dvs at et behov for informasjon oppstår rundt et problemområde. Dette behovet kan være internt eller eksternt motivert. Internt motivert kan være at man er nysgjerrig på detaljer omkring et spørsmål som oppstår i form av en tanke eller ide, eller det bare er område som en interesser seg for. Eksternt motivert betyr *f.eks at man får en oppgave av en lærer eller har oppgaver i forbindelse med en jobbsituasjn som en må utføre.*

Fase 2: Definere og forstå problemet.

I denne fasen prøver en å forstå omfanget av hva problemet faktisk er. Dersom det er et prosjekt eller lignende som skal gjenommføres er det i denne fasen man skriver et problembeskrivelse av problemområdet, og mulige måter å løse problemet på. Dette er en kritisk fase i informasjonssøkeprosessen, fordi forståelse av problemområdet er avgjørende for utfallet av søkeprosessen, og for at man skal forstå problemområdet avhenger det av din kunnskap og erfaring på området og det en faktisk skal finne svar på. I denne fasen kan en si at man lager et "rammeverk" for problemområdet ditt.

Fase 3: Velge en kilde/Søkesystem

Her skal man velge en kilde til informasjon hvor en kan finne den informasjon man søker. Ofte har man et hav av kilder å velge imellom. Også her vil en kunne dra nytte av tidligere erfaringer til kilder og søkesystemer en allerede kjenner, særlig dersom man kjenner til kilder som en har brukt tidligere og som gir svar på det man søker. *F.eks dersom man skal finne når Knut Hamsun var født og døde og man vet at dette er det første en finner på oppslag på forfatteren i et Leksikon, er det mest sannsynlig at man velger denne kilden fremfor andre kilder. Når man da vet at leksikonet gir svaret, er det mest sannsynlig at man velger denne kilden fremfor å søke i andre kilder som Internett og SIFT, hvor man er usikker på tilbakemeldingen en får.* I praksis velger en den kilden en kjenner best, dersom man har den tilgjengelig. Kunnskap om kilder omkring problemområdet er nødvendig, og vet man ikke kilden selv må en kjenne en mellomkilde *f.eks kollega eller medstudent* som kan henvise deg videre til den rette kilden. Marchionini [26] sier at det er velkjent at en ofte fortrekker kilder som menneskelige ressuser før formelle ressurser som leksikon og databasesystemer. Dette kan igjen være beslutninger som tas basert på vurderinger som tid, arbeidsmengde m.m. Det at en fortrekker menneskelige ressurser fremfor databasesystemer kan være påvirket av flere ting. Menneskelige ressurser kan ofte være mindre tidkrevende enn å benytte et databasesystem. Vi velger kilder etter "minste motstands vei" strategien. Marchinini refererer selv til undersøkelser som han har gjort på High-

School elever hvor han fant ut at elevene valgte leksikon og bøker for å få videre referanse til kilder som kunne hjelpe dem videre i informasjonssøkeprosessen.

Fase 4: Formulere en forespørsel / (query)

Å formulere en forespørsel til et informasjonssystem involverer å matche din egen forståelsa av oppgaven til informasjonssystemets forståelse. For å gjøre dette må det være en felles måte å kommunisere på som begge forstår gjennom det som omtales som et queryspråk. I denne fasen må det skje en mapping med hensyn på semantikk og hendelser (Action). Semantikk har med syntaks i formulering av queryet å gjøre, dvs qyeryspråket, som kan være *f.eks finn Hansen i fotograf eller "finn Hansen og (internett i bilen) ".* Ulike databasesystemer har ikke noen standarisert semantikk, og andre kan bruke helt andre måter for å uttrykke det samme. I tillegg må din forståelse matche de hendelsene som datasystemet er istand til å tolke, *f.eks hvor mange ord er det datamaskinen forstår, har den ordbok innebygd osv.* Kjenner man systemets totale vokabular vil man oppnå bedre effektivitet.

Å formulere et query til datamaskinen er ikke alltid like enkelt da datamaskinen ikke på langt nær kan spille på lik streng med menneskets forståelse av problemet og er avhengig av å bli matet med fullstendige opplysninger for å kunne gi noen som helst tilbakemelding. Alt etter hvilket system man bruker er det ofte at de hendelser datamaskinen kan forstå er svært begrenset og dermed ikke like lett å uttrykke en forespørsel innenfor disse begrensningene. Den kunnskap en har på det området man søker informasjon om vil også påvirke formuleirng av en forespørsel, da det er lettere å velge ut søketermer som representerer et tema man har mye kunnskaper om.

Ofte tilbyr søkesystemer hjelpemidler for å lette denne prosessen med å formulere en forespørsel, *f.eks ved bruk pull-down menues, emneindekser m.m.* Ved søk i emneindeks unngår en å søke på termer som ikke er brukt ved katalogisering Et eksempel på dette er at de dokumenter i basen som som kan katalogiseres både under emnet "Informasjonsvitenskap" og "Datavitenskap", bare blir katalogisert etter en av disse emneordene. Dersom "informasjonsvitenskap" blir valgt vil en ikke få treff ved søk på emneordet "Informasjonsvitenskap".

Bruk av operatorere og tegn kan også være problematisk. I web-grensesnitt har man ofte et vindu tilgjengelig for søk, hvor en kan bruke boolske operatorere. Søkesystem som Altavista forstår bare den engelske betegnelsen for boolske operatorene som "AND" og "OR", mens i andre system som SIFT kan man de norske betegnelsene "OG" og "ELLER" for de boolske operatorene.

Problemstillinger som nevt ovenfor viser at det kreves en del forkunnskaper og erfaringer med bruk av søkesystemer før man kan ta det effektivt i bruk.

Fase 5: Utføre en forespørsel /query

Denne fasen er her valgt å settes inn som egen fase, men den henger nært sammen med fase 4. I fase 4 legger en i det å formulere ett søkeuttrykk som en konseptuell prosess, der søkeuttrykket blir til i hodet ditt. Deretter tar en søkeuttrykket med seg til fase 5, hvor man setter seg til datamaskinen , skriver inn forespørselen som systemet forstår, og trykker SEND knappen for å sende forespørselen til datamaskinen. Deretter er det bare å vente på tilbakemelding. Utførelse av forespørselen (query'et) er her knyttet til de fysiske handlingene dine, som å skrive inn søkeuttrykket og trykke SEND eller ENTER knappen. Når en henter boken på hyllen og slår opp på innholdsfortegnelsen er dette oppslaget å betrakte som en forespørsel, mens i fase 4 tenkte man ut at dette skulle bli forespørselen. I realiteten foregår ikke dette så detaljert og med så fine grenser. Allikevel har Marchionini valgt å atskille disse to fasene for å vise at det er visse forskjeller her. I praksis kan man sitte å formulere et query mens man skriver på datamaskinen, men da er man i fase 5 og ikke i fase 4, i og med at fase 4 og 5 her skiller mellom tanke og fysiske handlinger. Dette vil si at man hele tiden mens man formulerer og utfører en forespørsel befinner seg i fase 4 og 5.

Fase 6: Undersøke resultatet

Når man har sendt avgårde en forespørsel til datasystemet får man tilbakemelding på forespøreseln din i form av en nummerert treffliste som representerer de dokumenter som er funnet. Hver representasjon i denne listen kalles for et dokument-surrogat av det dokumentet det representerer. Denne listen gjennomleses for å se om man har funnet det en er ute etter og er ikke dette tilfellet går man tilbake til fase 4, og gjentar prosess 5 og 6. Her vil omfanget av problemet og brukerens personlige informasjonsinfrastruktur bestemme informasjonssøkerens forventning om hvor mange treff vedkommende trenger for å fullføre en oppgave, selv om disse forvent- ningene ofte endrer seg ettersom informasjonsøkingen utvikler seg. Når man søker etter en spesifikk tittel for et dokument foventer man som regel 0 eller 1 treff, og når man søker på et spesifikk tema forventer man som oftes 0, 1 eller mange treff. Når en da forventer å få et lite utvalg av dokument for f.eks søk på tittel, kan en bli ganske overrasket når en får mange hundre treff. *F.eks dersom man søker på en fullstendig tittel som man vet finnes som f.eks "Information management systems" og får mange treff, har en kanskje ikke tenkt over at dette er en generell tittel som kan være tittel for mange flere dokumenter. En tittel er i utgangspunktet ikke unik for ethvert dokument.*

Fase 7: Trekke ut informasjon

I denne fasen har man funnet frem relevante dokumenter fra den trefflisten en ble servert i fase 5 og starter nå arbeidet med å trekke ut informasjonen man trenger til et spesifikt formål. Her kan et dokument man har funnet være relevant innenfor det området en søker etter, men etter nærmere studier av trefflisten finner en at dokumentene ikke dekker akkurat det man trenger innenfor dette området. En oppdager at en trenger å detaljspesifisere søket enda mer, for å se om en finner den informasjonen innenfor dette området som en savner. F.eks at en søker opp informasjon om Knut Hamsun og ønsker å vite detaljer rundt hans fødsel, men finner bare fødselsstedet til Hamsun og ikke fødselsår. En må da gå tilbake og må dermed gjennomløpe flere dokumenter, eventuelt gå tilbake til fase 4 og 5 for å formulere og utføre en ny forespørsel. Man beveger seg frem og tilbake mellom fasene 4,5, 6 og 7. Å trekke ut informasjon av et dokument man har funnet innebærer aktiviteteter som å lese, scanne, liste ut, klassifisere, ta utskrift (kopiere) og lagre informasjonen.

Fase 8: Reflektere/Gjenta / Avslutte

En informasjonssøkingsprosess blir sjelden avsluttet ved å bare gjennomløpe hver fase en gang. Ofte tjener siste fase som en tilbakemelding for videre utforming av en forespørsel til databasesystemet. Man kan velge om man vil gjenta formulering av forespørsel eller avslutte informasjonssøkingsprosessen. De valg man tar, er avhengig av hvilken informasjon man har funnet og den informasjon man ønsker å finne innenfor et område. Ofte kan man finne deler av informasjonen, men velger å avslutte før man har funnet alt det en trenger. Andre ganger trenger man ikke finne noe relevant selv etter flere forsøk.

En kan da velge andre kilder, eller velge å avslutte informasjonssøkingsprosessen helt avhengig av

- hvor lang tid man har brukt
- problemer underveis
- motivasjon for å fortsette
- kunnskap om oppgaven
- informasjonsøkingskompetanse
- datakunnskaper etc, etc....

Siden en informasjonssøkingsprosess ofte kan forsette i det uendelige alt etter hva man søker etter, er det helt naturlig at en for hvert søk reflektererer over resultatet, og vurderer det opp mot den innsats en er villig til å inve-stere for å finne informasjonen. Ofte viser det seg i undersøkelser at det tar for langt tid å finne det man søker etter som er årsaken til at en gir opp søket. Det at det tar lang tid vil også kunne avhenge av faktorer som individuell erfaring, kunnskaper om systemet for å formulere forespørsler, alder, kunnskap om problemområdet osv.

Som bruker i informasjonssøkingsprosessen er fasene 4 , 5 og 6 ofte besøkt. Det går i en runddans frem og tilbake mellom disse fasene helt til man oppnår det ønskede ønskede resultat. Fasene der man velger kilde, formulerer forespørselen, utfører førespørselen, undersøker resultatet og trekker ut resulatet er ofte i interaksjon med datamaskinen når man bruker et Online søkesystem.

4.4. Fritekstsøk

I fritekstsøk søker man på en term, flere termer eller kombinasjon av termer ved å benytte ulike boolske operatorer i tillegg til trunkering, *f.eks "finn Per Hansen og kongebryllup" , "finn kongebryllup* og dronning Sonja", "finn bil og hytte og havet", eller frasesøk som " finn (hytte ved havet)" hvor "hytte ved havet" er frasen.* Hvilken syntaks som brukes avhenger av systemet man benytter, men jeg har her valgt en naturlig språk tilnærming for å illustrere dette. Ulempen med fritekstsøk er at den kun sjekker dokumentet for de søketermer man oppgir i en forespørsel, og hvorvidt søketermen forekommer i metadataposten til dokumentet eller i fritekst har lik vektlegging. Dette til tross for at søketermer som forekommer i metadataposten har høyere relevans enn dersom den forekommer i råteksten til dokumentet ved søk i fulltekstdatabaser. Fritekstsøk sjekker bare om søketermene man oppgir finnes eller ikke i de dokumentene som er tilgjengelig i databasen, og ved søk på generelle søketermer vil man ofte få store trefflister, som man selv må gjennomløpe for å vurdere relevansen av. *Dersom man f.eks er ute etter informasjon om alle saker som journalist Geir Bakke har skrevet og en skriver "finn Geir Bakke" eller "finn Geir og Bakke", vil systemet lete etter dokumenter hvor termene Geir og Bakke i hele dokumentet. Systemet vil ikke skille mellom dokumenter som har Geir Bakke i journalist elementet, eller selve fullteksten av artikkelen, men rangere disse likt. Dermed kan man få opp dokumenter som er urelevante og omhandler info om en annen Geir Bakke enn han man var ute etter, fordi en ikke søkte direkte på metadataelementet for journalist.*

4.5. Metadatasøk

Metadata (også omtalt som metasearch) brukes i flere sammenhenger i informasjonsgjenfinning. Metadata er et hjelpemiddel til å finne den informasjonen man søker, dvs en god gammeldags index. Men når man benytter denne indexen i elektronisk sammenheng blir effekten av den så mye større, man kan f.eks få opp alle dokumenter som en forfatter har

skrevet i ett eneste søk, og få dem frem i fulltekst på skjermen, istedet for å lete dem opp på hyllen som ville tatt mye lengre tid.

I relasjonsdatabasesammenheng er metadataene attributtene i tabellene i databasen, hvor dokumenter er representert med tupler. Disse tuplene er dokumentsurrogater fordi det representerer det opprinnelige dokumentet. Dersom hvert dokument er representert med DC f.eks i en relasjonsdatabase, kan man benytte seg av metadatasøk, og få referanse til hele dokumentet.

Når man da skal utføre et metadata søk søker man spesifikt på det attributtet en ønsker, f.eks en har en relasjonsdatabase med følgende attributter

Dokument(DC.Identifier, DC.Creator, DC.Subject ....osv)

hvor forespørselen i naturlig språk blir *"**finn** Hansen **i** Creator"*. Creator er attributtet i relasjonsdatabasen som sammen med andre metadataelement beskriver dokumentet.

Søk på metadata gjør at man kan utføre raskere søk som gir bedre tilbakemelding. I kommandogrensesnitt forutsetter dette at man kjenner de metadataelement det søker på, men i mange webgrensesnitt er det bare å fylle inn søketermer i de aktuelle metadataelementene som er representert.

Ofte benytter en både metadatasøk og fritekstsøk i kombinasjon, noe som er veldig nyttig dersom en opererer med fulltekstdatabaser.

I fulltekstdatabaser kan en raskere bestemme relevansen av et dokument siden hele dokumentet er tilgjengelig og det i tillegg tilbys en sammendrag av artikkelen/dokumentets innhold. Ikke alle systemer, som f.eks BIBSYS, tilbyr Abstract /Sammendrag for sine dokumenter. Dette gjør at det er vanskeligere å vurdere dokumentets innhold bare ut fra tittel og emneord. BIBSYS brukes lett og raskt dersom man søker etter bestemte dokumenter som en vet finnes, og vet hva det inneholder. Vet en ikke hva en søker etter, men er ute etter informasjon innenfor et tema, vet en ikke sikkert om informasjonen en har funnet er relevant før en har mottatt boken/artikkelen og bladd igjennom den. Når man så mottar boken og finner at den ikke er relevant, må man fortsette søkeprosessen på samme måte. Dette er faktorer som kompliserer og forlenger informasjonssøkingsprosessen. Et sammendrag til dokumentet ville her ha bidratt til å effektivisere informasjonssøkingsprosessen.

Bruk av metadatasøk vil ha størst effekt i et større system hvor en har lagret store datamengder, eller dersom en har en database hvor en stor andel av dokumentene omhandler et tema som er forholdsvis likt. I mindre databaser vil en nok oppdage at man ofte får like gode treff eller samme

tilbakemelding som man ville fått i fritekst. Behersker man allikevel bruk av metadata ved søk vil denne metoden alikevel være mer pålitelig enn fritekstsøk.

Det er enklere for bruker å benytte seg av metadatasøk når metadata er representert i bokser, som de ofte er i Web-grensesnitt. Alt en da har å gjøre er å skrive inn verdien for de ulike metadataelement man vil søke på, *f.eks Geir Bakke i metadataboksen med navnet journalist*. I kommando grensesnitt må man skrive hele setninger helt ut hvor også navnet på de metadataelement man søker på må skrives. Dette er mer tidkrevende og krever en et minimum av kunnskaper for i det hele tatt få til å komme i gang med et søk. Man blir i større grad invitert til bruk av metadata i web-grensesnitt enn i kommando grensesnitt.

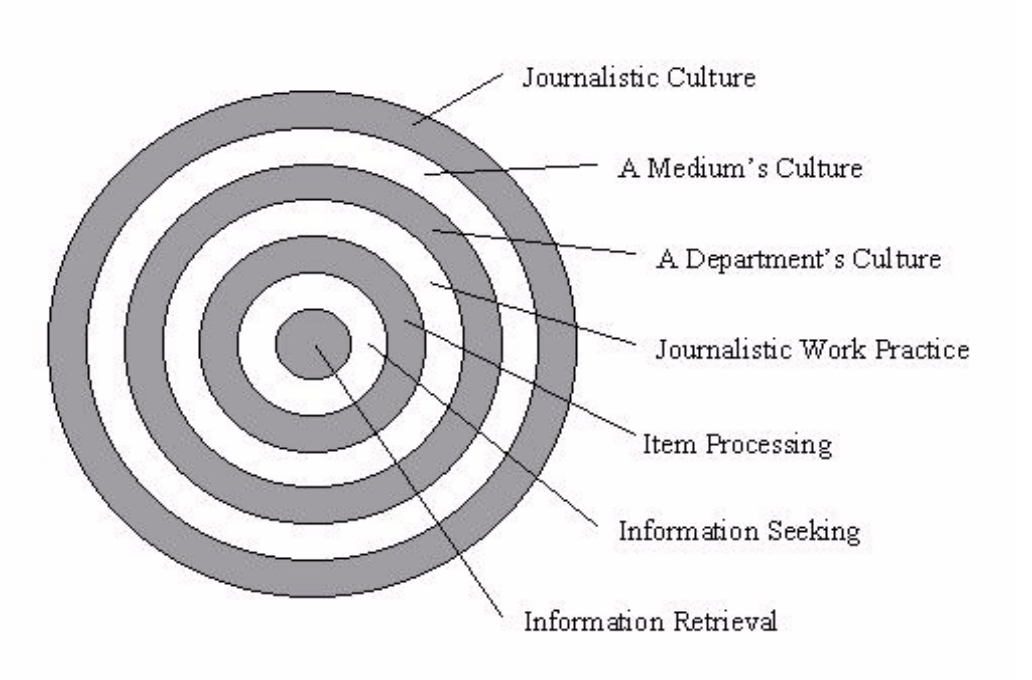
Metadatasøk forutsetter at man faktisk har en del informasjon å starte med, f.eks et emne man er interessert, artikler en bestemt journalist har skrevet osv. Når man skal søke etter informasjon har en iallfall tema eller emne klart, og da kan metadatasøk hjelpe deg dersom informasjonen en leter etter er indeksert etter emneord.

4.6. Journalistisk søkeoppførsel

Termen søkeoppførsel har referanse til både informasjonssøking og gjenfinningsprosesser, som Marchionini påpeker [26].

Innenfor den enkelte stilling er det bestemte oppgaver som skal gjennomføres. En journalist sin oppgave er å produsere stoff til avisen, og deres informasjonsbehov vil dermed følge de arbeidsoppgaver som skal utføres. I informasjonssøkingsprosessen vil en bli påvirket av flere faktorer, såvel som type informasjon en søker til personlige egenskaper og det arbeidsmiljø en er en del av.

Hannele Fabritius [29] har basert på sine undersøkelser utarbeidet en modell for journalistisk søkeoppførsel, og hva som påvirker den. Hun deler modellen inn i syv konsept som går fra det generelle til det spesielle som figur 4.2. viser :



FIGUR 4.2. Journalistisk søkeoppførsel

Konsept 1 : **Den journalistiske kulturen**.

Dette vil si normer, aktiviteter og rutiner som generelt kjennetegner journalistyrket.

Konsept 2 : **Kulturen innen en type medium**.

Dette er kulturen innenfor et spesifikt journalistisk medium, f.eks aviskulturen, radiokulturen, fjersynskulturen osv.

Konsept 3 : **Den enkelte avdeling sin kultur**.

Innenfor den enkelte bedrift vil det være en viss forskjell innenfor hver avdeling, *f.eks om man jobber i nyhetsavdelingen, kulturavdelingen eller*

sports-avdelingen. Hvilken type avdelingen man tilhører vil også si noe om type informasjon man er ute etter.

Konsept 4 : **Journalistisk arbeidspraksis**.

Her kombineres de generelle retningslinjer for journalistkulturen i tillegg til det medium og den avdeling en arbeider i. Den journalistiske arbeidspraksis her gir seg utslag i hvilke daglige rutiner og prosedyrer en utfører, som er i samsvar med bedriften sin overordnete målsetting.

Konsept 5 : **Arbeidsprosessen**.

Det vil si arbeidet med å forberede en nyhetsartikkel eller annet journalistisk arbeid. Om man skal skrive en artikkel, eller en filmanmeldelse vil påvirke hvilken informasjon en søker og hvordan man må gå frem for å innhente denne informasjonen.

Konsetp 6 : **Informasjonssøking**.

Kilder som skal brukes for å fullføre artikkelen eller film-anmeldelsen. Dersom man skal skrive en artikkel trenger en som oftest å foreta et intervju eller en observasjon. Om man skal skrive en filmanmeldelse er det nok å se filmen en skal omtale, i tillegg til at en må ha en kompetanse på å vurdere filmer. Ulike typer informasjonssøking kan være: intervjue personer, lese dokumenter, foreta observasjoner o.l. Informasjonen man finner kan også brukes for ulike formål, f.eks faktaopplysninger, sjekking av kilder og skrivemåter . En søker også informasjon for å oppdatere seg på en sak eller tema, f.eks om hvilke meninger andre ytret om akkurat dette temaet for noen år siden i forhold til dagens fokusering på det. Kildekritikk er et veldig viktig element i et grundig journalistisk arbeid, for det som produseres må komme fra pålitelig kilder som kan gå god for sannheten i det.

Konsept 7 : **Informasjonsgjenfinning**.

Dette innebærer hvilke informasjonsgjenfinningsteknikker og søkeferdigheter som er påkrevd for å søke etter informasjon som er digital lagret i databaser eller i mindre uorganiserte digitale ressurser.

Det er viktig å være klar over hva som kan påvirke din informasjonssøkingsprosess. Hvilke informasjonsbehov man har vil påvirke hvilke kilder man velger, og hvordan en oppfører seg i møte med den database man skal søke i. Dette er viktig å ta stilling til når en skal evaluere søkeprosessen. I tillegg til informasjonsbehovet vil også hvor gode søkeferdigheter man har i bruk av et system påvirke søkeprosessen, dermed kan utilstrekkelig opplæring være en av mange årsaker til at søkeprosessen ikke blir vellykket. Man kan allikevel kjenne systemet godt, men velge å ikke bruke det akkurat til den type informasjon man søker, alt etter tiden man har til rådighet og andre faktorer som må tas i betraktning. *F.eks: Har man det travelt og kjenner en kollega som har peiling på akkurat dette, spør man vedkommende på tur ut av døren, fremfor å velge andre kilder.* Årsaken er i dette tilfellet at man sparer tid. En journalist i en avis skal forsynes oss publikum med nyheter, og interne arkiv som SIFT vil dermed ha begrenset informasjon alt etter hvilke nyhetssaker det dreier seg om. En journalist sin hverdag er mye preget av arbeid "ute i felten", i form av intervjuer og observasjon, for å samle informasjon til å dekke en sak. *f.eks dekke ulike kulturarrangement, observere fotballkamper og andre sportsbegivenheter.* SIFT kan i en slik sammenheng fungere som et leksikon for internt bruk, til å samle fakta og oppdatere seg før en samler inn nytt stoff. I nyheter er det også vanlig å følge opp saker og da kan SIFT være grei å ha blant annet for å sjekke hvor mye har vi skrevet om dette temaet tidligere m.m.

Nettopp fordi SIFT inneholder "gammelt nytt", vil journalister alltid søke andre kilder i tillegg for å utføre sine arbeidsprosesser.

4.7. Søkemotorer generelt

Søkemotorer på nettet er kjent for å ofte forsyne oss med informasjon i form av kvantitet og ikke kvalitet. En side i disse søkemotorene er ofte indeksert flere ganger, dvs at man istedet for å få opp linken til URL'en som representerer hoveddokumentet til en hjemmeside, får opp alle sub-dokumentene som hoveddokumentet refererer til. Dette er fordi søkemotorene ofte vurderer relevans etter hvor mange linker det er til den siden, eller hvor mange ganger ord forekommer i et dokument.

Det er ikke alltid like lett å finne ut hvilke regler de ulike web-søkemotorer arbeider etter. Disse er kommersielle søkemotorer og er tydelig organisert etter dette formål. Hovedinntrykket ved bruk er at veldig få tar hensyn til metadata fordi dette er en kjent måte å "spamme" informasjon på. Spamming vil si "masseposting" i form av store mengder uadressert elektronisk post sendes mot

en server, program som er satt opp til å søke etter e-mail adresser for deretter å sende ut mange mail i en evig løkke. Spamming brukes ofte av hackere til å spre virus m.m.

De tre største søkemotorene som eksisterer på nettet i dag er Altavista, Northern Light og Fast Search. De aller fleste søkmotorer tar liten eller ingen hensyn til metadata, men fordi de rangere relevans etter 5-6 første linjene i dokumentet, kan føre til at dokumentet ditt før høyere prioritet med metadata integrert da de nettopp ligger først i HTML-koden. Men dette er tilfeldig og man har ingen garanti for at gjenfinning av dokumentet øker via søkemotorene.

Søkemotorer gjør lite bruk av metadata, og det er ikke alltid lett å vurdere kilder på det man finner på nettet og om det kommer fra en pålitelig kilde. Stuart Weibel har uttalt i en av sine mange publikasjoner at vi går fra en “where do I click” til en “who do you trust” mentalitet på Internett i dag. Dersom vi hadde fått en metadata legalisering, der alle brukte et standardisert metadataformat, ville det vært lettere for ulike søkemotorer å ta hensyn til dette. Allikevel er det et problem dersom en ikke benytter seg av representative metadata i indekseringen, dvs at en bevisst indekserer feil for å få treff fra søkemotorene på termer en vet det vil bli foretatt søk på. *F.eks kan man sette navn på kjente personer inn som metadata i metadata tag’ene, selv om dokumentet ditt handler om noe helt annet.* Dette kan en unngå ved at metadata i metadataelementene som ikke er representert i dokumentet ikke blir akseptert, men allikevel blir man ikke utelukke det faktum at en kan benytte emneord som representerer dokumentets innhold til tross for at dette emneord ikke er nevnt i selve dokumentet sin råtekst.

4.7.1. Altavista

Altavista kan forstå metadata av typen Description og Keywords, integrert i HTML koden [67]. Når Altavista da indeksere de 1000 første tegnende i dokumentet ditt, vil Keywords og Description tag’enes innhold få høyest relevansvurdering i forhold til annen informasjons innenfor de 1000 tegnene som blir indeksert. I tillegg kan man i Altavista sine avanserte søkemuligheter søke på språk og dato, men det er ikke noe her som tilsier at man spesifikt kan søke mot HTML dokument som har metadata for språk og dato inkorporert i sine dokumenter. *F.eks: Her vil søk på termene nyheter, kultur osv i keywords under META tag’en nedenfor gi dokumentet ditt høyerer relevans enn hvis disse termene kun forekom i resten av dokumentet og ikke i META tag’en.*

```
<META NAME="keywords" CONTENT="nyheter, kultur, underholdning, internett, avis,
news, norway, norge, nettavisen">
<META NAME="description" CONTENT="Temaet for siden din">
```

Mellom disse tag’ene kan man fint legge inn annet enn det siden i realiteten representerer. F.eks vil man at alle absolutt skal få opp hjemmesiden din, kan man legge inn ord her som en vet

"alle" søker på, f.eks vil ord som kjente personer ala Micheal Jackson lagt inn her føre til at mange blir "ledet feil" til siden din, i håp om å finne noe som de er interessert i.

4.7.2. Northern Light & Fast Search

Northern Light [68] indekserer hele dokumentet, og informasjon i META og ALT tag'er i HTML koden blir ikke indeksert. Relevansvurderinger blir her gjort på frekvens av søketerm, om term forekommer i tittel eller i resten av dokumentet, hvor mange andre sider i indeksen deres har linker til denne siden, og analyser av syntaks og semantikk av naturlig språk prosessering (NLP). Fast Search [69] tar heller ikke hensyn til metadatasøking pr i dag, så langt det har lykkes meg å finne informasjon om det via deres hjemmeside.

4.8. Indekseringsverktøy

Det finnes i dag et tyvetalls indekseringsverktøy tilgjengelig via Internett. Mange av disse er blitt til gjennom ulike metadata prosjekter. Flere av disse verktøyene støtter indeksering av DC metadata, deriblant verktøyet DC Dot [71] fra UKOLN [70]og Dublin Core Metadata Template [72] utviklet av Nordic Metadata Project [36]. DC Dot er helt automatisk ved at man gir inn en URL til dokumentet man vil ha indeksert og dermed får ut metadata, noe som fører til at en allikevel må inn å rette utvalget. Den ikke klarer å fange opp alle metadata som en trenger for dokumentet fordi utvalget gjøres på innholdet i HTML tag'er. Har man derimot DC metadata inkorporert i selve HTML dokumentet kan man bruke DC Dot til å sjekke om den finner alle DC metadata som en faktisk har lagt inn manuelt. Ikke alle disse fungerer like godt og en må evaluere verktøyene for å finne ut hvilke som passer best til sitt behov.

Kapittel 5

Metadataformater

"... verken Dublin Core eller andre metadataformater kan erstatte tradisjonelle katalogdataposter i bibliotekkataloger eller nasjonalbibliografiske databaser ..."
- Bendik Rugaas

5.1. Metadataformater generelt

Det finnes i dag en rekke ulike metadataformater eller metadataskjemaer, med den hensikt å beskrive ulike typer informasjonsressurser og informasjonsobjekt til støtte for gjenfinning. Som eksempel kan nevnes MARC, Dublin Core, SOIF, AARCII, ISBD, FGDC og mange flere. I dette kapittelet blir MARC, BIBSYS-MARC, Dublin Core og Ad-hoc format generelt beskrevet da det er disse formater som refereres til i denne avhandlingen. Dublin Core beskrives i detalj, da det er dette formatet som blir drøftet i mapping mot Ad-hoc formatene i kapittel 8.

5.2. MARC

Her følger historikk, anvendelsesområder for MARC formatet og en kort presentasjon av BIBSYS-MARC som er omtalt i dette informasjonsrommet.

5.2.1. Historikk

MARC (Machine readable Catalogue Format) [1] ble utviklet på 1960 tallet. Dette formatet ble utviklet med den hensikt å utveksle bibliografiske poster mellom biblioteksystemer over hele verden, slik at en kunne redusere inn på katalogiseringskostnader. Dette innebærer at et bibliotek i flere land (og da som oftest nasjonalbiblioteket) katalogiserer den litteratur som utgis i deres land, og så utveksler en katalogiseringsposter fra hverandre. *Dette kan eksemplifiseres ved at dersom et dansk bibliotek kjøper inn en norsk bok, vil de kunne søke opp boken i katalogiseringstjenesten til Nasjonalbiblioteket i Mo i Rana,*

for deretter å overføre de katalogiseringsopplysninger som ligger der. I tillegg settes på lokal klassifisering, som dokument identifikator og koder for fysisk plassering.

MARC formatet ble senere standardisert til ISO 2709. En MARC post har et hode som i tillegg til selve dataene også beskriver hvor dataene ligger i posten. De fleste bibliotek verden over har sitt avkom av MARC formatet, det vil si de har "arvet" fra MARC formatet, og tilpasset det til sitt behov. Vi har dermed fått ulike tilpasninger av MARC standarden, som NORMARC, DANMARC, USMARC, UNIMARC eller BIBSYS-MARC. Konsekvensen av dette er at alle land bruker de samme feltkodene, men ha ulike tolkninger av hva de legger i feltkodene, noe som igjen fører til at et MARC format ikke nødvendigvis er direkte kompatibelt med et annet. Selv om formatene ikke trenger å være direkte kompatible bygger de på samme grunnideer som gjør arbeidet med utveksling av poster betydelig lettere.

5.2.2. BIBSYS-MARC

BIBSYS er biblioteksystemet som benyttes av alle forskningsbibliotekene, Nasjonalbiblioteket og en rekke høgskole og fagbibliotek i Norge [19]. BIBSYS databasen er en samling av deldatabaser, og tilhører pr. i dag over 2 millioner bøker og annen type litteratur, som til sammen utgjør hele 6 millioner eksemplarer.

BIBSYS-MARC er formatet som benyttes til å beskrive de ulike informasjonsobjekter i BIBSYS databasen. BIBSYS-MARC er et surrogat av NORMARC formatet. I "mitt" virtuelle informasjonsrom er BIBSYS et av de medier som er tenkt integrert, men BIBSYS-MARC er ikke et aktuelt kommunikasjonsformat å benytte for felles kommunikasjonsformat i informasjons-rommet, nettopp fordi det er altfor avansert for den alminnelige bruker og som ikke umiddelbar gir bruker en intuitiv forståelse. Det tar lang tid å forstå MARC og BIBSYS-MARC i detalj, og å benytte vite hvordan disse formatene skal benyttes på en riktig måte krever lang erfaring av den som skal katalogisere.

5.2.3. Anvendelsesområder

MARC formatet brukes hovedsaklig av biblioteker verden over, og er mest brukt til beskrivelse av ulike typer litteratur, men det er også anvendelig for bruk til indeksering av bilder, selv om ikke dette er særlig utbredt. BIBSYS-MARC er brukt til å katalogisere manuskriptkart, og Det Kongelige Bibliotek i Danmark har brukt DANMARC til katalogisering av bilder. Siden MARC formatet er spesial-utviklet til bruk i bibliotek er det få bedrifter/organisasjoner som benytter seg direkte av MARC, men flere systemer som disse benytter i sitt arkiv er utarbeidet med utgangspunkt i MARC formatet. MARC formatet er såpass detaljert at et fåtall organisasjoner og bedrifter vil se seg tjent med å benytte dette formatet til sitt behov. Det krever god erfaring i

MARC formatet tilsvarende kunnskap som tilegnes i stillinger som bibliotekar eller katalogisator, for å forstå formatet godt nok til å bruke det i sin helhet. I hvilken grad bibliotek i bedrifter og organisasjoner benytter seg av MARC i sine elektroniske arkiv vil også være avhengig av om hvem som arbeider i arkivet. Dersom bedriftsarkiver har ansatte som har arkiv og biblioteksutdannelse øker sannsynligheten for at MARC formatet benyttes, siden de kjenner formatet forholdsvis godt.

5.3. Dublin Core

Først følger historikk og anvendelsesområder for Dublin Core (DC) formatet. Videre introduseres:

- DC Kvalifikatorer
- DC sine sub-elementer
- Søkemotorer som støtter DC
- Hvordan lagre DC metadata

5.3.1. Historikk

Arbeidet med formatet startet opp i 1995 og det foregår stadig forskning på formatet gjennom prosjektet som vi kjenner til som "The Dublin Core metadata intitiative" (DCMI) [43].

DC har internasjonalt engasjement og interesse. Utviklingen av DC formatet samler fagfolk fra ulike disipliner innen IT og Bibliotek, hvor bibliotekarer, forskere innen digitale bibliotek, katalogiseringseksperter og eksperter på text-markup deltar og samarbeider. Det er viktig å komme frem til felles tilnærminger for gjenfinning av informasjon på Internett. Stadig flere vurderer derfor DC som metadataformat i sine digitaliserte samlinger, og flere prosjekter pågår i universitets og forskningsmiljø verden over. Flere av disse forskningsprosjektene utvikler verktøy som skal gjøre det enda lettere for deg som bruker å ta i bruk DC for indeksering av dine web-dokumenter, som blant annet "DC Dot" verktøyet utviklet av UK Office for Library and Information Networking (UKOLN) [70] og Dublin Core Metadata Template [72], utviklet av Nordic Metadata Project [36].

Dublin Core er et metadataformat som er utviklet med den hensikt å tjene som et katalogiseringsformat for publisering av informasjonsressurser på Internett for å støtte gjenfinning av og tilgang til elektroniske dokumenter distribuert over Internett. "The Dublin

Core metadata element set", referert til ved forkortelser som Dublin Core eller DC har til hensikt å være enkelt å bruke og forstå uten forhåndskunnskaper til katalogiseringsregler.

Navnet Dublin Core stammer fra den første workshop som ble holdt i Dublin, Ohio i USA hvor utviklingen av DC startet i kraft av DC direktoratet som holder til ved "Online Computer Library Center"(OCLC). Navnet "Core" betyr kjerne og definerer det DC er ment å tjene som, nemlig en "kjerne" av basiselement for katalogisering. OCLC er en ikke kommersiell organisasjon som ivaretar alle bibliotek i USA og 70 andre bibliotek sine interesser. DCMI har allikevel fått en internasjonal fokusering og interesse. Da DCMI startet i 1995 besto DC av forslag til 13 metadataelement, men siden elementene coverage og rights ble innført i 1996 har alle 15 element vært uendret.

Helt siden DCMI startet i 1995/1996 har det vært snakk om utvidelser av DC, til tross for at hele hensikten var å foreslå et minimum av beskrivelseselementer for å oppnå interoperabilitet mellom systemer. En har imidlertid sett behov for utvidelser i form av nettopp sub-elementer, men DCMI sine arbeidsgrupper har enda ikke oppnådd enighet om hvilke sub-elementer som skal foreslås. Som en følge av dette eksisterer det ikke noen ferdig forslagsrapport på dette pr. dato.

DC er ikke ment å skulle erstatte metadataformater som MARC og AACR2, men å tjene som en kjerne av beskrivelses elementer som kan brukes av alle uten kjennskap til standardiserte katalogiseringregler for å kunne beskrive ressurser på en enklest mulig måte. Målet er å få "allmuen" av de som publiserer sine informasjonsressurser på Internett til å bruke DC nettopp fordi det er enkelt, og dermed bidra til at informasjonsressuser på Internett har et minimum av katalogi-seringsinformasjon til støtte i gjenfinning og relevansvurdering. En ønsker også at DC skal være såpass fleksibelt at det i tillegg til å passe for den vanlige bruker også kan tilpasses de bedrifter og organisasjoner med mer spesialiserte behov.

For de som trenger videre katalogisering ut fra sine behov kan DC være med å danne grunnmuren for katalogisering hos disse, dvs DC er utvidbar etter behov i form av innføring av sub-element.

Det arbeider stadig for å få DC gjennom de ulike prosesser for standardisering, og det er en egen arbeidsgruppe som arbeider for dette. Nylig har DC fått sin tilslutning i European Commitee for Standariszation (CEN) [59], og det arbeides nå videre for også å få DC akseptert som standard i National Information Standard Organization (NISO) og International Standard Organization (ISO).

I tillegg til at det er ulike arbeidsgrupper som arbeider med videreføring og utvidelse av de 15 basiselementene, er det arbeidsgrupper som også arbeider for hvordan DC kan fungere for å katalogisere utdanningsressurser som er anvendelig for mange nasjonale utdanningsinstitu-sjoner (f.eks K-12, videre og høyere utdanning og livslang læring)

I tillegg samarbeider DCMI kontinuerlig med W3C og andre organisasjoner som er opptatt av å løse problemer med katalogisering på Internett. Status om DCMI arbeid er tilgjengelig via deres hjemmeside. Hjemmesiden oppdateres kontinuerlig hvor de ferskeste arbeidsutkast og publikasjoner de ulike arbeidsgrupper er tilgjengelig til enhver tid.

Hovedideen om at DC skal være så enkelt at forfatter eller utgiver av en elektronisk publikasjon selv kan utforme beskrivelsen og integrere dette i selve dokumentet er viktig å ha i minne ved utvikling av sub-elementer. DC må ikke bli utvidet i den grad at det står i strid med dens grunnide, nettopp at DC skulle være enkelt å bruke for alle.

5.3.2. Anvendelsesområder

Dublin Core er utviklet med den hensikt å skulle brukes til å beskrive web-dokumenter, og dette er som kjent multimediadokumenter som kan bestå av både bilder,tekst, video, lyd m.m.

I flere prosjekter er Dublin Core brukt for ulike typer informasjonsobjekt. Anne Marie Vercoustre har brukt DC for fotografier [37] i sitt prosjekt.

Hun har funnet ut at DC har sine begrensninger når det gjelder definering av surrogat dokument og når flere fotografier representerer samme objekt. DC vil selvfølgelig har sine begrensninger på en del Internett ressurser, men hele hensikten med DC er at det skulle være enkelt å bruke for Internett ressuser flest. Det arbeides for at DC skal være mer fleksibelt og ta hensyn til stadig flere behov, men DC har sine begrensninger og er likevel ikke ment å tilfredstille de med svært spesialiserte behov. For de med spesialiserte behov kan en bruke DC som grunnmur for katalogisering og spe på med de metadata som en måtte trenge i tillegg.

Privatpersoner som velger å publisere sine dokumenter på Internett vil allikevel ikke ha de samme detaljerte behov for å katalogisere sine dokumenter som bedrifter og institusjoner ønsker. Det er naturlig å velge ut hvilke brukergrupper DC kan fungere for, noe som vil bli dokumentert etterhvert som de ulike prosjekter som benytter DC gjør sine erfaringer med dette. Ved hjelp av XML (Extended markup Language) kan du lage en Document Type Definition (DTD) for din dokumenttype som har DC som basiselementer for indeksering, men samtidig definere egne element utover DC dersom man har behov for det.

Det kan nesten virke som det er umulig å oppnå enighet fra DCMI om hvilke sub-element som skal foreslås her nettopp fordi det er så mange ulike behov. Drøfting av DC sub-element eller "qualifiers" som de omtales som, er blitt gjort ut i fra ulike behov som prosjekt tilknyttet DCMI har gitt uttrykk for. Det er forventet at det skal komme en rapport om oppnådd enighet på sub-element området i løpet av våren 2000 i følge sentrale aktører i DCMI.

Sub-element blir kort presentert i 5.3.4, mens detaljert redegjørelse for de sub-element som foreligger pr. dato er lagt i Appendix A.

The Dublin Core Element Set 1.1.

Her følger de 15 elementer i Dublin Core Element Set 1.0 [7] i form av en kort beskrivelse til hvert element. Den norske oversettelsen bak er basert på Ole Husby sin oversettelse av DC Element Set 1.0 [22]. Alle elementene i DC kan repeteres, *f.eks har du to forfattere dekkes dette ved å gjenta DC.Creator elementet.*

Tabell 5.1. The Dublin Core Metadata Element Set [43]

Elementnavn	Elementbeskrivelse	Norsk oversettelse av elementet
Title	Navnet på ressursen gitt av Creator eller publisher	Tittel
Creator	Personen eller organisasjonen primært ansvarlig for å ha laget det intellektuelle innholdet i ressursen.	Forfatter eller Opphavsmann
Subject	Temaet for ressursen, det være nøkkelord eller fraser som beskriver tema eller innhold i ressursen, inkludert kontrollert vokabular og klassifkasjonsskjemaer.	Emne
Description	En tekstlig beskrivelse av innholdet i ressursen, f.eks sammendrag, innholdsfortegnelse, grafisk representasjon av innholdet eller bare muntlig eller skriftlig fritekst rapport av innholdet.	Beskrivelse
Publisher	Den som er ansvarlig for å gjøre ressursen tilgjengelig i sin nåværende form, som f.eks et forlag, en universitetsavdeling eller en organisasjonsenhet.	Utgiver
Contributor	Personer eller organisasjoner som har vært bidragsytere til informasjonsressursen. Bidraget må ha betydelige intellektuell innvirkning på ressursen, men at bidraget er sekundært i forhold til den /eller de som kan regnes som forfattere eller her CREATOR av ressursen. Eksempel på bidragsytere kan være editorer, illustratører eller.	Annen Bidragsyter
Date	En dato som kan være involvert i hvilken som helst hendelse av ressursen livssyklus. Det er vanlig å oppfatte bruk av dette feltet som den dato ressursen ble gjort offentlig tilgjengelig eller publisert.	Dato

Tabell 5.1. The Dublin Core Metadata Element Set [43]

Elementnavn	Elementbeskrivelse	Norsk oversettelse av elementet
Type	Her etterspørres hvilken kategori eller genre som informasjonsressursen hører inn under. F.eks er det en hjemmeside, en roman, et dikt, arbeidsutkast, teknisk rapport osv.	Type
Format	Her er det datarepresentasjonen av ressursen som en vil spesifisere, om ressursen er representert i text/ html, ASCII, postcript file, word dokument, kjørende applikasjon, et jpg eller gif bilde osv.	Format
Identifier	Dette er den unike identifikatoren til ressursen, og kan bestå av tekst eller tall. Eksempel på en identifikator som brukes på Internett ressuser er URLs og URNs, og i tillegg har du andre globale unike identifikatorer som ISBN (International Standard Book Numbers.	Identifikator
Source	Her er en ute etter en referanse til en kilde som den nåværende ressursen stammer fra eller har sin opprinnelse fra. Denne kilden kan eksistere i trykt/fysisk form eller elektronisk.	Kilde
Language	Hvilket språk er ressursen tilgjengelig på.	Språk
Relation	Relasjoner ressursen har til andre ressurser, men hvor ressursene eksisterer som uavhengige ressurser.	Relasjon
Coverage	Begrensninger som ressursen måtte ha i tid og rom spesifiseres her.	Dekning
Rights	Henviselse til den som har opphavsretten til ressursen, og kan bestå av en URL til en notis om opphavsrettigheter.	Rettigheter

Disse elementene er delt inn i følgende gruppene innhold, intellektuell eiendom og instansering som vist i tabell 5.2.

Tabell 5.2. Grupperinger av DC.Elements etter hva de beskriver [55].

Innhold	Intellektuell eiendom	Instansering
Title	Creator	Date
Subject	Publisher	Format
Description	Contributor	Identifier
Type	Rights	Language
Source		
Relation		
Coverage		

5.3.3. DC Kvalifikatorer

For å ivareta både de som er minimalister og holder på DC som bestående av 15 element, og de som ønsker en finere struktur for å ivareta en mer detaljert katalogisering er det opprettet 3 kvalifikatorer som er ment å spesifisere innholdet av elementene ytterligere.

1. Sub-element
2. Scheme
3. Language

Et SUB-ELEMENT er en presisering av hovedelementet, f.eks DC.Creator.Agentrole. Dette blir videre forklart i neste avsnitt.

SCHEME kvalitfikatoren spesifiserer om et bestemt regelverk eller et kontrollert vokabular blir benyttet. Dette tillatter at en element sin verdi kan tilhøre et bestemt klassifikasjonssystem, *f.eks Dewey Decimal Classification eller en Art & Architecture Thesaurus for cultural heritage.*

LANGUAGE kvalifikatoren, forkortet til LANG spesifiserer hvilket språk innholdet i elementet er på.

For å illustrere bruken av disse kvalifikatorene tar vi utgangspunkt i DC elementet Subject og sub-elementet Classification:

DC.Subject.Classification="Beskrivelse i tekst av ressursen" SHCEME=Dewey Decimal Code (DDC) LANG=no.

5.3.4. DC Sub-elementer

DC er som sagt ment å være enkelt, og en skal derfor passe seg for at det blir for spesialisert ved å innføre flere utvidelser i form av sub-element. Arbeidsgruppen for sub-element har i sin rapport fra 1998 [56] sett på det som en nødvendighet med sub-element for noen DC element, såkalte "Qualifiers", for at DC skal bli effektiv nok for gjenfinning av web-ressurser. Deres definisjon på et sub-element er at det skal bringe klarhet i DC elementes innhold. *F.eks til dato elementet kan det oppstå spørsmål om hvilken dato det gjelder i ressursens levesyklus fra den er skapt til produsert. Er det dato for den dagen ressursen var ferdig skapt, eller er det den dagen ressursen ble publisert. DC.Date.Issued viser til publiseringsdagen til ressursen, mens DC.Date.Created er den dato ressursen var ferdig skapt.*

Arbeidsgruppen viser også til at kvalifikatorer som SCHEME og LANG også er nødvendige i tillegg til bruk av sub-element.

I mappingen i kapittel 8. blir flere sub-element benyttet. DCMI har kommet med et utkast til sub-element for DC [58] pr. dato, men dette er enda ikke helt formalisert. Jeg velger å legge presentasjon av sub-element i appendix A, samt en kort statusrapport for arbeidet med de ulike sub-element.

Eksempel på sub-element som DCMI foreslår er f.*eks innføring av Agentrole, Agentname, AgentAffiliation osv for DC elementene Creator, Publisher og Contributor som refereres til som DC.Creator.Agentrole, DC.Creator.Agentname osv.*

Flere prosjekt tilknyttet DCMI har tatt i bruk sub-elementer etter sitt behov. De ulike prosjekters anvendelser av sub-element og deres erfaringer drøftes i arbeidsgruppene for eventuelle nye sub-element eller endring av allerede eksisterende sub-element.

5.3.5. Søkemotorer som støtter Dublin Core

Som tidligere nevnt i kapittel 4 om informasjonssøkeprosessen er det veldig få søkemotorer som indekserer informasjon i <meta> tag'ene i web dokumentetet ditt, p.g.a spamming. De søkemotorere som støtter Dublin Core pr. i dag er:

1. Ultraseek [60]
2. Swish-E [61]
3. Microsoft's Index Server [62]
4. Autonomy Knowledge Server [63]
5. Blue Angel Technologies MetaStar [64]

6. Verity Search 97 Information Server [65]

Dette er søkemotorer som er tilgjengelig for søk mot en spesifikk database, ikke for bruk over hele web'en som Altavista, Northern light osv.

5.3.6. Hvordan lagre DC metadata?

DC Metadata Element Set er et sett av 15 beskrivelses element med definert semantikk. Det betyr i praksis at mange metoder kan brukes for å innhente eller overføre DC metadata. Vanlige metoder her inkluderer bruk av HTML, XML, RDF og relasjonsdatabaser. Hvilke metoder som brukes må vurderes ut fra behov og hvilke begrensninger de ulike metodene gir.

Ved bruk av HTML er metadata inkorporert i selve dokumentets innhold, det som omtales om inkorporerte metadata. Disse metadata er en del av HTML koden som dokumentet representerer. DC metadata er da satt inn mellom <HEAD> og </HEAD> tag'ene i HTML koden. Her forekommer det en liste av ulike metadatalement, hvor det er ingen gruppering eller annen struktur.

RDF har som hensikt å fungere som et fundament for å prosessere metadata, hvor flere personer fra DCMI har vært sentral i utviklingen av RDF. RDF støtter interoperabilitet mellom applikasjoner som utveksler maskinforståelig informasjon på Web'en. Hovedmålet til RDF er å definere en metode som er passende for å beskrive informasjonressurser uavhengig av domene og et domenes semantikk.

XML tillatter at en langt flere strukturerte metadata kan bli innhentet. Svært få programvare og nettlesere støtter bruk av XML kodede data på nåværende tidspunkt, og dersom en skal bruke XML må en konfigurere programvare og løsninger for dette selv. Støtte for dette vil imidlertid komme etterhvert med nyere programvare. En av fordelene med XML er at du kan definere egne metadataelement som skal være gyldige for den type dokument man benytter ved f.eks en artikkel hvor ingen andre element enn de som er definert for den dokumenttypen blir akseptert. Dette gjør at man ved hjelp av XML kan definere en dokumenttype for et dokument som inkluderer DC Metadata Element Set, og i tillegg kan man definere metadata utover DC sine element dersom det er behov for det. XML har den fordel at det er lett å utveksle metadata mellom bedrifter med samme behov *f.eks utveksling av Metadata for artikler mellom aviser.* XML metadata er informasjon som er forståelig både for maskiner og mennesker. XML har syntaks for å kode utsagn i RDF og i RDF kodet DC Metadata.

For best mulig ytelse i informasjonsøking og gjenfinning, er det best å laste DC metadata inn i selve databasesystemet direkte, men dette forutsetter at metadata knyttes til det dokumentet det beskriver, som igjen stiller krav til konsistens og synkronisering mellom prosedyrer.

5.4. Ad-hoc formater

Ad-hoc formater er formater som ikke er standardisert, men som er blitt til på grunnlag av katalogiseringsregler som den enkelte bedrift eller organisasjon har definert etter sitt behov. Disse formatene omtales også som proprietære format eller "in-house rules" i flere artikler. Ad-hoc formater har stor utbredelse, da svært få bedrifter og organisasjoner bruker standardisert format og heller velger sine egne løsninger. Ad-hoc formatet store utbredelse skyldes til dels at standardiserte format som MARC, er for detaljerte og kompliserte til at bedrifter og organisasjoner ser seg tjent med å benytte seg av disse. Ad-hoc formater har til hensikt å fungere til internt brukt, og er i utgangspunktet ikke ment for publisering av informasjon. Via Internett publiserer stadig flere bedrifter og organisasjoner sin informasjon. I denne sammenheng er det tid for å revurdere sine formater og vurdere om de er tilfredsstillende nok også for dette formål, eller om mer formaliserte formater bør benyttes. Det bør være et mål at en slik vurderingsprosess resulterer i en metadata-ontologi for indeksering og gjenfinning av bedriftens informasjonsressurser. Ofte opererer bedrifter med flere Ad-hoc formater internt i bedriften og en endring av dette er ikke helt smertefri eller gjort over natten.

Ved publisering av informasjon vil andre behov for metadata melde seg enn de du trenger til internt bruk. Det kan også hende at de ulike Ad-hoc formater som benyttes til internt bruk har konflikter med hensyn på navn, granularitet, syntaks og struktur. Margareth E. Graham har gjort en undersøkelse [17] hvor 60 medlemsinstitusjoner i "Art Library Society (ARLIS)" i Storbritannia deltok og hvor det blant annet var fokusert på indekseringspraksis. Institusjonene var bibli-otek, arkiv, museum osv. Undersøkelsen omfattet kun indeksering av bilder, men i denne under-søkelsen var det hele 68 % som benyttet in-house rules for indeksering av fotografier. I motsetning brukte 16 % MARC formatet og 26% AACR (Anglo american Cataloging rules) for indeksering av sine billedsamlinger. Dette gir et bilde på at Ad-hoc formatene foretrekkes av de fleste, selv innen flere forskningsarkiv og forskningsbibliotek

Kapittel 6

Case studiet

6.1. Innledning

Adresseavisen er her valgt som case-studium. Valg av bedrift ble tatt på bakgrunn av egne erfaringer jeg gjorde meg som ansatt ved arkivet på Adresseavisen sommeren 97 og 98 og som gjorde at jeg fikk ideen til problemstilling for min hovedoppgave innen informasjonsforvaltning og digitale bibliotek.

Adresseavisen er norges eldste avis med 455 ansatte (pr. 1 januar 1999), grunnlagt i 1767. De har et mangfold av samlinger i arkivet sitt, med mikrofilm helt tilbake til 1767, samt indeksering av bilder, stoff, filmer fra tidlig på 1960 tallet. Først i 1993 begynte de med digitalisering av noen av sine samlinger, og det er disse som blir presenteret nærmere her, i og med at det er gjenfinning av informasjon i digitale medier som er interessefeltet her.

Målgruppen min for undersøkelsen er arkivpersonell og ansatte i redaksjonen, dermed vil arbeidsprosesser fra disse avdelingene bli presentert her.

I dette kapittelet skal du få innblikk i hvilke søkemuligheter du har som bruker av SIFT systemet, bli kjent med de ulike metadataformater som Adresseavisen benytter og hvordan disse brukes, og hvordan kommunikasjonen mellom de ulike databasesystemer fungerer, samt de ulike arbeidsprosesser som er i kontakt med informasjonsressursene på en eller anne måte. Kjennskap til dette er viktig når en skal drøfte hvorvidt Adresseavisens behov kan tilfredsstilles i DC, fordi bruken av de ulike metadataelement og konteksten de står i er relevant for eventuelle forslag til løsning. Det er også viktig å få oversikt over hvordan Adresseavisens system og dokumentenes levevei i ulike prosesser fungerer. Slik situasjonen er i dag er det lite samsvar mellom de ulike arbeidsprosesser noe som gjør at indeksering av dokument utføres i flere ledd, ikke flyt i metadata fra arbeidsprosess til arbeidsprosess, flere

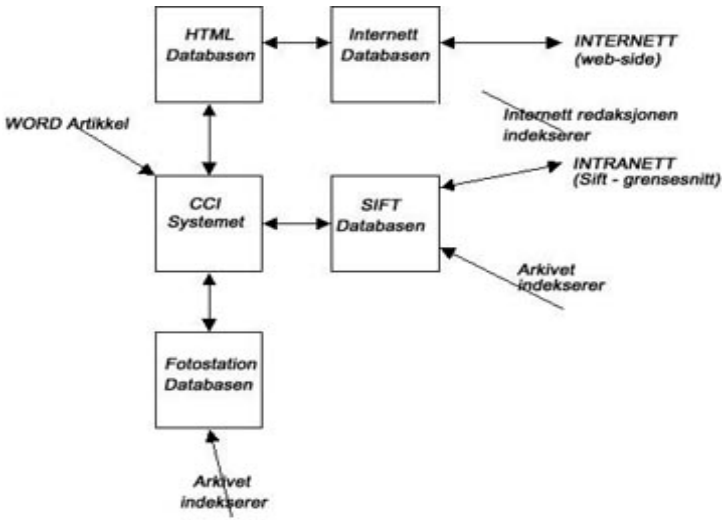
dokument lagres flere ganger i ulike databaser, og enkelte dokument blir også indeksert med to ulike formater osv. Dette igjen fører til tap av tid og effektivitet. Denne nåsituasjonen vil jeg konkretisere ved å presentere et scenario som viser problemstillingen en her står ovenfor. Ved innføring av et felles format kan prosesser samsvares og gjøres mer effektiv.

6.2. Adresseavisens datanettverk

I dette avsnittet får du vite mer om hvordan de ulike metadataformater benyttes, hvordan dokumenter indekseres, søkemuligheter, og hvilke arbeidsprosesser som utføres. Jeg ønsker at du skal kjenne til hele nåsituasjonen slik den er i dag, og som er det grunnlag jeg benytter for analyse, drøfting og forslag til løsning.

6.2.1. Hvilke databaser snakker sammen

Noen databaser kommuniserer med hverandre, mens andre overhodet ikke er på talefot. Dette illustreres i figur 6.1.



FIGUR 6.1. Adresseavisens databaser

HTML Databasen: Alle artikler i fulltekst som er produsert til trykk i avisen og de vedlegg i form av bilder m.m som er relatert til artiklene

FOTOSTATION Databasen: Alle digitale bilder som er scannet inn for å brukes på trykk i avisen det være fra papirform, digitalt kamera eller fra fysisk film m.m. Byråbilder kommer også inn her. Lagres i forskjellig mapper.

SIFT Databasen : Artikler i fulltekst, metadataposter for fysisk filmmapper, illustrasjoner, og bøker.

Internett Databasen: Alle artikler i fulltekst og eventuelle vedlegg til artikkelen som inngår i Adresseavisens Internettutgave. Alle artikler og vedlegg som ligger her er utdrag fra HTML databasen.

CCI Systemet: CCI står for Comment Computer Interface. Dette er produksjonssystemet som Adresseavisen bruker for å produsere hele avisen. Alt av bilder, artikler fra redaksjonen går inn her. I tillegg kommer reklame, annonsesider m.m fra annonseavdelingen og markedsføringsavdelingen.

6.2.2. Indekseringspraksis

Slik systemet for indeksering fungerer i dag er det 3 instanser som indekserer for sitt behov. Disse er arkivpersonalet, Internettredaksjonen og delvis journalistene som skriver artiklene. Journalistene som skriver sine artikler i CCI Word hvor metadata som tittel, mellomtittel, ingress, osv lagt på dokumentet. I tillegg benytter Internettredaksjonen sitt indekseringsformat for publisering av Internettartikler. Arkivet indekserer SIFT databasen, hvor alle artikler, illustrasjoner og fysisk film som har vært brukt i avisen er lagret og indeksert. I tillegg blir alle bilder som er brukt i avisen, både digitale bilder og fysisk film, scannet inn i Fotostation og lagret og indeksert der. Dette gjør at artikler blir indeksert av to formater, det arkivet har i SIFT systemet og det Internett bruker når et utvalg av artikler legges ut på Internett. Dersom fysisk film er brukt i avisen blir denne indeksert både i SIFT systemet og i Fotostation, mens digitale bilder kun blir lagret i Fotostation.

Dersom du ønsker å knytte et bilde til artikkelen gjøres det i word, bildet følger da artikkelen helt til den står på trykk i avisen. Når artikkelen importeres til SIFT, blir hele artikkelen i råtekst overført, men ikke de samme metadata som ble påført i word. I overføringen er det ikke samsvar i metadatafeltene, f.eks TITTEL påført i word blir til STIKKORD i SIFT formatet. Her påfører arkivet metadata som emneord, fotograf, hvilken illustrasjon det er til saken og også merknads elementet, de øvrige metadata som produksjonsdato, artikkelnavn, produktet artikkelen har stått i samt artikkelen størrelse i form av antall linjer blir automatisk generert ved import fra CCI systemet.

Når artikkelen er ferdig skrevet i CCI WORD ligger den i CCI systemet hvor den så blir import over til en mellomdatabase som sender artikkelen videre. Alle artikler blir lagret her, og alle importes også over til SIFT. Videre går artikkelen til HTML databasen hvor alle artikler blir konvertert til HTML kode. Alle metadata som er påført blir med til CCI systemet også fra Fotostation. Dette blir gjort for at det skal være lettere for Internettredaksjonen å foreta et utvalg av artikler som skal publiseres på web og som er ferdig konvertert til HTML. Vedlegget til artikkelen er som oftest et bilde og er lagret som eget objekt i basen, med referanse til artikkelen som det er en del av. Bildet blir i tillegg lagret som eget objekt i Internettdatabasen for Internettutgaven. De artikler som utvelges og publiseres i Internettutgaven av Adresseavisen blir indeksert med form-atet de benytter her og lagret i en ny database.Denne Internett databasen er det du som Internett bruker har tilgang til når du søker via Adresseavisens hjemmeside, da du ikke har tilgang til å søke i hele SIFT databasen som de ansatte har mulighet til via Intranett.

Dersom en artikkel i tillegg til å stå i papirutgaven av avisen, også kommer i Internettutgaven, er det altså lagret i 4 databaser. Dette er HTML databasen, SIFT, CCI systemet (som har en database) og Internettdatabasen.

De ansatte søker mot SIFT databasen via ett Web-grensesnitt i Intranettet. Tidligere søkte de ansatte fra et kommando grensesnitt i UNIX mot SIFT. Dette kommandogrensesnittet benyttes fremdeles av arkivansatte. Det er utviklet et web-grensesnitt for alle de tre arkivene, d.v.s stoffarkivet, filmarkivet og biblioteksdatabasen. Alle disse er surrogat av sine opprinnelige metadataformat.

Arkivet forholder seg strengt til indekseringsregelen HVEM, HVA, HVOR, NÅR og HVORFOR når de indekserer. Ved utfylling av emneord forholder en seg strengt til en liste på 55 emenord, altså bruk av et kontrollert vokabular her. Nye emenord utover listen kan ikke brukes før det er diskutert i fellesskap med resten av arkivpersonalet. Merknadsfeltet her fungerer som et felt for frie emneord, der du forholder deg til en liste men hvor du i større grad kan legge til nye ord selv, f.*eks under emnord UTDANNELSE legges det ord som skoler, lærere, mobbing under feltet* ***/merknad/***.

En problemstilling kan eksemplifiseres med dette scenariet her:

Scenario:

1. Du skriver en artikkel i Word
2. Etter at artikkelen er ferdig skrevet ønsker du å knyttet et bilde til artikkelen.
3. Du leter i mapper i Fotostation både i egne bilder og NTB, AP og finner et bilde du ønsker å ha.

4. Du knytter bildet til artikkelen
5. Artikkelen er ferdig, men før du sender den til produksjonssystemet (CCI) setter du på meta-data som nevnt for Word.
6. Noen dager senere skal du skrive en artikkel som følger opp denne artikkelen og du går i SIFT for å søke via web-grensesnittet.
7. Du husker at tittel på dokumentet er "Internett i bilen" men du har ikke mulighet til å søke på tittel i web-grensesnittet. Her må du søke som søkeord1 eller søkeord2.

Årsaken til at vedkommende ikke kan søke på tittel er at innholdet i metadataelementet tittel fra WORD blir lagt i stikkord elementet i SIFT ved import av artikkelen fra CCI.

Ideen her er at kunnskapen du har om metadata i Word skal kunne videreføres til til de andre systemene, for samsvar mellom arbeidsprosesser. I praksis betyr dette at du som bruker av systemene skulle kunne søke på tittel for en artikkel både i SIFT, Internettdatabasen og CCI sys-temet.

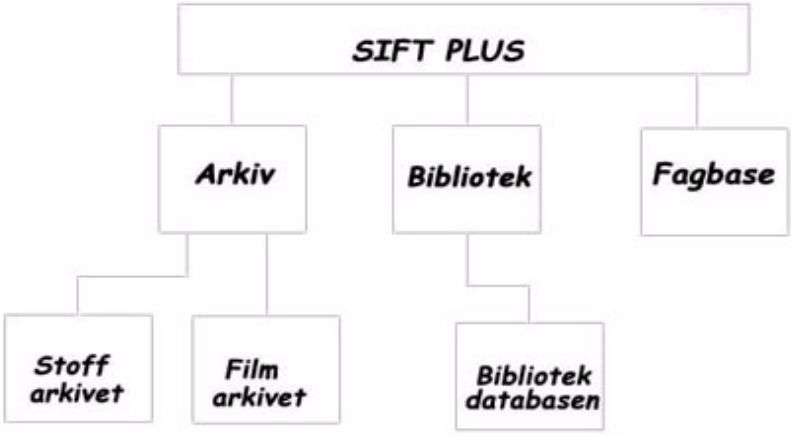
6.3. Introduksjon til SIFT - databasen

SIFT (Searching in free text) er et omfattende IT-system med mange muligheter tilpasset den enkelte brukskonteksten. Jeg vil ikke ta for meg en teknisk beskrivelse av systemet ut fra en systemutviklers interesseprofil, men presentere systemet ut i fra brukerens muligheter i sys-temet når det gjelder søking og browsing etter informasjon. Videre fokuseres det på hvordan SIFT sine ulike metadataformater brukes. SIFT er utviklet av Statens datasentral tidlig på 80 tallet, og et et system som også brukes av Dagbladet og Stavanger Aftenblad samt en rekke andre offentlige organisasjoner som NHO, Utenriksdepartementet og Næringsdepartementet, Kyst direktoratet, Stortingsbiblioteket m. fl. SIFT ble utviklet med hovedhensikt å støtte gjenfinning av informasjon i fritekst, da det var stor etterspørsel på denne tiden fra ulike medier nettopp etter et slikt system. Systemet håndterer rent tekstlige objekter, strukturerte og formaterte data og de fleste kombinasjoner av disse. SIFT systemet finnes i både kommando grensesnittutgave og webgrensesnittutgave hvor Adresseavisen benytter begge grensesnittene.

Adresseavisen bruker en database som heter SIFT Plus og som tilbyr funksjonene:

- søking
- lagring
- registrering

og er organisert i databaser etter et hierarki som figur 6.2 viser. I en SIFT plus database vil dokumenter registrereres i skjemaer, det jeg omtaler som metadataskjemaer.



FIGUR 6.2. SIFT Systemet

Adresseavisen har alle disse modulene forutenom modulen for fagbasen. De enhetene eller informasjonsobjektene som legges inn i basene kalles for dokumenter. En registrering av et dokument betraktes som en enhet. Vi ser nærmere på de ulike databasene. SIFT operer med tre modus, spørremodus, kommandomodus og skjemamodus. Spørremodus er bruk av ordbok, kommandomodus er søkemodus og skjemamodus er for registrering. Av disse modusene er det søkemodus som blir videre presentert her i avsnitt 6.3.4.

6.3.1. Stoffarkivet

I stoffarkivet blir alle artikler og annet tekstlig stoff som har stått på trykk i Adresseavisen registrert og indeksert. Indekseringsformatet for stoffarkivet er utarbeidet av Statens Datasentral som en "standard" for artikkelindeksering for aviser. Dette formatet er dermed spesialtilpasset tl avisers behov og ikke spesielt til Adresseavisen, men er allikevel å regne som et ad-hoc format. Andre aviser som benytter dette formatet er Stavanger Aftenblad og Dagbladet.

Stoffarkivet startet registrering av stoff på data i 1982 (Prime systemet), men i 1993 ble SIFT innført som databasesystem. Alt stoff som ligger i basen fra 1993 er indeksert, mens stoff tilbake til 1982 er uindeksert. Alle enheter som lagres i SIFT er organisert i databaser etter årstall, dvs at du må åpne de årstall som er av interesse dersom du ønsker å søke paralellt med flere databaser

Denne databasen har over 296276 enheter lagret pr. i dag og antallet vokser med vel 30-50 indekserte enheter pr. dag. Artikler og annet stoff er her lagret i fulltekst i databasen. Dersom du skal ha tilgang til artikkelen slik det sto på trykk i avisen, lagres aviser opptil et år etter at artikkelen sto på trykk. Etter ett år finnes den på microfilm.

6.3.2. Filmarkivet

I filmarkivet har vi to indekseringsformater, et for fysisk film og et for illustrasjoner som kan være dias, tegninger, plakater m.m. Alle enheter både for film og illustrasjoner blir regnet som en del av filmarkivet. Samlet utgjør filmarkivet over 91194 enheter pr. idag, og tilveksten er på 30-40 indekserte enheter pr. dag. Formatene som benyttes her regnes som ad-hoc formater, da de er behovsstilpasset i samarbeid med arkivleder på Adresseavisen, samt en representant for Statens Datasentral. Det som er spesielt med filmarkivet er at svært mange av postene i film-arkivet har referanse til et fysisk objekt. Postene i filmarkivet er kun metadataobjekter. Film-arkivet består av en enhetlig database.

6.3.3. Biblioteksdatabasen

Biblioteksdatabasen inneholder enheter som utgjør Adresseavisens bibliotekdatabase. Det omfatter bøker, kart og andre dokumenter fra de ulike biblioteksenhetene rundt om på Adresseavisen. Disse er spredt rundt på de ulike avdelinger, og også plassert på kontoret til enkelte ansatte. Biblioteksdatabasen har 4732 indekserte enheter pr. idag. Tilveksten av dokumenter til biblioteksdatabasen kan variere fra 4 dokumenter i 1998, i forhold til hele 184 dokumenter indeksert i den første måneden av år 2000. Det indekseringsformat som benyttes for biblioteksdatabasen er utformet på grunnlag av NORMARC formatet, og har dermed sitt grunnlag i et standarisert format. På grunnlag av dette definerer vi ikke dette som et ad-hoc format, og denne databasen sitt format blir dermed ikke gjenstand for mapping i denne oppgaven. Siden formatet bygger på NORMARC formatet og BIBSYS MARC allerede har foretatt mapping mellom DC og BIBSYS MARC burde det ikke være store endringer knyttet til en mapping fra dette formatet over til DC. [se kapittel 5 for mer informasjon om DC]. Mange offentlige institusjoner som NHO, Utenriksog Nærings-departementet, Stortingsbiblioteket m.fl. bruker SIFT biblioteksmodul. For mer info om SIFT biblioteksmodul se SDS sin brukermanual.

6.3.4. Søkemuligheter i SIFT

SIFT basen tilbyr mange og avanserte søketjenester. Disse mulighetene kan deles opp i 4 grupper:

1. SIFT vanlige søkespråk. Kan brukes både i søkemodus og skjemamodus. Egner seg til det meste.
2. Søkeskjemaer. Her har du tilgang til hele formatet og fyller ut de feltene du ønsker å se på. Forskjellen her er at du slipper å huske de ulike metadatanavnene før du gjør et metadatasøk. Du får et grafisk grensesnitt på søkene dine. Dette egner seg derfor best for sporadiske brukere av katalogen.
3. Makrosøk. Her definerer du ferdige søkeoppsett. Egner seg for uttak av lister med spesielle utvalg, og kompliserte søk som foretas jevnlig.
4. Nøkkeloppslag (spesialbruk av søkeskjema). Egner seg for gjenfinning av katalogposter som ennå ikke er indeksert.

Arkivet benytter seg av makrosøk når de produserer dagsrapporter over hvor mange enheter som er blitt registrert og indeksert i stoffarkivet og filmarkivet. Dette er det kun behov for ved spesialiserte arbeidsoppgaver, og SIFT vanlige søkespråk er det som er mest aktuelt å benytte for den gjennomsnittlige bruker av SIFT.

I SIFTs vanlige søkespråk kan du benytte deg av fritekstsøk eller metadatasøk.

6.3.4.1.Metadatasøk

Når det gjelder metadatasøk er det er mulig å definere alle element i de ulike metadataformat som benyttes søkbare i SIFT, selv om ikke dette er gjort for alle de som bruker systemet. Det er vanlig at en definerer hvilke felt som skal være søkbare etter behov. Ved presentasjon av metadataformatene vil jeg komme tilbake til hvilke metadataelement som er gjort søkbare i Adresseavisens formater. Du kan også her kombinere de boolsk operatorene i søkene dine.

Eksempel 6.1 Eksempel på metadatasøk:

finn *(Hansen i fotograf)* ***OG*** *(forbrytelser i emneord)*

eller

finn *fotograf=Hansen* ***OG*** *emneord=forbrytelser*

finn *(Antonsen i fotograf)* ***OG*** *(hansen i journalist)* ***OG*** *(prod-dato=20000404)*

eller

finn *fotograf=Antonsen* ***OG*** *journalist=hansen* ***OG*** *20000404 i prod-dato*

I kommandogrensensnittet forutsetter dette at du er godt kjent med de felt som er mulig å søke på og vet litt om hva du leter etter. Dersom du har noen opplysninger som fotograf, dato det sto kan du spesialisere søkene dine og veldig konkret definere hva du er ute etter. Dette gjør at trefflisten som regel blir betydelig mindre. I mindre databaser med få registrerte dokumenter, kan godt fritekstsøk gi like godt resultat som metadatasøk. I web-grensesnittet blir du mer invitert til å søke på metadata, da elementene **"/dato/", "/journalist/", "/emneord/"** er mulig å søke på her som du kan se av figur 6.10. og 6.11.

Kombinert med fritekst blir det slik: ***finn*** *(Hansen* ***i*** *fotograf)* ***OG*** *Anne Lise Ryel*

Når du har gjort dine søk har du også mulighet til å velge hvilke metadatafelt du vil at trefflisten skal vise. Her har Adresseavisen sin ferdige liste over element som vises, som vises med kommandoen ***vis felt.*** Ved å skrive

vis felt doknr, element 1, element 2, element 3 osv

kand du definere din egen liste som viser deg akkurat de felt som gir deg den informasjon du ønsker.

Ut over dette har du mulighet til å sammenligne verdier i metadatafeltene som største og minste verdi m.m. Du kan også bruke en assosiativ operator, som *f.eks finner dokumentet med en gitt tema og velger ut kun de med dette tema som har samme forfatter.*

6.3.4.2.Fritekstsøk

I fritekstsøk benytter du deg av de boolske operatoene" OG", "ELLER" og "OG IKKE". Fritekstsøk kan deles opp i gruppene:

1. Søke på en term, *f.eks finn* ***bil***
2. Søke på flere termer, *f.eks finn* ***bil*** *OG* ***hytte.***
3. Søke på sammensatte uttrykk, *f.eks* ***finn*** *vernet arbeid,* ***finn*** *Gro Harlem Brundtland*
4. Flere termer i kombinasjon med trunkering. *F.eks* ***finn*** *bil** ***OG*** *vernet arbeid.*(* er trunkering-stegnet)
5. Søke på en term i samme avsnitt som et annet. *F.eks* ***finn*** *narkotika* ***avsn*** *idrettsutøver.*

Når du søker i fritekst i SIFT søkes det mot alle råtekst og alle metadatafelt i hver registrerings-enhet. Det angis ikke hvilket felt man ønsker at søketermene skal forekomme i, og oppfører seg

etter prinsippet "Exist or not Exist" når dokumenter gjennomløpes på leting etter de termer du søker på.

6.3.4.3. Andre tjenester til hjelp i søkeprosessen:

I tillegg til vanlige navigeringstjenester har du også tilgang til disse tjenestene:

- Listing av tilgjengelige databaser.
- Listing av metadatafelt
- Tilgang til ordbok for hjelp til å finne ord å søke på innenfor et felt.
- Markere søketermer ved treff
- Sortere trefflisten din etter det metadatafelt du ønsker, *f.eks sak eller dato*.

Du har mulighet til å definere din egen sekvens av databaser, slik at dersom det er databaser du bruker ofte kan du innkludere dem i en sekvens og ved da å skrive : åpne sekvens1 som er definert til å være database for 1999 og 2000. Dette kan være til hjelp i stoffarkivet hvor databasene er organisert etter år, og hvor du kan definere sekvenser som søker i henholdsvis 2, 3, 4 og 5 år tilbake, helt til du finner det du er ute etter. Slike sekvenser er definert i web-grensesnittet. Dette vil ikke ha noen betydning for slik filmarkivet er organisert som består av en enhetlig database.

6.3.5. Adresseavisens Metadataformater:

Adresseavisen har 6 metadataformater som er :

1. SIFT format for indeksering av fysisk film
2. SIFT format for indeksering av artikler/stoff til avisen.
3. SIFT format for indeksering av illustrasjoner (lysbilder, plakater, tegninger m.m)
4. SIFT format for indeksering av bøker, og andre dokumenter til biblioteket
5. Fotostation sitt format for indeksering av digitale bilder
6. Internett redaksjonen sitt format for indeksering av mulitmediadokumenter (artikler med vedlegg)

I tillegg har vi tre surrogatformater av SIFT sine metadataformat som er brukt i WEB-grensesnittet. Disse er :

1. SIFT web-format for indeksering av fysisk film

2. SIFT web-format for indeksering av illustrasjoner (lysbilder, plakater, tegninger m.m)
3. SIFT web-format for indeksering av bøker, og andre dokumenter til biblioteket.

Web grensesnitt 1 og 2 blir presentert i eget avsnitt senere i kapittelet.

6.4. Adresseavisens dokumenttyper

Adresseavisen har 5 elektroniske arkiv, som er biblioteksarkiv, filmarkiv, stoffarkiv, billedarkiv og arkiv for Internettutgaven. Noen av dokumentene er lagret i flere arkiv, noe som fører til redundans. I disse arkivene er følgende dokumenttyper lagret:

- fysisk film i SIFT filmarkiv
- artikler/stoff i SIFT stoffarkivet og Internettarkivet
- bøker og kart i SIFT bibliteksarkiv
- digitale bilder i billedarkivet Fotostation
- illustrasjoner, dias, tegninger m.m i SIFT filmarkiv

6.5. Definering av arbeidsprosesser og informasjonsprosesser i Adresseavisen

Her omtales de ulike arbeidsprosesser ved redaksjonen i Adresseavisen og ved arkivet, som er de avdelinger som er representert i undersøkelsen. Arkivet sine arbeidsprosesser kjenner jeg veldig godt til og disse er beskrevet i detalj. De øvrige avdelingers arbeidsoppgaver blir presentert mer generelt.

6.5.1. Arbeidsprosesser ved Arkivet

Arkivet betjener kunder eksternt og sine egne ansatte via telefon og mottak i skranken. Ekstern kundebetjening er hovedsakling bilder til media og privatpersoner, samt artikler / bakgrunnstoff til lokale saker. I tillegg er det en del forespørsler på microfilm. Internt er det hovedsaklig journalistene som benytter seg av arkivfunksjoner. Å betjene kunder og ansatte innebærer

indekseringsprosesser og søkeprosesser. De indekserer bøker til biblioteksarkivet, artikler til stoffarkivet, og film og illustrasjoner til filmarkivet. Arkivet er delt opp i to seksjoner, der noen ansatte arbeider med stoffarkivet og andre med filmarkivet. De som arbeider ved arkivet bruker SIFT kontinuerlig hele dagen i alle sine arbeidsprosesser.

Arbeidsprosesser:

- Ta imot bildebestillinger fra kunder eksternt og internt
- Lete opp bilder og artikler fra arkivet som jornalister spør etter
- Finne bakgrunnsstoff om en sak på oppdrag fra journalister
- Finne bakgrunnsstoff om lokale saker, lokale personer fra andre medier i norsk presse
- Indeksere alle artikler som har stått på trykk i avisen
- Indeksere alle filmstriper som har vært brukt i forbindelse med en sak i avisen
- Indeksere alle digitale bilder som er brukt i avisen i Fotostation
- Indeksere bøker til biblioteksdatabasen når nye bøker og andre dokumenter kjøpes inn

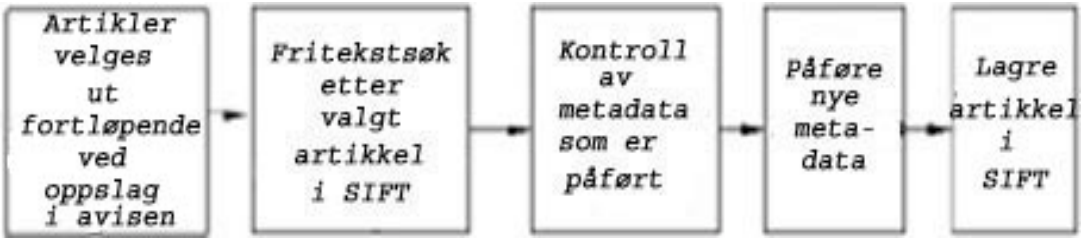
Dette er et utvalg av arbeidsprosesser og det er umulig å nevne alle arbeidsprosesser et arkiv vil ha, da det nesten ikke er grenser for hvilke oppgaver de kan bli satt til å gjøre. Av de arbeidsprosesser som er nevnt over kan disse deles inn i 2 hovedoppgaver:

1. Søk etter informasjon i SIFT og ikke-digitaliserte samlinger
2. Indeksering av dokumenter i SIFT og Fotostation

Søk etter informasjon i SIFT kan være hva som helst av søkeoppgaver og blir for omfattende å gå inn på her. Indekseringsprosessene i SIFT og bildebestilling blir presentert her i de følgende avsnitt.

6.5.1.1. Arbeidsprosess: Indeksering av artikler/stoff

Artikkelen kommer som import fra CCI systemet i råtekst og arkivet sjekker her om metadata som er påført er riktig, før de påfører metadata om bidragsyter og setter på emneord og eventuell merknad, som skissert i figur 6.3



FIGUR 6.3. Indekseringsprosessen for artikler/stoff ved arkivet

6.5.1.2. Arbeidsprosess: Indeksering av fysisk film

En registrert enhet i skjemaet for fysisk film referer til et fysisk informasjonsobjekt hvor tre filmstriper blir lagret pr. enhet. Film fra fotografer som er brukt kommer i plastlommer klippet opp i filmstriper. Selv om det er brukt bare et bilde i forbindelse med en sak, lagres det 3 striper, dersom disse er tatt i forbindelse med saken. Overskuddet går tilbake til fotografen. Det er gunstigst å lagre hele filmstriper, og dersom det er brukt flere bilder som går over flere enn 3 filmstriper, vil saken få 2 enheter i det fysiske arkivet. Saken blir da en serie, hvor flere enn 3 filmstriper er lagret. Hvor mange enheter en sak kan ha er avhengig av hvor mange filmstriper som er brukt. Det er sjelden at en sak har brukt bilder fra mer enn 3 fysiske filmstriper lagret på en filmmappe. Hver fysiske filmmappe har en unik identifikator i form av et femsifret tall. Disse film-enhetene har et hode som tallet som er identifikatoren blir hullet ut i ved hjelp av et fysisk verktøy. Filmmappene blir så satt inn i arkivskuffer hvor mappene henger i kolonner med filmstripene hengende loddrett. Tallet for identifikatoren blir også skrevet med tusj på selve mappen, slik at identifikatoren også er synlig for det blotte øyet. Over disse film-mappene er det en "leser" som du drar over de taggete hodene til filmmappene og som leser identifikatoren til filmmappene, *F.eks er identifikatoren på mappen din 54434 gir du dette tallet inn i "hodeleseren" (som består av 5 hjul med tall med verdi 1-9).* Når du da har gitt inn tallet 54434 dras "leserhodet" (som er i enden av arkivskuffen) over alle "hodene" til film-mappene i arkivet. Filmmappen med tallet 54434 vil da stikke litt opp fra alle de andre film-mappene som er lagret. *F.eks når du da søker opp en sak i SIFT filmarkiv, og videre skal hente filmen i det fysiske filmarkivet er alt du trenger de 5 siste sifrene av elementet **/film-id/**. **/Bilde-nr/** og **/motiv/** hjelper deg deretter å finne frem til den aktuelle filmstripen på filmmappen, hvor bildet du er ute etter er. Det kan også være at du ser etter bilder som er egnet til gjenbruk.*

6.5.1.3. Arbeidsprosess: Indeksering av illustrasjoner m.m

Illustrasjoner blir lagret fortløpende dersom de har stått på trykk. Når illustrasjonene samtidig blir lagret ved arkivet, *f.eks i forbindelse med en filmanmeldelse kan det godt hende at samme illustrasjon blir lagret som to enheter i filmarkivet.* Dette er fordi hver illlustrasjon som er brukt i forbindelse med en filmanmeldelse, tv-serie o.l. ikke blir beskrevet hva motivet er (fordi dette er saker som Adresseavisen sjelden selv har rettighetene til å bruke fritt, da det ofte dreier seg om byråbilder som NTB og AP) og heller ikke har en fysisk identifikator. Arkivet for illustrasjoner er dermed mer enn referansearkiv om hvor illustasjonen kom fra, hvilken dato den sto på trykk i avisen og i hvilken kontekst den var brukt. Illustrasjonene blir arkivert etter emnegrupper fra en emnegruppeliste på 55 ord som arkivet går etter. Hver illustrasjon får ikke sin unike identifikator slik at du kan søke opp nøyaktig det bildet som er brukt i tidligere saker. Dette skiller indeksering av illustrasjoner i forhold til indeksering av fysisk film og artikler, hvor hver artikkel og hver filmmappe har sin unike identifikator. Illustrasjoner som enhet lagret i databasen har en **/ill-id/**, som blir generert automatisk ved indeksering av en ny post. Det kan godt være at en og samme illustrasjon har mange ulike **/ill-id/** i basen, fordi hver illustrasjon nettopp ikke har en unik identifikator. Ved Adresseavisen er det bare tegninger av Adresseavisens tegner som har en unik identifikator. Identifikatoren for dette bildet (som består av en tallkode) blir lagret i **/merknad/** selementet, Illustrasjonene blir deretter lagret i en arkiveske for tegninger. Motivet til tegningene blir beskrevet i feltet for sak i metadataskjemaet. Dette viser at det er behov for element for unik ID og beskrivelse av motivet for enkelte typer illustrasjoner.

6.5.1.4. Arbeidsprosess: Bestilling av bilder

I denne sammenheng må bilder som kunden er interessert i søkes opp i SIFT basen, eventuelt andre baser. I SIFT blir det da notert "utlånt til bestilling" og dato for mottakelse av bestillingen legges inn. Deretter skriver en bestillingen ned på et bestillingsskjema. Bestillingsskjemaet legges inn til fotolaben for fremkalling av bildet, sammen med filmen bildene er på. Når filmmappen kommer i retur søkes det opp på **/film-id/** i SIFT basen, og utlånskommentaren fjernes. I de tilfeller hvor det er bestilling fra digital film er det billeddesken som utfører eksponering av bildene.

6.5.2. Arbeidsprosesser i redaksjonen

Her vil jeg ta utgangspunkt i de avdelinger som er representert i undersøkelsen over hvilke hovedoppgaver som inngår i de ulike avdelinger av redaksjonen, dvs Kultur, Internettredaksjonen, Desken og Nyhet/Sekreatariatet.

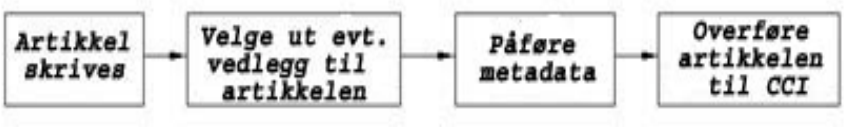
Desken er det området i redaksjonen hvor journalistene sitter og lager neste dags avis. Der sitter vaktsjef og flere redigerere, vaktsjefen har ansvaret for hele avisen, mens de journalistene som redigerer stoffet har hver sine sider i avisen som de har ansvaret for. Typografene sitter også på desken, de plasserer stoffet på sidene med tekst og bilder. Billeddesken er bemannet med fotografer og en fra arkivet, som er behjelpelig med bilder til neste dags avis. Desken er altså det området i reaksjonen som har ansvaret for at avisen går i trykken og er ferdig til neste morgen. Redigerer har på samme måte som arkivet en type indekseringsarbeid siden de setter på metadata.

Journalister og redigererere arbeider med artikler i CCI Word, hvor de på mange måter på lik linje med arkivet har en arbeidsprosess hvor de indekserer noe av artikkelen, som senere blir overført til avisproduksjonssystemet CCI. Her påfører de metadataene:

- tittel
- undertittel
- ingress
- tekst
- mellomtittel
- bildetittel
- bildetekst
- bildesignatur
- vignett
- undervignett

hvor alle untatt vignett, undervignett, undertittel og mellomtittel blir overført som råtekst til SIFT, pluss metadata som dato m.m som genereres automatisk. Metadataene bildetekst, bildetittel, og bildesignatur havner i **/tekst/** i SIFT, tittel i **/stikkord/**. Arbeidsprosessen i CCI Word illustreres av figur 6.4.

På alle avdelinger unntatt arkivet sitter det journalister som produserer stoff til avisen, som regel innen for sin avdeling, men også innen andre områder i avisen. De skriver sine artikler i CCI word. Internett indekserer også artikler som de legger ut på Web'en etter Internettformatet som presentert i avsnitt 6.8.



FIGUR 6.4. CCI Word Arbeidsprosess

Stort sett ligger informasjonsbehovet på fakta og bakgrunnsopplysninger om saker, samt sjekke kilder, skrivemåter m.m i forbindelse med skriving av artikler til avisen etc. En journalist i sine arbeidsprosesesser trenger ikke bare SIFT til å samle informasjon, men må ofte søke andre kilder, samt samle inn informasjon via intervju og observasjon. Fremgangsmåte her vil være individuell, og jeg går ikke nærmere inn på disse.

Den enkeltes individuelle fremgangsmåte i søkeprosessen vil være påvirket av faktorer som Hannele Fabritius [12] viser til i sitt prosjekt og som er omtalt i kapittel 4. En person som arbeider ved kulturavdelingen, og stort sett skriver film og spill anmeldelser har ikke samme type informasjonsbehov i forhold til en som skriver utenrikspolitiske saker. Vi ser da at den enkeltes søkeoppførsel er påvirket av den avdeling du tilhører sin kultur, den journalistiske arbeidspraksisen innen denne avdelingen og den type "skrivejobb" du skal gjøre. Dette er konsept 3, 4 og 5 i Hannele Fabritius sin modell [figur 4.2] over journalistisk søkeoppførsel. Variasjoner i journalistisk søkeoppførsel hos den enkelte vil ligge innenfor denne modellen.

6.6. Presentasjon av Adresseavisens format for informasjonsobjekter i SIFT

Nedenfor følger hver av formatene med forklaring til hvordan metadataelementene brukes, og et eksempel følger for hvordan en enhet lagret i SIFT vil bli presentert.

Følgende metadataskjema blir introdusert i de følgende seksjoner:

1. Metadataskjema for indeksering av artikler
1. Metadataskjema for indeksering av illustrasjoner
2. Metadataskjema for indeksering av film

6.6.1. Metadataskjema for indeksering av artikler

Metadataelementene **/skjermes/retur-adr/** er helt nye tilleggsfelt som er blitt lagt til formatet etter at undersøkelsen ble gjennomført. Elementet **/navn/** erstatter elementet **/art-nr/**, og det er **/art-nr/** som er referert til i undersøkelsen. Elementene presentert i tabell er slik formatet eksisterer pr. idag, hvor eventuelt gammelt felt er angitt i parantes.

Tabell 6.1. Metadatasett for indeksering av artikler

METADATA NAVN	FORKLARING
Skjermes	Her skrives verdien JA eller NEI dersom det stoffet skal eller skal ikke skjermes fra innsyn.
Prod – dato	Genereres automatisk ved overføring til SIFT.
Navn **(Art-Nr)**	Her legges navnet på artikkelen inn. Dette feltet erstatter det gamle feltet Art-nr, og er den unike indentifikatoren til artikkelen. Navnefeltet består kun av tekst og kan ikke forstå tall informasjon. Viktig å være oppmerksom på at dette feltet er endret etter at spørreundersøkelsen ble gjennomført, og det elementet som referes til der er **/Art-nr/** og ikke /**Navn/**.
Ant-linjer	Genereres automatisk ved overføring til SIFT.
Produkt	Her har vi Adr'ut (Ut-magasinet), Adr'Lille (Lille-Sportsavis), Adr'Uke (Uke Adressa), Adr'Nyhet, Adr'Kultur (kulturbilaget) m.m. Dette er sub-bilager som utgis i Adresseavisen. ADR oppgis dersom det ikke står under noen av de andre produktene av Adresseavisen.
Illustrasjon	Dersom det er et bilde som er brukt skrives det bare inn FOTO, ellers skrives det inn illustrasjon, tegning m.m. Type illustrasjon vi snakker om.
Kilde	Her skrives det inn hvor illustrasjon kommer fra, dvs hvilken organisasjon som, *f.eks "EGET ARKIV", "AP", "NTB" etterfulgt av navnet på opphavspersonen av illustrasjonen. Dersom det er illustrasjon som er returnert skrives kun inn navn på fotografen.*
Original	Ikke i bruk av Adresseavisen.
Journalist	Navnet på journalisten som har skrevet artikkelen.
Emneord	Her legges det inn kontrollerte emneord som er utarbeidet. Noen nye er kommet til etter hvert på listen. Over 30 Emnegrupper totalt.
Merknad	Felt som brukes til å beskrive artikkelen ytterligere for å kunne finne den igjen. *F.eks. andre mulige søketermer, noe som er spesielt med artikkelen m.m.*
Stikkord [a]	Må være veiledene for artikkelen dreier seg om. Det er nemlig kun dette feltet som vises på trefflisten etter et søk i databasen. Desto mer treff du har på et emne, bør denne være så pass beskrivende at du slipper å gå inn p alle andre ulike treff dersom de er urelevant. Denne benyttes ikke til frie emneord slik den er tenkt, men tittel fra CCI blir overført hit. Dette er forvirrende, med hensyn på betydningen av stikkord. Stikkord er stikkord, og ikke tittel. Følgelig skal ikke tittel forekomme her.
Tekst	Artikkelen i fulltekst, med tittel og det hele.

a. Ikke søkbart ved metadatasøk i Adresseavisen

Legg merke til at dette formatet ikke har tittel felt. Tittelen legges på første linje i tekst feltet. Årsaken til hvorfor denne vurdering er tatt, i og med at tittelelementer blir påført i WORD er ikke kjent. En tittel i avissammenheng er ofte veldig generell, og trenger ikke nødvendigvis representere innholdet i artikkelen. Allikevel kan tittel være nyttig å søke på dersom det er den du husker. Thorbjørn Wale [38] sier at tittel i aviser har to formål:

- Å fortelle hva meldingen eller artikkelen inneholder
- Gi leseren lyst til å lese det som står i artikkelen

Ut fra dette ser vi at tittel skal informere om saken, og kan brukes i stedet for **/sak/** elementet eller brukes paralellt med **/sak/**.

F.eks dersom du har en artikkel som har stått på førstesiden av Adresseavisen, blir saken lagret som to enheter i SIFT. Dette vil si en registrering der saken er på førstesiden og en registrering der saken står på side 4.

```
2000ARK:artikler  Res 26 dok 1 av 112  Lin 1-20 av 20    Søkbart      , Over NumT
 Adresseavisen  tekstarkiv                                Artikler  IKKEINDEKSERT

Skjermes   : 1.........
Prod-dato  : 20000216
Navn       : ØieHøgsnes
Side       : 41                                          Produkt : ADRESSEAVISEN
Ant-linjer : 7
Illustrasj.: Nei  Kilde: .... ......................... Original: ............

Journalist : Sæther, Tore
Emneord    : ..........................................................................
Merknad    : ..........................................................................

Stikkord   : Øie/Høgsnes leder finalen
Tekst      : Svein Olav Øie og Bjørn Høgsnes er satt på jobben med å lede
             herrefinalen i håndball-cupen. Drammen og Viking er finalister,
             og kampen spilles i Oslo Spektrum lørdag 4. mars. Samme dag og
             samme sted spilles også kvinnefinalen mellom Larvik og
             Nordstrand. Der får vi finale-debutanter som dommere, Andre
             Hansen og Øistein Pettersen fra Østfold.
```

FIGUR 6.5. Indeksering av artikkel i SIFT stoffarkiv

6.6.2. Metadataskjema for indeksering av illustrasjoner

Tabell 6.2. Metadatasett for indeksering av illustrasjoner

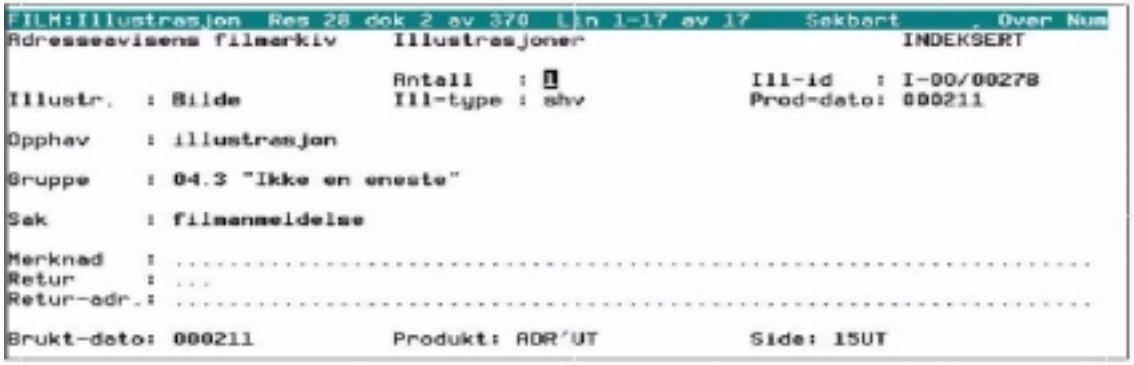
METADATA NAVN	FORKLARING
Antall	Du skriver antall illustrasjoner det dreier seg om. Dersom de kan skrives under samme sak lagres de samlet i basen. *F.eks har du en sak som har 2 tegninger som illustrasjoner, uavhengig av illustrasjonens motiv, blir disse lagret som 2 i dette feltet.*
Ill-id	Genereres automatisk
Illustrasjon	Her legges type illustrasjon inn. Er det et bilde, tegning, bok m.m. det gjelder.
Ill-type	Her legges det inn om bilde er i svart- hvit , farger eller om det er dias film.
Prod-dato	Dato illustrasjonen sto i avisen.
Opphav	Her skriver vi inn hvem som er opphavet til illustrasjonen. Er det et bilde tatt av NTB, og andre eksterne kilder lagres fotografens navn her, eller bare NTB. Er det fra privat personer skriver vi privat. Er det innsendt og ment at vi skal lagre det, eller ta vare på det, skrives det innsendt i feltet.
Sak	Saken illustrasjonen er brukt i sammenheng med.
Gruppe	Dersom illustrasjonen skal lagres i vårt arkiv, setter vi på gruppe. Dersom det er en illustrasjon som ikke er vårt eget, står dette feltet tomt. Dette er tilfelle med NTB, Ap sine bilder, og bilder, tegninger som er returnert til privatpersoner.
Merknad	Er det noe spesielt som må meddeles, *f.eks. at vi ikke har mottatt dette, eller at det mangler, skriver vi inn det her.*
Retur	Skriver hvor illustrasjonene er returnert, eventuelt med tlf-nr. og adresse.
Brukt-dato	Når illustrasjonen sto i avisen
Produkt	Hvilket produkt illustrasjonene sto i , var det ADR (hovedavisa), Adr'lokal (lokal (-avisa), Adr'lille (Sportsmagasinet), Adr'UT (UT-magasinet) m.m.
Side	Hvilken side i avisen illustrasjonen sto på.

/Brukt-dato/Produkt/side/ har utvidbare felt dersom illustrasjonen brukes flere ganger i forbindelse med en sak.

En kan umiddelbart se flere svakheter med dette formatet sett i forhold til god indekseringsskikk. Disse svakhetene må imidlertid ses i sammeheng med hvilken hensikt dette formatet har som

hensikt å tjene. Ved registrering av illustrasjoner i Adresseavisen er det HVEM som har illustrasjonen til saken, HVOR mange illustrasjoner som var til saken totalt, HVILKEN sak illustrasjonene sto til, NÅR illustrasjonene var brukt i forbindelse med saken, og eventuelt OM Adresseavisen har illustrasjonen lagret i sitt arkiv eller at illustrasjonene er returnert. Illustrasjoner som blir lagret i SIFT filmarkiv i dette formater er ofte illustrasjoner som Adresseavisen ikke har selv, eller rettighetene over. Ofte blir disse illustrasjonene returnert til opphavsmann, er til vedkommende som hadde saken, som fungerer som mellomledd til en evnetuell opphavsmann. Dette formatet brukes som en kontrollinstans dersom det skulle komme forespørsler til Adresseavisen om illustrasjoner m.m som er returnert til opphavsmann men ikke kommet frem. Da kan Adresseavisen se at illustrasjonene faktisk er returnert fra dem, og at da eventuelt forsendelsen er "tapt" i annet distribusjonsledd. Hva illustrasjonene faktisk beskriver er ikke relevant i denne sammenheng. For tegninger produsert av en av Adresseavisens egne tegnere, og som en ønsker å skille fra hverandre, blir påført en identifikator i form av entallkokde og deretter arkivert. Dette nummeret blir plassert i illustrasjonsformatet sitt element **/merknad/**. I dette formatet vil en ikke klare å skille de illustrasjoner som er brukt unikt fra hverandre. Dersom samme illustrasjon er brukt to ulike dager i forbindelse med en sak vil denne illustrasjonen få to registreringer i SIFT. Det er altså saken som blir identifisert med en enhet, mens illustrasjonen ikke blir det. Metadataelementet **/ill-id/** gir kun registreringsenheten fysisk i databasen en unik i dentifikator. I realiteten kan en og samme illustrasjon ha mange **/ill-id/** alt ut fra hvor mange ganger illustrasjonen har vært brukt i forbindelse med en sak. Hver sak har en unik **/ill-id/**.

Dersom det var illustrasjonen som hadde blitt unikt identifisert og ikke saken den hadde stått i forbindelse med, kunne en ha hatt utvidbare metadataelement for **/sak/** og **/brukt-dato/** slik at nettopp hver illustrasjon hadde fått en registrering i databasen. Dette ville da gjelde de illlustrasjoner Adresseavisen hadde i sitt eget arkiv, og ikke de illustrasjoner som blir returnert til opphavsmann. På grunn av indekseringspraksis har ill-id feltet ingen funksjon ved søk. Det er ikke mulig å ta en en illustrasjon fra en konvolutt, og deretter søke direkte på den unike identifikatoren, fordi illustrasjonen er bare utstyrt med dato den sto på trykk og emnegruppe. Det er derfor ikke definert at **/ill-id/** skal være mulig å søke på ved metadatasøk.



FIGUR 6.6. Indeksert illustrasjon i SIFT

6.6.3. Metadataskjema for indeksering av film

De fem siste elementene i tabellen som er **/bilde-nr/brukt-dato/produkt/side/motiv/** er gjentakende felt og gjentas for hver gang et fotografi på filmmappen er brukt. Saken blir dermed bare nevnt første gang fotografi fra filmmappen brukes. Selv om motiv elementet når fotografiet brukes for andre gang vil få samme innhold trenger dette strengt tatt ikke gjentas, men blir gjort av oversiktsmessige grunner for ikke å forstyrre rekkefølgen. Dersom saken er interssant for hver gang fotografiet blir brukt er dette ikke ideelt. I figur 6.3 ser du hvordan film organiseres i filmmapper, og i figur 6.7. ser du en indeksert enhet for en filmmappe i SIFT. Som oftes brukes det ett eller to fotografier fra hver filmstripe som lagres, dvs har det vært brukt 5 bilder i forbindelse med en sak og disse er fordelt på tre filmstriper, vil dette blir lagret i en registrert enhet som i figur 6.7, hvor da **/bilde-nr/ brukt-dato/produkt/side/motiv**/ vil være gjentatt for hvert Bilde-nr.

Elementene /**journalist/** og /**gruppe/** kan også gjentas dersom tema kan klassifiseres etter flere emneord, eller at det er flere enn en journalist som har skrevet artikkelen som er knyttet til bruk av filmen

Tabell 6.3. Metadatasett for indeksering av film

METADATA NAVN	FORKLARING
Jobb-id	Her settes dato og ett påbegynnende jobbnummer opp, *f.eks. 30.12.1998/500, osv.*
Film-nr	Det lagres tre filmstriper p.r. filmmappe. Dersom du har behov for å bruke mer enn enn filmmappe antyder du her hvilken av mappene du holder på med på den saken det gjelder. *F.eks. 1 av 2, 2 av 2, 1 av 3, 2 av 3 osv.*
Film-id	.Identifikatoren er et dato+ et nummer som velges fra en rull for hver ny filmmappesom registreres. Det kan se slik ut 30121998/44520, 30121998/44552. Identifikatoren fem sifret. Dette er et felt som brukes ved søking.
Film-dato	Dato filmen ble tatt (dvs når filmen var oppbrukt).
Film-type	Hvilken type film er brukt, er det farge neg, svh, eller farge pos.
Prod-dato	Den dato filmen og bildene fra filmen første gang var brukt i avisen.
Fotograf	Fotografen som har tatt bildene. Her ligger fotografene inne i rullemeny.
Journalist	Journalisten som skrev artikkelen filmen er knyttet til.
Gruppe	Her velges gruppe fra en liste med emenord, altså et kontrollert vokabular. f.*eks Emnegruppe legges inn som referanse for å vise at vi har emnegrupper som går på dette og er dermed en referanse til fotografier som finnes i papirform i konvoluttarkivet.*
Sak	Saken filmen er brukt i forbindelse med. Saken nevt her gjelder første gang filmer fra filmmappen er brukt.
Merknad	Dersom film mangler og er blitt borte skrives det her, og dersom den er utlånt til bestilling. Andre ting som skrives her er portrett, dersom det er portrett bilder på på filmen i stor grad, men dette er ikke like konsekvent gjennomført i dette feltet. Om det er portrett skal alltid nevnes enten her eller i motiv feltet.
Retur	Dersom filmen er blitt returnert vil det stå JA her, ellers NEI.
Retur-adr	Dersom verdien er JA i retur feltet vil Adressen til den det er returnert til stå her. Dersom adresse ikke er kjent, eller verdien i retur feltet er NEI vil feltet være tomt. Som oftest returneres film og eventuelt overskudd til fotografen.
Bilde-nr	Her skriver vi opp nr. på negativet på negativstripen, *f.eks. 1a-2, 2a-3, 4, 5-5a osv.*
Brukt-dato	Hvilken dato ble fotografiet brukt i avisen.
Produkt	I hvilket produkt ble fotografiet brukt.
Side	Sidetallet bildet sto på den datoen det sto på trykk. Sidetall følger produktet fotografiet sto i. *f.eks 7 UA som er side 7 i produktet UKE- Adressa.*
Motiv	En utførlig beskrivelse av motivet på bildet . Navn skal alltid nevnes, samt si om det er action bilde, portrett m.m. Har en begrensning på to linjer.

/Jobb-id/ fungerer som en teller og er bare et tall for å se hvor mange registrerte enheter du har

den dagen. Tallet består av dato pluss ett tresifret tall, som begynner enten på 900 eller 950 og telles oppover for hver registrerte enhet. Årsaken til at dette gjøres er at filmer til hovedavisen og filmer til f.eks Uke-Adressa, Ut-Magasinet osv registreres parallelt. Da starter noen på jobbnr 900 og teller opp og andre vedkommende starter på 950. **/Jobb-id**/ er like unik for den registrerte posten som **/film-id/**, og det hadde egentlig holdt å ha en av dem.

/**Film-dato/** fylles svært sjelden ut, da fotografene ikke oppgir når filmen ble tatt. Denne informasjon fylles ut når det er snakk om filmer som er fra det gamle filmarkivet, og da vil dato være før 1993. Det gamle filmarkivet er organisert etter dato og i dette tilfellet vil ikke filmmappen få en identifikator og bli lagret som er tilfelle med ny film.

En sak kan ha flere filmmapper. Dersom 4 fotografier er brukt og hver befinner seg på hver sin filmstripe, blir ikke filmstripene klippet opp, men alle 4 blir lagret. Hver filmmappe tar 3 filmstriper på hver filmmappe, og i dette tilfellet må det derfor indekseres en mappe til, som også får 3 striper. Dersom det skal lagres 3 filmstriper pr. filmmappe i slike tilfeller forutsetter det at det er film nok innenfor denne saken. I praksis vil dette si at dersom en har totalt 5 filmstriper som er tatt i forbindelse med saken, men hvor bilder fra 4 striper er brukt på trykk i avisen, så blir en mappe lagret med 3 filmstriper og en mappe med 2 filmstriper. Det kan aldri være en filmmappe som har to saker, siden den nye saken ikke blir nevnt når bildet brukes for andre gang. Dersom en og samme sak har hatt bidrag fra to ulike fotografer blir bidraget fra dem lagret i hver sin mappe. De er da å betrakte som en serie av filmmapper, da de følger samme sak.

I og med at bilder som kommer fra fysisk film også er scannet inn i Fotostation og indeksert og lagret, har Fotostation referanse til **/film-id/** til filmen som bildet er scannet fra. På grunnlag av dette hadde det ikke vært nødvendig å i tillegg indeksere filmen i SIFT som illustrert i figur 6.8. Figuren viser at flere bilder kan peke til samme fysiske filmmappe, fordi hvert bilde i Fotostation er et seperat objekt. Finner du et bilde i Fotostation vil du få referanse til den fysiske filmen og trenger dermed ikke gå innom SIFT for å få denne referansen, men direkte til det fysiske arkivet for tilgang til filmen. Det er snakk om at Fotostation på sikt skal erstatte SIFT, men i en overgangsfase lagres altså bilder fra fysisk film i to arkiv. Dersom Fotostation skal overta for SIFT forusetter dette at filmmappens unike identifikator legges i et eget element i Fotostation slik at det er direkte søkbart for arkivet. Dette er nødvendig når de skal fjerne utlånskommentarer for de filmer det gjelder når de kommer i retur til arkivet etter at de har vært utlånt til scanning eller bildebestilling. Slik det er nå utgjør /film-id/ og /bilde-nr/ i SIFT elementet /film-nr + motiv-nr/ i Fotostation.

6.6.3.1.Hvilken type informasjonen får du fra formatet:

Den registrerte enheten for beskrivelse av filmmappen, dvs posten i SIFT, har metadataelement knyttet til sak, selve filmen i filmmappen, og fotografiet hvor metadataene fordeler seg slik:

Sak: **/journalist/gruppe/sak/prod-dato/brukt-dato/produkt/side/**

Film: **/film-dato/fotograf/film-id/film-nr/film-type/retur/returadr/**

Fotografi: **/bilde-nr/motiv/**

Det du ikke får vite er sakbeskrivelsen til andre gang et fotografi blir brukt, dvs når **/bilde-nr/** får lik verdi som annet **/bilde-nr/** på filmmappen. Skal du vite mer om saken andre gang bildet er brukt må du bruke **/brukt-dato/** og **/side/** for å finne dette i stoffarkivet. Videre har du ikke direkte referanse til filmmapper som er fra samme sak, du vet bare om filmmappen inngår i en serie på 2 eller flere filmmapper som film-nr gir beskjed om.

Her ville det vært bedre når alle filmmappene likevel indekseres samtidig, og i stedet for å skrive dettte, har direkte referanse til den unike identifikatoren til de andre mappene som inngår i serien. Dette vil vi komme tilbake til under mapping i kapittel 8. Du ville da ha sluppet å søke opp de andre enhetene i serien, men gått rett til arkivet og hentet alle filmmappene. Dette kan konkretiseres med et eksempel:

Eksempel 6.2

Dersom en har 3 mapper i en sak, vil vi få to forekomster av relasjonselementet:

Filmmappe 1

film-id: F-000124/59091 .

relasjon: F-000124/54422

relasjon:F.-000124/54420

Filmmappe 2

film-id: F-000124/54422

relasjon: F-000124/59091

relasjon: F.-000124/54420

Filmmappe 3

film-id: F.-000124/54420

relasjon: F-000124/59091 (filmmappe 1)

relasjon: F-000124/54422 (filmmappe 2)

På samme måte kunne det har vært et gjentakende relasjonselement for hver gang et foto- grafi var brukt på nytt som referanse til artikkelen. I den gjentakende sekvensen i dette tilfellet kunne en ha lagt til et element som vist i eksempel 6.2. I tillegg er det ønskelig at relasjonselementet skal referere til artikkelen i SIFT Stoff slik:

Eksempel 6.3

Bilde-nr: 15A

Brukt-dato: 980530

Produkt: Adr'uke

Side: 7 UA

Motiv: Heidi Sørensen - portrett

relasjon: Artikkel 1

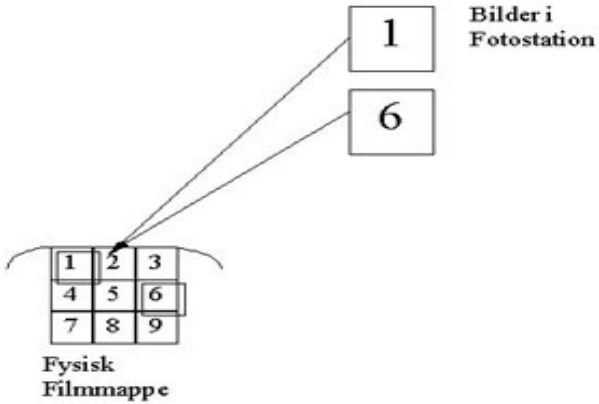
Denne sekvensen av element gjentas for hver gang et motiv i filmmappen er brukt i en sak

6.6.3.2. Hvilken informasjon etterspørres fra formatet:

Ofte vet en hva en er ute etter i forbindelse med en sak. Informasjon som etterspørres er hvor lenge siden det bildet med akkurat det motivet er brukt, og hvor mange ganger det er blitt brukt tidligere. En kan også ut fra motivbeskrivelsen vurdere i hvilken kontekst bildet egner seg, og få en assosiasjon til det ved å se hvilken type sak bildet først var brukt i forbindelse med. Videre trenger en å se ut fra posten om dette er en film som er lagret i eget arkiv, eller returnert til fotografen. Selv om et fotografi er nylig brukt kan likevel filmen være av eldre dato. Det kan derfor være interessant å se om filmen kommer fra "gammelt" filmarkiv. Dette kan du lese ut av **/film-dato/** elementet som da vil vise dato før 1993.

```
FILM:film  Res 27 dok 1 av 1075  Lin 1-21 av 27    Søkbart      , Over NumTast
Adresseavisens filmarkiv    Film                                  INDEKSERT

Jobb-id    : J-000215/911    Film-nr  : 01 av 01       Film-id: F-000215/59208
Film-dato  : 000213          Film-type: farve neg      Prod-dato: 000215

Fotograf   : Stensås, Christer
Journalist: Stensås, Christer

Gruppe     : 14.0 Hansen, Stian M - alpint
Gruppe     : 14.0 Thorland, Tor Egil - alpint

Sak        : KM i alpint i Vassfjellet
             ..........................................................................

Merknad    : ..........................................................................
Retur      : ..........................................................................
Retur-adr  : ..........................................................................

Bilde-nr   : 16      Brukt-dato: 000215  Produkt: ADR'LILLE        Side: 2 SP
Motiv      : Freidig-duo med seiersglis - Stian M Hansen (tv.) og Tor Egil
             Thorland                                              Papirkopi: ..
```

FIGUR 6.7. En registrert filmmappe i formatet for film



FIGUR 6.8. Løsning for referanse til fysisk film i Fotostation

6.7. Presentasjon av Adresseavisens metadataformat for digitale bilder i Fotostation

Digitale bilder er lagret i et system som heter Fotostation. Adresseavisen startet arkivering av digitale bilder i Fotostation 1.juni 1998. Fotostation er et produkt som er utviklet i Norge av Fotoware [15]. Metadatafeltene i indekseringsformatet kan lett fjernes og legges til alt etter behov, og med tilleggsproduktet Index Manager (søkemotoren) kan du gjøre hvilket som helst metadatafelt søkbart som du ønsker. I tillegg har du også mulighet til å endre navnene på de overordete overskrifter for hver gruppe av metadatafelter, som f.*eks Date & Time gruppen, men da med noe mer avanserte endringer i oppsett filer.* Fotostation kan på grunnlag av dette tilpasses det nødvendige behov som bedriften måtte ha, og trenger dermed ikke operere med overflødige indekseringsfelt som ikke gi logiske meninger ut fra deres hverdag. Når det gjelder publisering på Internett, er det lett å tilpasse bildearkivene i Fotostation til søking via Intranett og Internett. Nedenfor følger de inndelinger i metadata som Adresseavisen har organisert Fotostation formatet. Alle bilder/illustrasjoner osv som blir scannet og som står på trykk i avisen blir lagret i fotostation og arkivert av arkivpersonalet i egne digitale arkivmapper i Fotostation.

Min første tanke til dette formatet her er at de metadataelementer som ikke er i bruk for indeksering kan fjernes fra grensesnittet, siden dette kan være forvirrende for brukeren. I tillegg får en ikke ivaretatt det at et bilde/vedlegg kan brukes flere ganger i forbindelse med ulike saker, dvs at det er behov for å gjenta element som er knyttet direkte til saken bilde var brukt i forbindelse med og ikke beskriver selve bildet, noe som vil si nesten alle feltene. Den samme problematikken for dette finnes i SIFT, men der har en gjentatt elementene **/bilde-nr/brukt-dato/ produkt/side/motiv/** for hver gang bildet er brukt i andre saker, slik at **/sak/** kun blir lagret for bildet første gang det blir brukt. En må da vurdere hvilke element som er viktig å ta vare på dersom et bilde er brukt flere ganger i ulike sammenhenger. Slik jeg ser det er **/brukt-dato/ produkt/side/** viktig, for da har du de referanser du trenger for å finne saken til bildet ved å søke i SIFT stoffarkiv. Jeg ser ingen annen god måte å løse dette på for øyeblikket. Som oftest når et bilde blir brukt flere ganger har det tilknytning til lignende saker den allerede er blitt brukt i sammenheng med. Dette er ikke tilfelle med f.eks portrettbilder som i for seg kan ha vært brukt til alle typer saker. Digitale dok-ument som blir scannet inn i Fotostation og som er fra byråbilder fra NTB, AP blir ikke arkivert i TIL-ARKIV mappen som finnes på skrivebordet i Fotostation. Enheter som ligger i TIL-ARKIV mappen vil være de som blir regnet som fullstendig indekserte dokument.

Tabell 6.4. Metadataskjema for digitale bilder/dokument i Fotostation

HOVED GRUPPE	UNDER GRUPPE	METADATA	BRUK AV METADATAFELTET
Bildetekst	Ingen definerte	Gruppe (Objekt-navn)	Den frie emnegruppe.
		Sak (overskrift)	I forbindelse med saken bildet er brukt, dvs artikkelens tema/ tittel.
		Fotograf	Hvem som er fotograf bak bildet, eller hvem som har tegnet tegningen m.m. Blir bare fylt ut dersmo det dreier seg om fotografier tatt enten digitalt eller med vanlig kamera.
		Motiv	Beskrivelse av billedmotivet.
		Instruksjoner	Dette feltet brukes til å skrive inn /brukt-dato/, /produkt/ og / side/ for hver gang bildet er brukt på nytt i forbindelse med en sak.
Kategorier & Nøkkel-ord	Ingen definerte	Kreditt	Brukes ikke av Adresseavisen.
		Hva er scannet	Her blir det lagt inn om det er fotografi, tegnng m.m som er scannet.
		Journalist	Opphavsperson til artikkelen.
		Opprinnelig over-føringsrefereanse	Ikke i bruk av Adresseavisen.
		Returnert til: (Opphavsrett)	Fylles ut dersom illustrasjonen/bildet er returnert til opphavs-mann. Navn legges altså inn i dette feltet.
		Nøkkelord (Gruppe)	Emneord som velges fra en kontrollert liste av utvalgte emne-ord.
	Kategorier	Produkt (kategori)	En forkortelse emnegruppe innenfor produkt kategori blir her lagt inn og kan bestå av maksimalt 3 bokstaver, f.eks MOT= MOTOR.
		Produkt (tillegg)	Identisk med /produkt/ elementet i SIFT, f.eks Utmagasinet er et produkt. Her har en en liste med verdier som er lovlige å bruke.

HOVED GRUPPE	UNDER GRUPPE	METADATA	BRUK AV METADATAFELTET
Dato og Status	**Dato & Tid**	Fotodato	Den dagen bildet er tatt, tegningen tegnet osv.
		Tid opprettet	Benyttes ikke av Adresseavisen.
		Utgivelsesdato	Den dagen bildet/illustrasjonen står på trykk i avisen.
		Utgivelsestid	Benyttes ikke av Adresseavisen.
	Status	Edit Status	Ingen av disse elementene blir benyttet av Adresseavisen.
		Priority	
		Object cycle	
	Plassering	Brukt dato (by)	Identisk med /utgivelsesdato/ ovenfor. Dersom bildet er brukt flere ganger i forbindelse med ulike saker blir ikke det ivaretatt her for annet enn første gangen bildet var brukt.
		Side (Provins/stat)	Hvilken side i avisen bildet/illustrasjonen er plassert på.
		Landskode	Benyttes ikke av Adresseavisen.
		Land	Benyttes ikke av Adresseavisen.
		Original referanse	Benyttes ikke av Adresseavisen.
	Diverse	Merknad	Andre opplysninger som er relevant, f.eks "filmen mangler" osv. Brukes på samme måten som **/merknad/** i SIFT formatene.
		Film-nr + Motiv-nr	Disse elementene tilsvarer **/film-id/** og **/Bilde-nr/** i SIFT filmformat.
Egendefinerte felter	**Ingen definert**	Egendefinert felt 1	Benyttes ikke av Adresseavisen.
		Egendefinert felt 2 .osv	Benyttes ikke av Adresseavisen.

FIGUR 6.9. Digitalt bilde indeksert i formatet i Fotostation

6.8. Presentasjon av Adresseavisens format for publisering på Internett

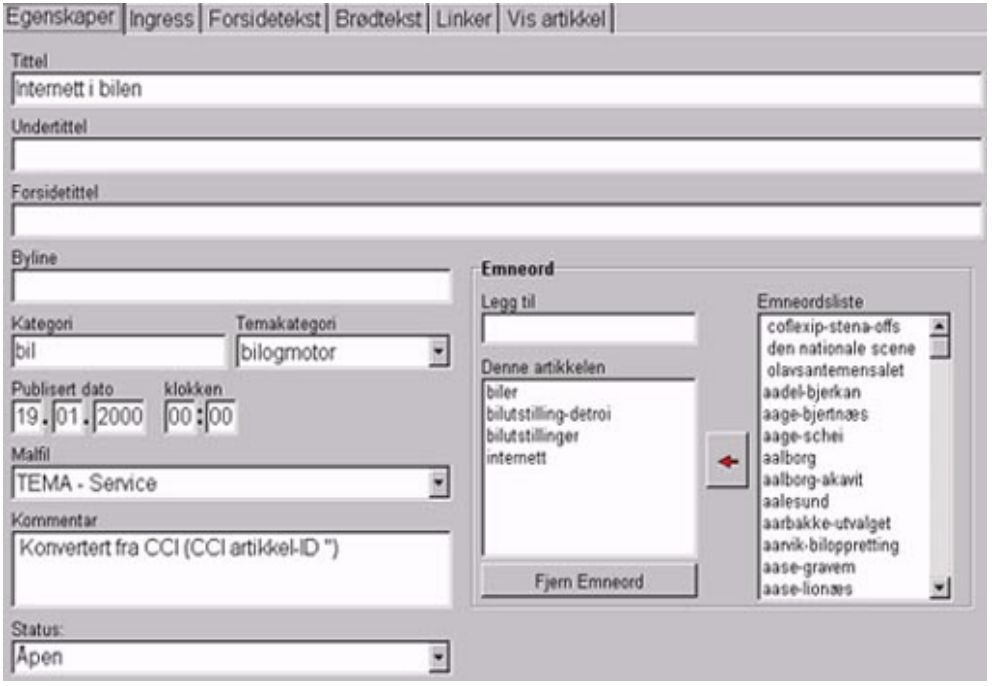
Dette systemet håndterer en artikkel med et påfølgende vedlegg, hvor vedlegget hovedsaklig er bilder. Hver artikkel kan ha flere vedlegg. Her brukes et system hvor artikkel har sine metadata, og vedlegg har sine metadata. Metadatafeltet /vedlegg/ har funksjon som en relasjon til bildet som hører til artikkelen. /**vedlegg/** referer til et mindre metadatformat for selve vedlegget til artikkelen. Emneord referer til en liste av emneord, og har dermed repeterende funksjon. Alle felt under er knyttet til Internettutgaven av Adresseavisen og ikke papirutgaven, og det er Internettjournalistene som tar seg av indekseringen her.

Tabell 6.5. Metadata for artikkel for Internettformatet

Metadatanavn	**Forklaring**
Artikkelref	Artikkelref er primærnøkkel for en artikkel i databasen, og brukes også i URL'en.
Tittel	Hovedtittel som artikkelen får ved publisering.
Undertittel	Fylles ut dersom artikkelen har en undertittel som henger sammen med hovedtittel.
Kategori	Hvilken hovedkategori artikkelen går innunder.*f.eks bil.* Det er ikke en kontrollert liste det velges etter. Nye kategorier lages ved behov.
Temakategori	Utdyping av kategori, på en måte en undertittel f.eks bilogmotor.
Publisert dato	Datoen artikkelen står på trykk i Internettutgaven av avisen. Som oftest er det samme dag som den sto på trykk i papirutgaven av avisen.
Malfil	En mal for hvordan selve artikkelen vil se ut layout-messig i skjermbildet på web-siden til adresseavisen.no.
Kommentar	Frie kommentarer til artikkelen som har intern betydning. Kan brukes som CLIBOARD.
Sperredato	Her legges det inn den dato for når artikkelen vil blir sperret frem til, f.eks hvis vi legger ut stoff før det kommer på trykk i avisa. F.eks. dersom UT-magasinet legges inn på torsdag, da stoffet finnes i CCI-basen, men ikke skal publiseres før fredag, vil da dato for fredag legges inn.
Status	Om artikkelen er publisert eller ikke. Vi legger også fakta-biter skjult, da kan de ikke finnes som egen artikkel, men finnes som lenke. Status kan settes til "Åpen inntil sperredato" eller "Åpen etter sperredato".
Byline	Hvem som har skrevet artikkelen, dvs byline er identisk med journalist.
Ingress	Dette er en kort innledning til artikkelen.
Forsidetekst	Dette er teksten som har stått på forsiden av Adresseavisen, den dagen artikkelen ble publisert i Internettutgaven. Dette feltet kan også se på som en slag forsideinnledning, men er noe lengre enn ingressen.
Brødtekst	Dette er selv tekst-kroppen, dvs hele artikkelen i råtekst.
Forsidetittel	Dette er tittelen artikkelen har fått på forsiden av www.adressa.no, og hvor du må følge linken til hele artikkelen. Ofte er forsidetittel identisk med ingress.
Emneord	Liste med frie emneord.
Vedlegg	referanse til en eller flere vedlegg som har egne metadata.

Tabell 6.6. Metadata for artikkelvedlegget for Internettformatet

Metadatanavn	**Forklaring**
Filnavn	Hvilken filnavn vedlegget har, dvs det det ligger som i databasen, f.eks "bilmotor.jpg".
Vedleggsnummer	Den unike identifikatoren til bildet.
Vedleggstype	Hvilken type vedlegg er det, et bilde, tegning m.m.
Bredde	Hvort stort er vedlegget i bredden.
Høyde	Hvort stort er vedlegget i høyden.
Byline	Hvem er opphavspersonen til vedlegget.
Tekst	Eventuelt billedtekst til bildet, som skal være med i Internettutgaven.



FIGUR 6.10. Metadataformatet for publisering for Internett

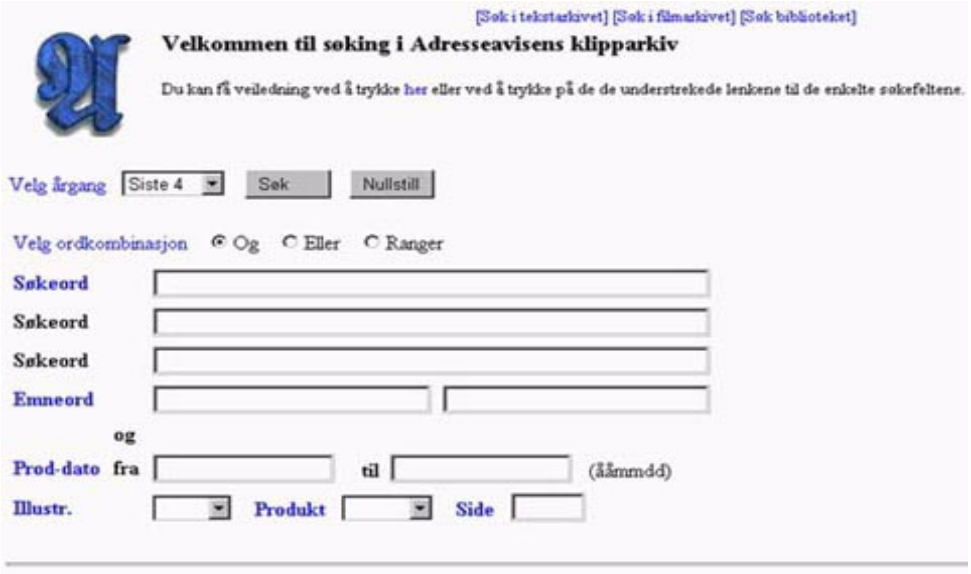
I tillegg til de nevnte metadata over for Internettartikkelen kan du i tillegg legge til linker i form av URL'er til Internettartikkelen. Dette er kun eksplisitte linker til relaterte saker og til sider

utenfor Adressa.no. Relaterte saker genereres automatisk ut i fra emneord. Metadata for artikkelvedlegget for Internettformatet

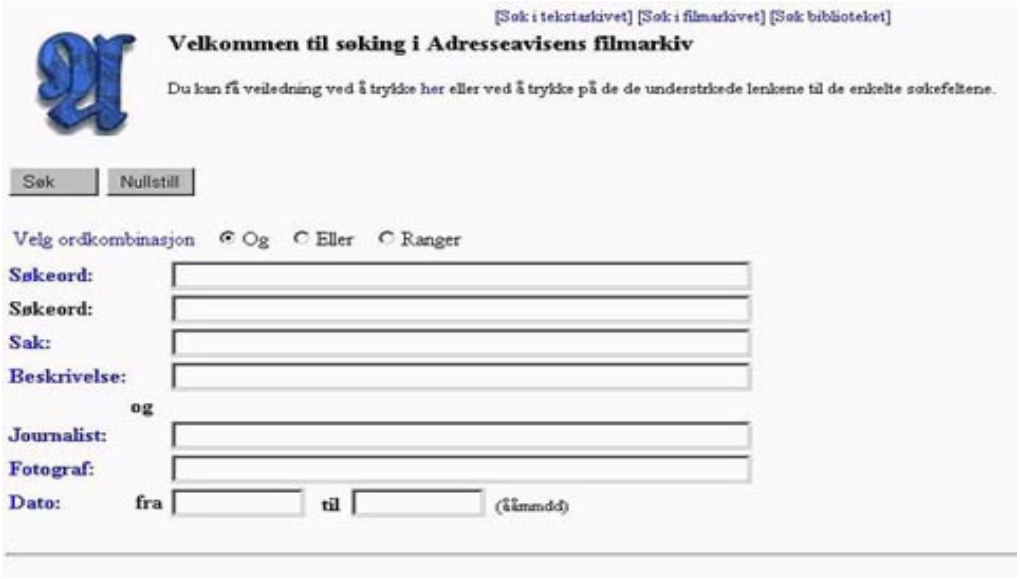
6.9. SIFT Grensesnitt

Som nevnt i de innledende avsnitt har Adresseavisen tre web-grensesnitt til SIFT basen, for filmarkivet, stoffarkivet og bibliotekdatabasen. Disse grensesnittene er surrogater av de formatene som finnes innenfor de ulike arkivene i SIFT. Web-grensesnittet for stoffarkivet og filmarkivet er presentert i figur 6.11 og 6.12. Vi ser at det ikke er mange metadataelement som er tatt med for søk her. For stoffarkivet er det mulighet til å søke på emneord, datointervall, hvilken illustrasjon det er til saken, samt produkt og side. For illustrasjon kan du søke på alle illustrasjoner, eller bare foto, tegning eller grafikk. For filmarkivet kan du søke på metadata som sak, beskrivelse, datointervall, samt fotograf og journalist. I tillegg har filmarkiv og stoffarkiv mulighet til å kombinere henholdsvis 2 og 3 termer i fritekstsøk med boolsk operatorer.

Tidligere brukte både journalistene og arkivet kommandobasert grensesnitt mot SIFT, men etter at web-grensesnittene ble innført bruker journalistene kun dette grensesnittet. Arkivet bruker fremdeles kommandogrensesnittet, fordi de kjenner det godt og gir flere muligheter i informasjons-søkingen enn web-grensesnittene. Arkivet har også mulighet til å benytte alle tjenester i SIFT, noe ikke journalistene kan, og dermed blir web-grensesnittet ikke tilfredsstillende nok for arkivet sitt behov.



FIGUR 6.11. Webgrensesnittet for SIFT stoffarkiv



FIGUR 6.12. Web-grensesnitt for filmarkivet

Kapittel 8

Mapping av de ulike formater

8.1. Innledning

I dette kapittelet mappes Ad-hoc formatene som Adresseavisen benytter mot DC formatet. Målet er å finne ut om DC kan brukes som et felles format for indeksering av Adresseavisens informasjonsobjekter. I avsnitt 8.2, (s149). er det et kort sammendrag til resultatet av mappingen gjort mellom DC og BIBSYS-MARC. Deretter følger mapping av DC og Ad-hoc formatene på rad og rekke i resten av kapittelet.

Hvert element blir drøftet ut fra brukskontekst mot DC sin brukskontekst, og resultatet av mappingen blir presentert i tabeller. Dersom DC kan tilfredsstilles for de ulike informasjonsobjekt, vil jeg ta utgangspunkt i et informasjonsobjekt indeksert med ad-hoc format og vise hvordan dette kan indekseres i DC.

Til slutt oppsummeres de DC element, pluss tilleggselement, som kan fungere som metadatasett for indeksering av informasjonsobjekter ved Adresseavisen. Det jeg da ender opp med er mitt forslag til felles kjerneformat for indeksering, med medieavhengige variasjoner. Siden jeg ikke har oversettelser av sub-element til norsk å referere til, bruker jeg de engelske betegnelsene for dette i mappingen. Norsk forklarelse til disse finnes i kapittel 5, (s57) om metadata-formater.

8.2. Mapping mellom Dublin Core og BIBSYS-MARC

Mapping mellom Dublin Core og BIBSYS-MARC er et arbeid som er gjort i prosjektet "BIBSYS Digital bibliotek". Mapping fra Dublin Core til BIBSYS-MARC [4] viser at Dublin Core fullt ut kan tilfredsstilles i

BIBSYS-MARC, i og med at BIBSYS-MARC har flere strenge regler til innhold enn hva som er tilfelle med Dublin Core og er et rikere format. BIBSYS-MARC sitt behov vil derfor ikke kunne tilfredsstilles i DC [5].

8.3. Mapping av Ad-hoc formatene i SIFT til Dublin Core

DC er et indekseringsformat for publisering av informasjon via distribuerte nettverk som Internett. Via Intranett og Internett kan Adresseavisen utvide sin tilgang til sine informasjonsressurser, og i så måte blir det her vurdert om DC kan fungere effektivt nettopp med tanke på denne type publisering og i samsvar med Adresseavisens behov på dette området.

Ad-hoc formatene som her blir mappet er som følger:

1. Formatet for illustrasjoner i SIFT
2. Formatet for fysisk film i SIFT
3. Formatet for stoff/artikler i SIFT
4. Formatet for digitale bilder i Fotostation
5. Formatet for Internettpublisering (med vedleggformatet for bilder)

For hvert format som mappes vil det bli gitt eksempel på mapping av et konkret eksempel ut fra de resultatet som fremkommer av hver mapping.

8.3.1. Mapping av Ad-hoc formatet for illustrasjoner

Illustrasjonsformatet slik det fungerer i dag passer best til internt bruk. Informasjonsobjektene har ingen interesse eksternt så lenge de ikke er utstyrt med motivbeskrivelse, og brukeren kan skille mellom de ulike illustrasjonene. Oppsummering av mapping er vist i tabell 8.1, (s152)

8.3.1.1. Drøfting av mappingen

Ill-id Formatet har en unik identifikator for hver registrert enhet i SIFT og kan følgelig mappe DC. Identifier. **/Ill-id/** her følger saken illustrasjonen var brukt og ikke selve illustrasjonen. **/ill-id/** har ingen funk-

sjon i dag, da du trenger å vite tallet som representerer **/ill-id/** for å kunne søke på elementet.

Det anbefales her at hver illustrasjon som er lagret i Adressesavisen arkiv utstyres med en unik ID på samme vis som det er gjort for filmmappene. Dette vil lette adminstrasjon og kontroll med hvilke illustrasjoner som faktisk er brukt. **/ill-id/** er nødvendig å ha med dersom beskrivelseselement for motivet til illustrasjonen innføres, for å kunne spore seg tilbake til illustrasjonen.

Illustr

Viser til type illustrasjon som beskrives, dvs er det et bilde, tegning etc. Dette mappes av DC.Type.

Ill-type

Dette viser til om illustrasjonen i papirutgaven er svart hvitt eller farget. DC har ikke et element som matcher dette direkte, men en mulighet her er å benytte DC.Type og innføre sub-elementet Farge, dvs DC.Type.Farge hvor verdiene "svart hvitt bilde", "fargebilde" og eventuelt andre verdier som kan være aktuelle innføres som lovlige verdier her. Anne Marie Vercoustre [37] har brukt DC.Type for dette formål til sine fotografier, men uten sub-elementet farge. Det viser at å bruke dette elementet kan fungere også uten sub-elementet. Jeg mener det er passende å innføre sub-elementet Farge her for DC.Type, siden DC.Type uten sub-element brukes for å definere om illustrasjonen er et bilde, tegning o.l.

Tabell 8.1. Mapping av DC mot illustrasjonsformatet

DUBLIN CORE META-DATA	**SIFT METADATA FOR ILLUSTRA-SJONER**	**KOMMENTARER**
DC. Identifier	Ill-id	Skjema her er knyttet il Illustrasjon og kan defineres SCHEME=Illustrasjon.
DC.Type	Illustr.	Hvilken genre av bilder hører ressursen innunder, er det en tegning, en plakat, et cd-cover, spill-cover, bokomslag osv.
DC.Type.Farge	Ill-type	Verdier her blir "Svart/hvitt" og "Farge".
DC.Subject.Classification	Gruppe	
DC.Creator.Agentrole DC.Creator.Agentname DC.Creator.AgenAffiliation	Opphav	
DC.Title DC.Relation.Sak	Sak	Tittel skal her relaters til illustrasjonen, men vil også si noe om saken. Tittel har ikke formatet fra før. DC.Relation.Sak brukes til å beskrive saken.
DC.Date.Issued	Bruk-dato, Prod-dato	Publiseringsdato.
DC.Relation.IsPartOf	Produkt	Verdier her er Ut-magasinet, Uke Adressa, Lille Sportsavis osv.
Element som ikke mappes		
- - -	retur	Beholdes.
- - -	retur-adr	Utgår. Ikke behov for å beholdes.
- - -	merknad	Beholdes.
- - -	Side	Beholdes.
- - -	Antall	Et felt som ikke er relevant. Antall kan nevnes i merknadsfeltet i de tilfeller en mener det er nødvendig.
Anbefalte element å bruke i tillegg [Se "Tilleggsbehov for bilder" avsnitt 8.4, (s184)]		
DC.Contributor.Agentrole DC.Contributor.Agentname DC.Contributor.AgentAffiliation		
DC.Subject		Frie emneord. Kommer i tillegg til kontrollerte emneord i DC.Subject. Classification.
DC.Description		"Beskrivelse av plaktamotivet".
DC.Rights		Kan fylles ut med f.eks "Fri benyttelse for Adresseavisen" hvis opphavsperson er ukjent.
DC.Type.Quality		Kvaliteten på bildet. Her må en bli enig om verdier det er behov for. *F.eks "God", "Middels" osv.*
DC.Format.Extent		Bildet sitt størrelse i cm for bredde og høyde. *f.eks 50*40cm.*
DC.Description.Genre		Dette innføres for å definere type genre på motivet. f.eks portrett, landskap m.m.
DC.Coverage.Place		Geografisk sted bildet dekker. *F.eks "Fosen".*

DUBLIN CORE META-DATA	SIFT METADATA FOR ILLUSTRA-SJONER	KOMMENTARER
DC.Format.		Hvilken type ressurs det er som beskrives, f.eks bilde, tekst, lyd, samling m.m. Her vil det da verdien være bilde for alle. f.eks "bilde/papir"
Repeterende element	**(Repeteteres for hver gang illustrasjonene er brukt)**	
DC.Date.Issued DC.Relation.IsPartOf Side DC.Relation.Sak		

I tillegg har Anne M.V. (AMV)[1] brukt et sub-element til DC.Type som definerer kvaliteten på fotografiet. Dette kan være nyttig å bruke for illustrasjoner ved Adresseavisen, da mange av dem foreligger i papirform med varierende kvalitet. Siden kvaliteten er varierende og en må kopiere fra papirformatet dersom illustrasjonen skal brukes på nytt, kan det være interessant informasjon for hvor godt resultatet vil bli.

Bildets kvalitet er viktig informasjon, da dette gir mulighet for å velge det bildet med best kvalitet til gjenbruk. Dette vil også gjelde for illustrasjoner i Fotostation, da scannede bilder også har varierende kvalitet alt etter om det er et negativ eller papir m.m som er kilden for scanning. For illustrasjoner kan det her være aktuelt å legge in DC.Source som får verdier som "negativ film", "papirkopi", "dias" m.m.

Dersom illustrasjonen skal brukes til scanning eller kopiering, vil den størrelse si noe om hvilken kvalitet en vil få ut av det. Størrelsen til bildet kan derfor defineres ved DC.Format. sitt Extent (Appendix A) element som DCMI har foreslått.

Høydebilder, dvs at f.eks bygninger eller annet er tatt i vertikal retning er noe som ofte etterspørres ved bestilling av bilder. Dette kan tilfredsstilles ved inn-førelse av sub-elementet Orientation som også A.M.V. også har brukt.

1. Heretter referert til som A.M.V [37]

I tillegg etterspør brukerne et genre element som beskriver hvilken type fotografi det er. f.eks portrett, landskap osv. Portrettbilder er bilder som ofte etterspørres for å bruke i forbindelse med intervju av personer som kjendiser og politikere. Dette er genre til motivet, og jeg mener at det kan legges som sub-element til beskrivelseselementet, dvs DC.Description.Genre.

Gruppe

Dette er emnegruppen som illustrasjonen er klassifisert etter. Dette velges fra en egen liste over kontrollerte emneord og emnegruppe mapper DC. Subject.Classification.

Opphav

Opphav er den som har frembringet ressursen, og mappes av DC.Creator. Her anbefales også bruk av sub-elementene for å gjøre dette mer detaljere innholdet. Opphavspersonen tilhører ofte en organsiasjon. Sub-elementet for rolle tas også med. Vi får elementene DC.Creator.Agentrole, DC.Creator.Agentname og DC.Creator.AgentAffiliation som skal dekke **/opphav/**.

Sak

Dette er tema for saken som illustrasjonen har stått på trykk sammen med. Det er ikke riktig å bruke DC.Description her, da sak er en relasjon til illustrasjonen som er brukt. DC.Description skal brukes for selve beskrivelse av ressursen, som da hadde vært illustrasjonsmotivets innhold.

Når det gjelder beskrivelse av saken kan dette dekkes både av et tittelfelt og et relasjonsfelt. Bruk av tittel felt her skal knyttes til bildets motiv, og ikke til saken, Tittel vil allikevel si noe om sakens innhold. I Margareth Graham sin undersøkelse [17] var det rundt 70 og 80% av de som 60 medlemmer deltok som brukte tittel eller navn på tegninger og malerier og 44% som brukte det på plakater. Dette viser at det er veldig vanlig å bruke tittel i indekseringen, og indikerer at den har en viktig betydning. Sak mappes best ved innføring av sub-elementet Sak til relation, d.v.s DC.Relation.Sak.

DC.Relation.Sak kan enten brukes hver gang illustrasjonen blir brukt, eller lagres bare første gang illustrasjonen blir brukt som det gjøres i dag. *Er det et motiv fra f.eks en bestemt film er det lite trolig at motivet blir brukt i andre sammenhenger enn omtale av filmen eller i omtale om en bestemt filmgenre.For denne type illustrasjon er det ikke det store behovet for gjentakelse av elementet.*

Brukt-dato/Prod-dato

/Brukt-dato/ er et element som gjentas for hver gang en illustrasjon har vært brukt, men elementet blir ikke helt brukt slik det er tenkt å brukes. Med en unik ID til illustrasjonen ville dette ha fungert veldig likt filmformatet, med gjentakelse av elementene /**dato**/**produkt**/**side/** hver gang illustrasjonen gjenbrukes. Nå er det ikke illustrasjonen som følger **/ill-id/**, men saken. Hver gang en illustrasjon har stått på trykk i forbindelse med en sak registreres dette.

Det er ingen mulighet til å vite om dette samme motivet har vært brukt tidligere siden motivet ikke beskrives. Når en og samme illustrasjon er brukt i forbindelse med to ulike saker blir de lagret som to uavhengige poster i SIFT sitt illustrasjonsformat. Dette fører til at metadataelementene **/brukt-dato/produkt/side/** svært sjelden blir gjentatt, selv om **/brukt-dato/** er ment å gjentas for hver gang illustrasjonen er brukt på nytt. Første gang illustrasjonen er brukt er **/prod-dato/** og **/brukt-dato/** identisk, så det er derfor behov for et element for publiseringsdato her. Dette mappes av DC.Date.Issued.

Dersom unik ID innføres for hver illustrasjon som anbefales, repteres DC.Date.Issued for hver gang illustrasjonen publiseres i ny sak.

Produkt

Dette er en relasjon til produktet, eller "hoved dokumentet" hvor illustrasjonen og saken sto på trykk. Produkt er her å regne som en type kategori eller tema som omtales som produkt, og hvor innholdet her er enten Ut-magasinet, kultur delen av Hovedavisen, eller Lille Sportsavis, d.v.s de ulike oppdelinger som tilsammen er en del av hele Adresseavisen. DC har et relasjonselement som dekker dette. Illustrasjonen er å regne som en del av artikkelen og artikkelen+bildet som en del av Ut-magasinet. Relasjonen her kan tilfredsstilles i DC.Relation.IsPartOf.

Resultatet av denne mappingen blir at elementene /**Antall**/**Retur**/**retur-adr**/**merknad**/**side**/ ikke kan mappes som vist i tabell 8.1, (s152).

Jeg ser ikke det noe behov for å beholde **/antall/** her. Ved å innføre /ill-id/ for egne illustrasjoner, vil **/antall/** illustrasjoner brukt i en sak vises ved antall treff i trefflisten, *f.eks får du opp tre ulike illustrasjoner for en og samme sak er det brukt 3 illustrasjoner til den saken.* Det er heller ikke behov for å søke på **/antall/.**

/Retur/ elementet sier om antall illustrasjoner er returnert eller ikke, og har verdiene "JA" eller "Nei". Her kan det være aktuelt å skille søk på de illustrasjoner vi har selv, og de som er returnert. Det er et behov for å avgrense treff ved å søke på **/Retur/**. Elementet **/retur-adr/** utgår, da det ikke er behov for å beholde dette. Opplysninger som legges her kan overføres til **/merknad/.**

/merknad/ elementet skal her brukes til frie kommentarer. Her bør en allikevel avtale et kontrollert vokabular for fraser som det aktuelt å søke på i dette elementet for å lette gjenfinning. *f.eks Verdiene "illustrasjon ikke mottatt", "returnert til opphavsperson", "illustrasjon mangler" betyr alle at illustrasjonen mangler. Hvis du da søker på "film mangler" i **/merknad/** vil du ikke få opp illustrasjonene med de andre verdiene.* Vær oppmerksom på at dette fortrinnsvis skal brukes til frie kommentarer, da forståelsen av /**merknad/** nettopp er dette.

Informasjon som ligger i disse /**retur/** og **/merknad/** er administrativ metadatainformasjon, som brukes internt.

Mange illustrasjoner er ikke fysisk innom arkivet før de blir lagret, dette gjelder som regel byråbilder. I tillegg har du en god del spillcover, bokcover m.m, men disse blir alltid sendt til vedkommende som har skrevet saken som da vil stå som opphavsperson. Disse illustrasjonene er det ikke aktuell å spore tilbake for gjenbruk, og trenger derfor ingen unik ID.

En har allerede mulighet som besøkende til arkivet å søke i stoffarkivet, og filmarkivet via web-grensenittet. Når en søker i filmarkivet, er det som oftest for å finne bilder som kan være aktuelle å bruke i forbindelse med en oppgave som skal skrives, da det ofte er skoleklasser på grunnskole, ungdomsskole og v.g.skole nivå som henvender seg til arkivet. Da vil du på søk se om denne posten i filmarkivet enten er en film eller illustrasjon, og om du er interessert i illustrasjonen må du vurdere dets motiv ut fra emnegruppe den er arkivert etter, siden motivet ikke er beskrevet. Her hadde det vært greit med en beskrivelse av motivet for å lette relevansvurderingen og dermed øke effektiviteteten i gjenfinning. Dette har du ikke mulighet til, men må lete deg frem til arkivkonvolutten det refereres til via navnet på emnegruppe, for deretter å studere alle illustrasjonene som ligger i konvolutten. Dersom hver enhet hadde en unik ID kunne du også fått sett hvor stor samlingen var, *f.eks hvor mange bilder har du i Adresseavisens arkiv av Ivar Aasen.* Slik formatet for illustrasjoner blir et og samme portrettbilde som er brukt 9 ganger, lagret som 9 registererte en-heter i SIFT. Betydningen av **/antall/** vil her være forvirrende for ekstern bruker.

Det er ikke alle illustrasjoner som det er relevant å beskrive motivet. Dette gjelder oftest byråbilder og bilder som returneres, og som ikke er aktuell å spore tilbake. I Fotostation blir disse scannet inn, og er dermed direkte tilgjengelig for gjenbruk. I SIFT blir bare metadataposten til illustrasjonen lagret.

Dersom det skal være aktuelt å tilby skoleklasser tilgang til SIFT databasen eksternt, er det nødvendig med en beskrivelse av motivet for de illustrasjonene som Adresseavisen har. De kan da på forhånd velge ut de motiv de er interessert i før de henvender seg til arkivet med sin bestilling. Brukeren gjør da søkejobben selv.

Det anbefales at **/ill-id/** blir lagret for alle illustrasjoner som Adresseavisen har lagret. Dette vil lette gjenfinning, og gi mulighet til å spore det bildemotiv som faktisk er brukt.

Elementet **/side/** er det behov for å beholde siden illlustrasjoner ofte har faste sider i avisen alt etter hvilke typer saker de illustrerer, og kan brukes til å avgrense et søk. *f.eks Debattsidene står alltid på side 8 og 9 i Adresseavisen, og her brukes det ofte tegninger. Dersom du da er interessert i en illustrasjon som er brukt her, kan du avgrense ved å søke på **/side/** i tillegg til andre opplysninger.*

Det er ikke her fylt ut informasjon for eventuell bidragsyter her, men i de tilfeller hvor det kan være aktuelt å fylle ut dette, bør metadata for bidragsyter også tas med. Bidragsyter vil da få samme forhold som /journalist/ har for filmformaet [se drøfting av dette 8.3.2, (s163)].

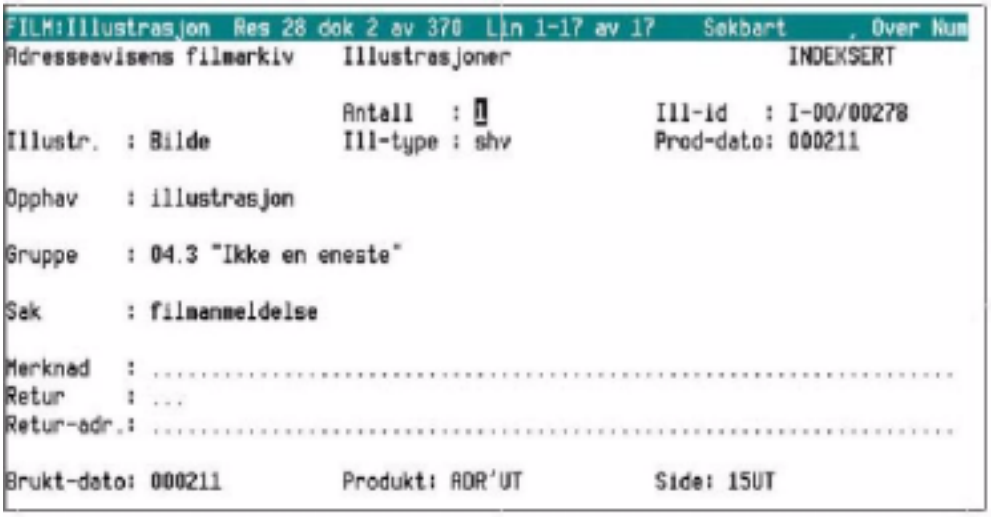
Element som DC ikke mapper, men som det er behov for å beholde:

- Side
- Merknad
- Retur

Eksempel 8.1 Mapping av en illustrasjon i SIFT til DC

Her mappes illustrasjonen som vist i figur 8.1, (s158). Dette er den samme som er vist i Casestudiet i kapittel 6, men for å unngå å bla tilbake er figuren gjengitt her. Her ville det også vært aktuelt å benytte DC.Descripton, DC.Type.Quality, DC.Source osv men det er informasjon som ikke er lagret i formatet for illustrasjoner i dag,

f.eks i eksemplet indeksert i tabell 8.2, (s158) kan DC.Description inneholde beskrivelsen "Paul og Tina som er hovedpersonene i filmen sitter ved siden av hverandre og snakker ansikt til ansikt på en benk i parken", DC.Type.Quality inneholde verdien "Middels" og DC.Format.Extent verdien "50*40cm".



FIGUR 8.1. Indeksert illustrasjon i SIFT

Tabell 8.2. Eksempel på mapping av illustrasjonen i figur 8.1, (s158).

DC METADATA	INNHOLD	KOMMENTARER
DC.Identifier	"I-00/00278"	Skjema her vil være "illustrasjon".
DC.Date.Issued	"000211"	
DC.Type.Farge	Svart/hvitt	
DC.Type	PlakatBilde	
DC.Subject.Classification	"4.3 Film"	Motivet er en scene fra en film.
DC.Title.	"Ikke en eneste"	Informasjon: Motivet er en scene fra filmen "Ikke en eneste".
DC.Relation.IsPartOf	"Ut-magasinet"	
Side	"15"	
DC.Creator. Agentname DC.Creator.Agentrole DC.Creator.AgentAffiliation	"Intet innhold" siden det er ukjent her hvem som er opphavsperson.	Her er "illustrasjon" fylt ut , noe som er helt usaklig og ikke gir mening for eksterne brukere. Når opphavsperson er ukjent bør dette elementet være tomt.

Tabell 8.2. Eksempel på mapping av illustrasjonen i figur 8.1, (s158).

DC METADATA	INNHOLD	KOMMENTARER
Merknad	Ingen merknad i dette feltet	
DC.Relation.Sak	"Filmanmeldelse"	Blir lagret første gang. Kan også gjentas for hver gang illustrasjon er brukt, dersom sak er vesentlig forskjellig.

Kommentar til eksempel: Sak kan gjentas for hver gang bildet er brukt i forbindelse med en sak. En får ikke da påvist et direkte forhold mellom hvilken sak som går til hvilken dato, men det får en ikke slik formatet er pr. i dag heller. Definering av relasjoner mellom de ulike element kan realiseres ved f.eks å bruke XML til å implementere metadata. Slik formatet er i dag påvises relasjon mellom elementene som gjentas for hver gang illustrasjonen er brukt ved å gjenta en frekvens av elementer. De som da kommer fortløpende hører sammen.

F.eks: Du har to frekvenser av elementer hvor hver frekvens består av 4 element som vist nedenfor. Elementene i en frekvens hører sammen.

Frekvens 1:
DC.Date.Issued DC.Relation.IsPartOf DC.Relation.Sak Side

Frekvens 2:
DC.Date.IssuedDC. Relation.IsPartOf DC.Relation.Sak Side

mens de i realiteten vil se slik ut:

DC.Date.Issued:.....DC.Relation.IsPartOf:..... Side:.....
DC.Relation.Sak:.....
DC.Date.Issued:.....DC.Relation.IsPartOf:..... Side:.....

DC.Relation.Sak:**.....**

8.3.2. Mapping av Ad-hoc formatet for fysisk film i SIFT

Adresseavisen legger ut et utvalg av artikler i sin Internettutgave, men de bruker ikke det samme format som er brukt for indeksering av artikler i SIFT.

8.3.2.1. Drøfting av mappingen

Jobb-id/Film-id

Dette er filmformatet sine to identifikator. DC har her et element for identifikator som er DC.identifier. **/Jobb-id/** fungerer her som en unik identifikasjon til posten i databasen. **/Film-id**/ er også en identifikator som er unik og som refererer til en fysisk identifikator som finnes på en filmmappe, *f.eks som ISBN nummer for bøker fungerer.* Dette formatet har heller ingen kommunikasjon fra andre databaser utenfor SIFT (dvs ingen import fra CCI[1] som SIFT stoffarkiv har). Dersom DC.identifier skal brukes her må en kunne definere hva som er ID i basen og hva som er ID til filmmappen. Nå har ikke DC identifisert noe sub-element, men i dette tilfellet kunne sub-element av typene DC.identifier.jobb og DC.identifier.film være aktuell dersom begge identifikatorene er nødvendig.

En annen løsning hadde vært å tilfredsstille dette ved SCHEME kvalifikatoren, *f.eks:*

name=DC.Identifier SCHEME="Film-id skjema" CONTENT="F-000102/45520"

og

name=DC.Identifier SCHEME="Jobb-id skjema" CONTENT="J-970828/903".

Tabell 8.3. Mapping av DC mot filmformatet

DC METADATA	SIFT META-DATA FOR FILM	KOMMENTARER
DC. Identifier.Film	Film-id	Her blir skjema="Film" som viser til at det filmformatet som benyttes her.
DC.Identifier.Jobb	Jobb-id	Skjema her blir også film.
DC.Source.Type	Film-type	Verdier her blir farge positiv (dias), farge negativ osv.
DC.Date.Created	Film-dato	
DC.Date.Issued	Prod-dato / Brukt-dato	DC.Date.Issued hver gang filmmappen brukes.
DC.Creator.Agentname	Fotograf	Her er det navnet på fotografen som oppgis slik det benyttes i dag. Agenten sin rolle og eventuell bedrift/organisasjon han eller hun tilhører
DC.Creator.Agentrole		
DC.Creator.AgentAffiliation		

1. CCI=Avisproduksjnssystemet

Tabell 8.3. Mapping av DC mot filmformatet

DC METADATA	SIFT META-DATA FOR FILM	KOMMENTARER
DC.Contributor.Agentname	Journalist	Her er det bare navnet på journalisten som oppgis slik det benyttes i dag. Agenten sin rolle og eventuell arbeidstedtil-knytning er også nødvendig å definere.
DC.Contributor.Agentrole		
DC.Contributor.AgentAffiliation		
DC.Subject.Classification	Gruppe	Kontrollert vokabular hvor skjema blir egendefinert liste.
DC.Relation.Sak	Sak	Kan gjentas ved gjenbruk av filmmotiv
DC.Title	Sak	Tittel rettet mot bildemotiv, men gir også info om sak.
DC.Relation.IsPartOf	Produkt	Som illustrasjonsformatet.
DC.Description	Motiv	Beskrivelse av hva vi ser på bildet.
DC.Relation.Filmmappe	Film-nr	Se kommentar for repterende element
Element som ikke mappes		
- - -	Merknad	Beholdes
- - -	Retur	Beholdes
- - -	Retur-adr	Ikke behov for å beholde. Utgår.
DC.Source.Bildenummer	Bilde-nr	**/Bilde-nr/** er referanse til motivet på negativ filmen. Negativ film mappes av DC.Source. Type. Legger **/bilde-nr/** som sub-element her, for å vise at dette er relatert til kilden. Velger å skrive det uten forkortelser.
- - -	Side	Beholdes slik det er.
Anbefalte element å bruke i tillegg [Se "Tilleggsbehov for bilder" avsnitt 8.4, (s184)]		
DC.Subject		Frie emneord
DC.Rights	- - -	Opphavsrett
DC.Type.Quality	- - -	Med DC.Type.Quality kan en si noe om kvaliteten på ressursen. Ressursen her er filmmappen, men det er kvaliteten på bildemotivet som er av interesse her, dermed må vi rette dette mot at det bildet som er ressursen vi vil si noe om kvaliteten på. Hvert bilde som lagres pr. filmmappe blir da en ressurs, og vi får her repitisjoner av dette elementet.
DC.Description.Genre	- - -	Elementet for å definere om bildet er portrett, landskap m.m.
DC.Format	- - -	Definerer hvilken type ressurs som beskrives, som her er en filmmappe med filmstriper.
DC.Coverage.Place	- - -	Elementet for å definere størrelse eller dimensjoner i x og y på bildet er ikke nødvendig her siden kilden er negativ film, og dermed har en fast str.
Repeterende element		
DC.Source.Bildenummer		Gjentas for hver gang bildet er brukt på nytt.
DC.Date.Issued		
DC.Relation.IsPartOf		

Tabell 8.3. Mapping av DC mot filmformatet

DC METADATA	SIFT META-DATA FOR FILM	KOMMENTARER
DC. Relation.Filmmappe		Gjentas ikke for hver gang bildet er brukt, men for alle de mappene som inngår i denne serien. Er ikke et element for bruk til relaterte saker, men til filmmapper som dekker samme sak.
DC.Relation.Sak		Gjentas for hver gang bildet er brukt på nytt.
Side		
DC.Coverage.Place		Ikke alle bilder på filmen som brukes trenger være av et som kan relateres til geografisk sted. Dette fordi det også kan finnes portrettbilder på filmen.
DC.Type.Quality		Gjentas en gang for hver DC.Source.Bildenummer

for å skille disse. Det er behov for å søke på **/Film-id**/ og ikke **/jobb-id/,** og jeg foreslår at sub-elementene jobb og film benyttes.

Film-type Her er det hvilken type film som er brukt som er informasjonen som fylles ut i dette feltet. Type negativ film som er brukt sier noe om mulighetene for eksponering og fremkalling av filmen, og er dermed å regne som en kilde til filmen. Dette mappes av DC.Source hvor negativfilmen da verdien. Vi ønsker i tillegg å definere om det er svart hvitt film m.m. og det kan gjøres ved innføring av sub-element Type som får verdier som "farge postitiv" (lysbilde film) og "farge negativ". A.M.V. bruker også dette sub-elementet.h

Film-dato **/Film-dato/** er den dagen filmen var ferdig, dvs den dagen bildene er knipset med et kamera. Allikevel kan det ta noen dager før filmen blir brukt på trykk i avisen i forbindelse med en sak. /**Film-dato/** mapper til DC.Date.Created.

Brukt-dato/Prod-dato Samme argumentasjon som for illustrasjonsformatet, [s155] Behovet for dato her mapper DC.Date.Issued som er dagen for formell utgivelse av ressursen.

Fotograf Fotograf er opphavspersonen til filmen, men andre ord den som har frembrakt ressursen. Dette mapper DC sitt Creator element. Her kan en også definere rollen til Creator som her er fotograf. Det dekkes av DC.Creator.Agentrole="Fotograf". Navnet på Creator legges i sub-ele-

ment DC.Creator.name. Fotograf kan også være knyttet til en arbeidsgiver, og det dekkes av DC.Creator.AgentAffiliation.

Journalist

Journalist er opphavsmannen til artikkelen, altså konteksten filmen er brukt i. Dette er ikke metadata knyttet til filmen, men til saken. Dette elementet henger sammen med /**gruppe**/ og **/sak/** som også er knyttet til saken. DC har et element for bidragsyter, dvs en person som har bidratt til frembringelse av ressursen. En journalist har produsert artikkelen, ikke filmen. Journalisten er knyttet til bildet sin brukskontekst. Dersom en gjør bruk av sub-elementet til elementet for bidragsyter i DC som definerer hvilken rolle bidragsyter har, kan kanskje feltet benyttes, i og med at "alle" vet relasjonen mellom en journalist og en fotograf og hvordan de forholder seg til hverandre. Dette er ikke helt enkelt, fordi når et bilde brukes i forbindelse med en artikkel vil fotograf være bidragsyter til artikkelen fordi bildet og artikkelen i ett anses som hele artikkelen. Dette fordi bildet visualiserer innholdet i artikkelen. På en annen måte kan vi kanskje si at journalist indirekte har bidratt til frembringelse av bildet i og med at mange bilder tas nettopp fordi en bestemt sak skal dekkes og at de henger naturlig sammen. Hadde det ikke vært fordi saken skal dekkes ville ikke bildene blitt tatt, så på den måten bidrar også journalisten til bildenes innhold fordi bildene nettopp skal illustrere saken.

Jeg velger her å benytte DC.Contributor for journalist, men da også med elementene DC.Contributor.Agentrole,DC.Contributor.Agentname, og DC.Contributor.Agent-Affiliation for å spesifisere innholdet så detaljert som mulig. Ofte er også journalist og fotograf den samme i avissammenheng, og jeg mener det derfor blir riktig å bruke dette elementet her. I denne situasjonen er det behov for å verifisere rollen til bidragsyter, i og med at det kan finnes andre bidragsytere enn en journalist, selv om det ikke er tilfelle for Adresseavisen.

Gruppe

Her brukes et kontrollert vokabular for emneord og det mappes av DC.Subject.Classification.

Sak

Her anbefales også bruk av DC.Title og DC.Relation.Sak, som foreslått for illustrasjonsformatet. *F.eks tittel kan være "Fotballkamp*

Strindheim-Nardo" som indikerer at bildene inneholder motiv av spillerne på banene. Kan også inneholde portrettbilder.

Pr. dato blir motivbeskrivelsen gjentatt på nytt for hver gang motivet brukes, noe som er helt unødvendig. Istedet for å gjenta motivelementet kan en heller bruke denne plassen til å gjenta saken for hver gang bildet er brukt. Dermed blir sak ikke bare lagret første gang filmen brukes.

Produkt Dette er "produktet" hvor artikkelen og fotografiet er brukt. Det er faste verdier for hva som regnes som et produkt. Produkt er de bilag som følger med utgivelsene av Adresseavisen, *f.eks Hovedavisen, Utmagasinet, Lokal-avisa, Lille-Sport osv.* Her kan DC.Relation.IsPartOf benyttes, som også foreslått for illustrasjonsformtaet.

Motiv Dette mapper DC.Description direkte, *f.eks DC.Description="Heidi Sørensen - portrett" som beskriver fotografiet.*

Vi har altså elementene /**Film-nr**/**Retur**/**retur-adr**/**merknad**/**bilde-nr**/**side/** som ikke kan mappes direkte av DC.

/**Retur/** feltet inneholder verdien JA eller NEI dersom film er returnert. Her vil jeg også foreslå samme løsning som for illustrasjonsformatet der /**retur/** og **/retur-adr/** tilfredsstilles i **/merknad**/ [Mer om dette i avsnitt 8.3.1, (s150)].

/Bilde-nr/ er referanse til negativnummeret som identifiserer motivet på filmstripene. Filmstriper lagret på samme filmmappe kan ha lik **/bilde-nr/**. Dette indikeres ved Bilde-nr="2a (2)", hvor tallet "2" viser til filmstripe 2 i filmmappen. Dette er informasjonen du trenger for å lokalisere fotografiet på selve filmstripene. /Bilde-nr/ er da relatert til kilden, og jeg velger å innføre et subelement for å vise at dette er relatert til kilden. Elementet blir da DC.Source.Bildenummer, uten bruk av forkortelser.

Det er også behov for elementet /**side/** da det brukes ofte ved metadatasøk, p.g.a av visse bilder er plassert eller antatt forekommet på visse sider i avisen.

/Film-nr/ viser hvor mange filmmapper som er brukt i forbindelse med en sak. Dette feltet dekkes av DC.Relation.Filmmappe der en direkte refererer til ID'ene til de andre filmmappene som er brukt i saken. Dette relasjonselementet refererer da til andre filmmapper og til de filmmapper som var brukt sammen første gang saken sto på trykk. Den viser dermed ikke relasjon til filmmapper med bilder av lignende saker, men det er også en mulighet å vurdere. Art Museum Image

Consortium (AMICO) [49] bruker blant annet DC.Relation.Image, men siden dette er en relasjon til en annen filmmappe, og ikke en sak eller noe, er det naturlig at vi her innfører vårt eget sub-element til Relation. Vi får da DC.Relation.Filmmappe der den unike ID til filmmappe som inngår er referert til. *F.eks DC.Relation.Filmmappe="F-980530/45881"*.

Metadata som ikke mappes direkte av DC, men beholdes:

- **/side/**
- **/DC.Source.Bildenummer/.**
- **/merknad/**
- /**retur/**

Kommentarer til implementering av DC:

Elementene som her ikke kan mappes, men som allikevel er nødvendig å ha med kan tilfredsstilles i XML ved implementering. Her har du mulighet til å lage grupperinger slik at en viser relasjoner mellom de ulike elementene. F.eks en filmmappe vil ha flere motiv beskrevet og da kan du gjenta f.eks <bilde> tag'en i XML og definere hvilke element du skal bruke til å beskrive bildet, som finnes på film-mappen. Du får da tre grupperinger i XML, d.v.s de opplysninger som følger filmmappen, saken og selve bildemotivet. Her har du også mulighet til å beskrive hvert motiv med de elementene som er uforanderlige som ID og beskrivelse, og opprettet et element for <brukt-bilde> som gjentas med DC.Identifier, side, produkt m.m for hver gang bildet var brukt. Se eksempel , (s165).

Indeksering i XML:

```
<Bilde>
    <DC.Identifier>...Negativnr.</DC.Identifier>                    = erstatter /bilde-nr/
    <DC.Date.Issued>....</DC.Date.Issued>                           = /Prod-dato/ første gang bildet ble brukt.
    <DC.Relation.IsPartOf> Ut magasinet </DC.Relation.IsPartOf>      = produkt hvor bildet var brukt første gang.
    <DC.Description>.... </DC.Description>                          = /Motiv/
    <Side>....</Side>
    <DC.Description.Genre> Landskapsmotiv </DC.Description.Genre>
    <DC.Relation.Filmmappe.1>"F-000102/45520"</DC.Relation.Filmmappe.1>
    <DC.Relation.Filmmappe.2>"F-000102/45522"</DC.Relation.Filmmappe.2>
</Bilde>
```

Gjentas for hver gang bildet står på trykk:

```
<brukt-bilde>
```

<DC.Source.Bildenummerr>....</DC.Source.Bildenummer>
<side>....</side>
<DC.Relation.IsPartOf>....</DC.Relation.IsPartOf>
<DC.Date.Issued>....<DC.Date.Issued>
<DC.Relation.Filmmappe.2>"F-000102/45522"</DC.Relation.Sak.2>
</brukt-bilde>

Det vil også være noen metadata som er felles for hele film-mappen.

Eksempel 8.2. Eksempel på mapping

Her vil jeg foreta en eksempel på mapping av DC for en indeksert filmmappe i SIFT filmarkiv. Den filmmappen som her mappes er illustrert i figur 8.2, (s166)., og selve mappingen i DC for denne vises i tabell 8.4, (s166).

```
Adresseavisens filmarkiv     Film                                    INDEKSERT

Jobb-id    : J-971117/912    Film-nr   : 01 av 01      Film-id: F-971117/47258
Film-dato  : 971115          Film-type: farve neg      Prod-dato: 971117

Fotograf   : Asphaug, Aage
Journalist: Eidsvåg, Terje

Gruppe     : 01 Hølmebakk, Gordon - forlegger

Sak        : Hovedgjest på seminaret "Nykle in memoriam"
Person     : portrett

Retur      :

Bilde-nr   : 21      Brukt-dato: 971117  Produkt: ADR              Side: 1
Motiv      : Gordon Hølmebakk - utendørs i hatt - portrett
                                                                  Papirkopi:

Bilde-nr   : 14      Brukt-dato: 971117  Produkt: ADR              Side: 27
Motiv      : På pub med et glass øl - portrett
                                                                  Papirkopi:
```

FIGUR 8.2. Indeksert film i SIFT filmarkiv

Tabell 8.4. Eksempel på Mapping til DC for filmformatet

DC METADATA	INNHOLD	KOMMENTARER
DC.Identifier.Film	J-971317/912	Skjema=Film
DC.Identifier.Jobb	F-97117/47258	Skjema = Film
DC.Date.Created	971115	Erstatter /film-dato/

Tabell 8.4. Eksempel på Mapping til DC for filmformatet

DC METADATA	INNHOLD	KOMMENTARER
DC.Relation.Filmmappe	"ingen relasjon til andre filmmapper her"	Erstatter /film-nr/
DC.Source.Type	farve negativ	
DC.Creator.Agentrole	Fotograf	
DC.Creator.Agentname	Asphaug, Åge	
DC.Creator.AgentAffiliation	Adresseavisen	
DC.Contributor.Agentrole	Journalist	
DC.Contributor.Agentname	Eidsvåg, Terje	
DC.Contributor.Agentaffiliation	Adresseavisen	
DC.Subject.Classification	01 Hølmebakk, Gordon - forlegger	
DC.Relation.Sak	Hovedgjest på seminaret "Mykle in memo-riam"	Gjelder for begge motivene
DC.Source.Bildenummer	21	
DC.Date.Issued	971117	
DC.Relation.IsPartOf	"Adresseavisen"	
DC.Description	Gordon Hølmebakk - utendørs i hatt	
Side	1	
DC.Source.Bildenummer	14	
DC.Description.Genre	Portrett	
DC.Date.Issued	971117	
DC.Relation.IsPartOf	Adresseavisen	
DC.Description	På pub med et glass øl	
Side	27	
Repeterende element		
DC.Date.Issued	Intet innhold her da ingen av motivene i dette eksemplet er brukt mer enn en gang. Ingen gjentakelse av DC.Relation.Filmmappe her, da saken har bare en filmmappe.	Gjentas neste gang motivet er brukt. Denne elementet kan der-for få flere repitisjoner knyttet til hvert bildenummer. Alle disse elementene gjentas rett bak DC.Source.Bildenummer.
DC.Relation.IsPartOf		
Side		
DC.Description.Genre		
DC.Relation.Sak		
DC.Relation.Filmmappe		

Eksemplet som vist i figur 8.2, (s166) ble hentet ut av SIFT arkivet nylig, og her er elementet **/person/** lagt til. Det er ment å dekke kommentar som "portrett" som tidligere ble lagt i **/merknad/,** eller tatt med i selve motivbeskrivelsen. I denne eksemplet er det gjentatt i begge element, noe som er unødvendig. Dersom det er ønskelig at dette skal gjøres søkbart er det bedre å opprette et element for genre, som legges som sub-element til beskrivelseselement. Vi får DC.Description.Genre.

8.3.3. Mapping av Ad-hoc formatet for stoff/artikler til Dublin Core

Mappingen er her oppsummert i tabell 8.5, (s168).

Tabell 8.5. Mappingoversikt av formatet for artikler/Stoff i SIFT til Dublin Core

DC METADATA	**SIFT METADATA FOR ARTIKLER**	**KOMMENTARER MED HENSYN PÅ SKJEMA OG VERDIER SOM BENYTTES**
DC.Date.Isssued	Prod-dato	Følger utgivelsesdato.
DC.Identifier	Navn (Art-nr)	Skjema="Artikkel".
DC.Relation.IsPartOf	Produkt	Artikkelen er en del av Ut-magasinet, Uke-adressa osv.
DC.Creator.Agentname DC.Creator.Agentrole DC.Creator.AgentAffiliation	Journalist	Opphavsperson til artikkel.
DC.Subject.Classification	Emneord	Skjema="Egen liste med 55 emneord".
DC.Title	Stikkord	Her kommer tittel inn fra CCi, og brukes ikke til frie emneord. Elementet bør også gjøres søkbart.
DC.Subject	Merknad	Merknad brukes her til frie emneord, og det mappes av DC.Subject.
Element som ikke mappes		
- - -	Side	Beholdes.
- - -	Ant-linjer	Unødvendig element. Utgår fra anbefalingen.
- - -	Tekst	/tekst/ viser her til selve artikkelteksten. I HTML er dette tilsvarende med tag'en <Body>. Beholdes slik det er.
Bidrag:		
DC.Relation.Bilde DC.Relation.Tegning DC.Relation.Illustrasjon DC.Relation.Film osv.	Illustrasj	Viser til type bilde som er brukt og er dermed å regne som en relasjon til artikkelen. Relasjonselementet anbefales her å vise direkte til relasjonsobjektet sin ID. Vil også indikere hvilket arkiv bildet er lagret.
Tilleggselement som anbefales brukt (Se element for Internettpublisering av artikler i avsnitt 8.3.5, (s176).)		
DC.Contributor.Agent-Affiliation DC.Contributor.Agentname DC.Contributor.Agentrole	Kilde	Innholder hvem som har bidratt med illustrasjon, fotografi til bildet. Her kan opphavsperson være både en person og en organisasjon derfor bør også Agentrole benyttes. Om kilden tilhører egen organisasjon eller ikke blir også lagt i Kilde feltet i dag. Dette behovet kan erstattes ved å fylle ut Opphavspersonens arbeidsorganisasjon.
Ingen repeterende element (Artikler brukes ikke flere ganger på trykk i avisen)		

8.3.3.1. Drøfting av mappingen

Elementer som her ikke mappes er: /**Side/antall-linjer/tekst/**.

/Antall-linjer/ er et element som det ikke søkes på, og en trenger derfor ikke beholde dette elementet. **/Side/** og **/tekst**/ er det behov for å beholde, og disse elementene kommer i tillegg til DC sine element. I tabell 8.5, (s168). er de elementene **/Prod-dato/Produkt/Journalist/** mappet på tilsvarende måte som for filmformatet og illustrasjonsformatet.

Navn Navn er her den unike identifikatoren til artikkelen. Mappes av DC.Identifier.

Stikkord **/Stikkord/** er her ikke i her brukt til frie emneord, men er tittel som overføres fra CCI systemet når artikkelen importeres til SIFT. Siden tittel importeres her, samsvarer det med tittel som brukes i CCI WORD og i Internettformatet. Tittel tilfredsstilles i DC.Title.

Emneord **/Emneord/** velges fra kontrollert vokabular, og DC.Subject.Classification kan her brukes.

Merknad Elementet /**merknad/** brukes ikke til det ordet betyr, men til frie emneord. Frie emneord mappes av DC.Subject. /**merknad**/ elementet kan beholdes, men da er det tenkt å brukes nettopp til "merknader".

Tekst Artikkel er et fulltekstdokument som elementet **/tekst/** viser til. Dette er tilsvarende <body> tag'en som brukes i HTML dokumentet for å referere til hvor i dokumentet teksten forekommer. DC mapper ikke dette direkte. Jeg synes ikke det er riktig å legge det i DC sitt beskrivelseselement, siden det ikke er en beskrivelse av ressursen, men ressursen i seg selv. /**tekst/** beholdes derfor slik det er. Elementet brukes også i CCI Word, og i Internettformatet heter det /**brødtekst/.**

Illustrasj Dette elementet viser til bidraget til artikkelen i form av en tegning, et bilde eller lignende. Dette er å regne som en relasjon til artikkelen. DC har flere sub-element for relasjon som her kan benyttes. Anbefalt bruk av relasjonselementet fra DCMI, er at verdien til relasjonselementet skal vise til relasjonsobjektet sin ID[1].

IsPartOf er et sub-element som kan benyttes, men dette benyttes for elementet **/produkt/.** Her kan en innføre et sub-element for hver av de relasjonsobjekter vi har med å gjøre, det være et bilde, illustrasjon eller film. Vi får da DC.Relation.Bilde, DC.Relation.Film og DC.Relation.Illustrasjon som viser til den unike ID for hver av disse, f.eks DC.Relation.Film="F-000302/45536" som viser til filmen bildet som er brukt er å finne på. Dersom bildet også er lagret i Fotostation, kan en også har referanse til det. Når en søker opp saken her, kan en direkte hente filmen i arkivet hvor bildet til artikkelen befinner seg. Her kan en også finne eventuelle bilder som kan tenkes gjenbrukt. Med denne relasjonen er det også raskt å søke opp den aktuelle filmmappen i SIFT filmarkiv, og dermed få en beskrivelse til alle bildene som er brukt her.

Her kan en også vurdere om det er aktuelt å innføre et relasjonselement som viser til andre artikler som er relatert til artikkelen på noe vis. Dette vil være en utvidelse av formatet i forhold til slik det brukes p.r. i dag.

Internettformtaet er et rikere format enn SIFT formatet for artikler, da det har flere beskrivelseselement. Flere av de samme behovene som her tas opp, vil også være tilfelle når Internettformatet mappes i avsnitt 8.3.5, (s176). I dette avsnittet vil jeg også foreta en eksempel på mapping av en artikkel hvor relasjoner til begge formatene blir påvist.

8.3.4. Mapping av Ad-hoc formatet for digitale bilder i Fotostation

Her har en forsøkt å gjenbruke endel av de samme feltene som er brukt for indeksering av film og illustrasjoner i SIFT formatene, men det er ikke samsvar overalt. En må se dette formatet i sammenheng med SIFT formatene for indeksering av bilder (illustrasjoner og film), fordi noe er likt og annet er ulikt. Dette fordi det i prinsippet dreier seg om dobbellagring av bilder, og at det hadde vært tilstrekkelig med ett av verktøyene til dette. Det er også snakk om at Fotostation skal erstatte SIFT filmarkiv på sikt. Dette vil være en fordel siden Fotostation lagrer bildene som egne objekt i motsetning til SIFT filmformat. Det er likevel fullt mulig å tilpasse Fotostation til å ivareta arbeidsprosesser som tidligere har vært utført i SIFT, *f.eks utvide til flere muligheter for metadatasøk*. Oppsummering av mappingen vises i tabell 8.6, (s172).

1. ID=Identifikator

8.3.4.1. Drøfting av mappingen

Ut fra det behov som eksisterer allerede er det feltene **/motiv-nr/side/returnert til/**, **instruksjoner/merknad/** som ikke mappes av et element i DC direkte.

Også her er det behov for å beholde elementet /**merknad/** og **/retur/** elementet. Elementet **/returnert til**/ blir da kalt for **/retur/** slik det gjøres i SIFT formatene. Fotostation har i tillegg egendefinerte felt som kan brukes for gjentakelse av **/side//DC.Date.Issued/DC.Relation. IsPartOf/** for hver gang bildet var brukt. Disse egendefinerte feltene i dette formatet brukes ikke pr. dato. Det er også mulig å definere element for **/side/** og andre gjentakbare element, som kan ta imot nye verdier for hver gang de ble brukt. Disse må stå etter hverandre (loddrett eller vannrett) for å påvise relasjoner dem i mellom, *f.eks hvilken siden som hører til hvilken datoangivelse osv.*

I det tilfellet når bildet er scannet inn fra negativ film, er negativ filmen å regne som en kilde som har ført til frembringelse av bildet. Referanse til filmmappen med filmstripene i vil da kunne beregnes som en relasjon til det digitale bildet. Elementet **/Film-nr/** her kan derfor tilfredsstilles av DC.Source og vi kan utvide med DC.Source.Type="negativ film". Her er det viktig å være oppmerksom på at SIFT filmformat også bruker et element ved navn **/film-nr/**. Dette er ikke tilsvarende Fotostation sitt **/film-nr/** som er det samme som /**film-id/** i SiFT film format.

Sub-elementet Type som benyttes her er også brukt av Anne Marie Vercoustre [37] i hennes prosjekt. Videre har også Anne Marie Vercoustre foreslått et sub-element til DC.Format som identifiserer verdier for kvaliteten på bildet. Dette synes jeg er en god ide å benytte også her, siden bildene har forskjellig kvalitet alt etter hva som er brukt for å produsere bildet. Alt etter om bilder er scannet fra negativ film, fysisk tegning, papirkopi osv. vil også dette ha innvirkning på kvali-teten av bildet. Elementet som tilfresstiller dette er innføring av subelementet Quality for DC.Type som gir DC.Type.Quality. I tillegg vil jeg også foreslå bruk av DC.Type for å angi hvilken farge som er på bildet med sub-elementet Farge, f.eks DC.Type.Farge = "farge fotografi". Bruk av DC.Rights er også veldig aktuell å bruke ved publisering via Internett, og også elementet DC.Coverage.Place som refererer til bestemte områder som bildet er tatt i. *F.eks er bildet fra Fosen eller et annet sted i Trondheim kan det godt være aktuelt å bruke DC.Coverage.Place="Fosen".* I forbindelse med bestillinger av bilder etterspørres ofte portrettbilder, høydebilder (Vertikale bilder) eller breddebilde (horisontale bilder). Dette kan tilfredsstilles ved å opprette et sub-element, f.eks DC.Format.**Orientation** hvor verdiene er "vertikal" og horisontal. Portrettbilder går mer på hvilken type bilde det er med hensyn på motivet, og

Tabell 8.6. Mapping av formatet for digitale bilder i Fotostation til DC

DC METADATA	METADATA FOR FOTOSTATION	KOMMENTARER MED HENSYN PÅ SKJEMA OG VERDIER SOM BENYTTES
DC.Identifier	Gruppe (Objektnavn)	
DC.Relation.Sak DC.Title	Sak	Tittel fungerer bedre her enn i SIFT, da tittel her følger hvert enkelt bilde. Tittel her blir mer oversiktlig enn i SIFT hvor tittel må repeteres for hvert bilde når du har opptil 5 bilder som kan gjenbrukes som er registrert på en enhet.
DC.Creator.Agentname DC.Creator.Agentrole DC.Creator.AgentAffiliation	Fotograf	
DC.Description	Motiv	
DC.Type	Hva er scannet	Hvilken genre av bilder hører ressursen innunder, er det en tegning, en plakat, et cd-cover, spillcover, bokomslag osv.
DC.Contributor.Agentname DC.Contributor.Agentrole DC.Contributor.AgentAffiliation	Journalist	
DC.Relation.IsPartOf	Produkt	Produktet som bilde sto i første gang det ble brukt
DC.Subject.Classification	Gruppe (Nøkkelord)	
DC.Date.Created	Fotodato	Det samme som /Film-dato/ i SIFT
DC.Date.Issued	Utgivelsesdato	
DC.Date.Issued	Brukt-dato	Først gang bildet blir brukt blir /side/ og /brukt-dato/ lagret i egne felt
DC. Source.Type	Film-nr (Utgjør ett felt i Fotostation sammen med motiv-nr)	Denne benyttes når kilden har en unik ID som du kan referere til, som er tilfelle for negativ film. DC.Source vil da ha samme verdi som **/film-id/** i SIFT filmformat. f.eks " 45520", hvor 45520 er ID'en.
DC.Source.Type	- - -	Verdier her blir om det er digital film, negativ film, papirkopi m.m Dette er kilden bildet scannes fra.
DC.Source.Bildenummer	Motiv-nr + (Utgjør ett felt i Fotostation sammen med Film-nr)	Det samme som **/bilde-nr/** i SIFT. Ikke alle kilder trenger å bruke DC.Source.Bildenummer, men det trenger negativ film.
- - -	Returnert av	Beholdes, men får navnet **/retur/** som for filmformatet og illustrasjonsformatet.

Tabell 8.6. Mapping av formatet for digitale bilder i Fotostation til DC

- - -	Merknad	Beholdes.
- - -	Instruksjoner	Ikke nødvendig. Utgår.
- - -	Side	Beholdes.
Anbefalte element å bruke i tillegg [Se "Tilleggsbehov for bilder" avsnitt 8.4, (s184)]		
DC.Subject		Frie emneord. Benyttes av artikkelformatene.
DC.Description		Dette er de samme element som beskrevet og anbefalt i tilegg for illustrasjonformatet i SIFT. Forklaring i tabell 8.1, (s152).
DC.Rights		
DC.Type.Quality		
DC.Format. Min		
DC.Format.Max		
DC.Description.Genre		
DC.Coverage.Place		F.eks her brukes Geografisk navn etter et kontrollert vokabular.
Repeterende element	**(Repeteteres for hver gang illustrasjonene er brukt)**	
DC.Date.Issued DC.Relation.IsPartOf Side DC.Relation.Sak		Opprettes element for dette som er mulig å utvide med flere felt av samme type, og som kommer etter hverandre i rekkefølge for å påvise relasjon.

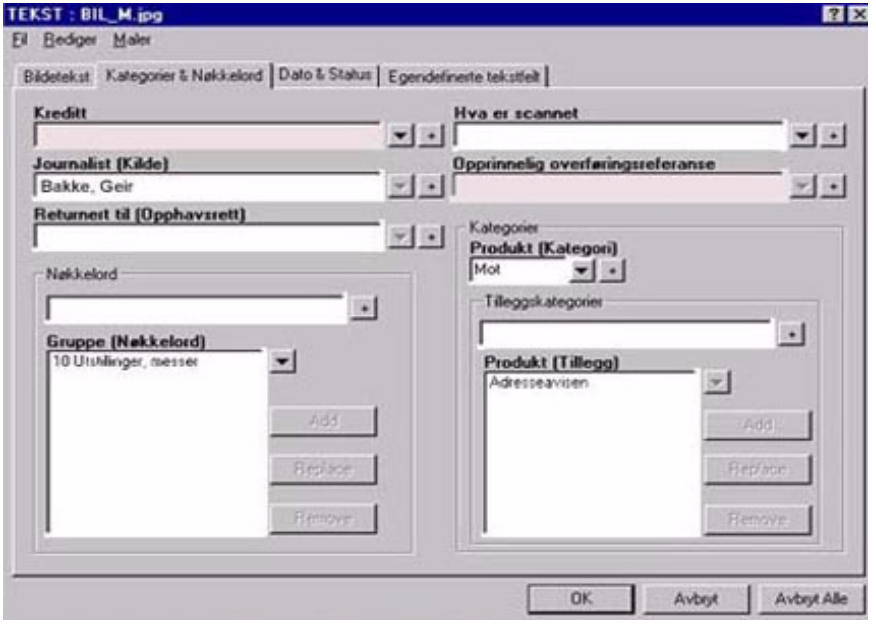
representrer mer en genre for motivet, dvs du har portrettbilder, du har landskapsbilder m.m. Her foreslås innføring av elementet DC.Description.Genre for hvert motiv.

/Sak/ mappes ikke direkte mot DC.Title, men mot DC.Relation.Sak. Deretter innføres også DC.title for bildet som er en av de metadataelement som er veldig ofte benyttet ved indeksering av bilder som tidligere nevnt i dette kapittelet.

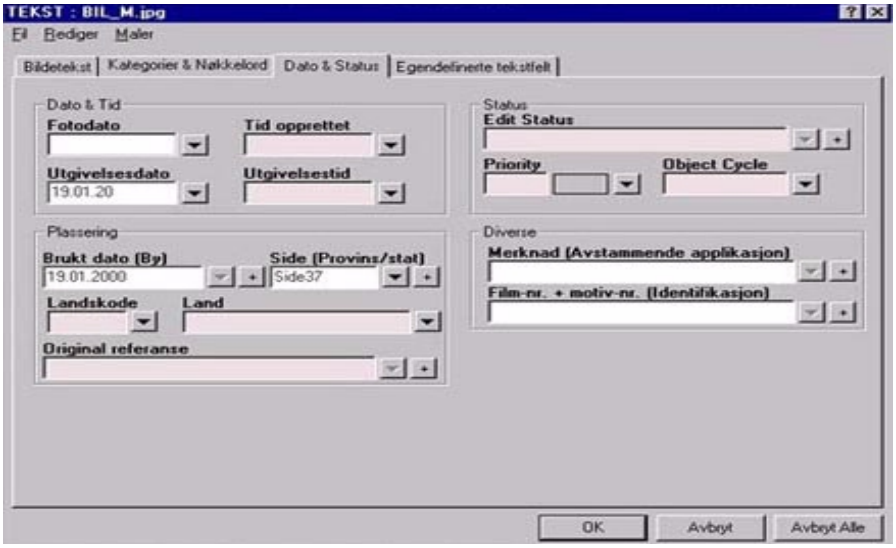
Eksempel 8.3.

Eksempel på mapping her vil ikke være mye forskjellig fra eksemplifisering som er vist for SIFT formatene. Jeg tar i dette eksempelet utgangspunkt i det digitale bildet som er indeksert i Fotostation, og vises i figur 8.3, (s174), 8.4, (s175) og 8.5, (s175).Mappingen vises i tabell 8.7, (s175).

FIGUR 8.3. Digitalt bilde indeksert i Fotostation, (skjermbilde 1)



FIGUR 8.4. Digitalt bilde indeksert i Fotostation, (skjermbilde 2)



FIGUR 8.5. Digitalt bilde indeksert i Fotostation, (skjermbilde 3)

Tabell 8.7. Eksempel på Mapping til DC for formatet for digitale bilder i Fotostation

DC METADATA	INNHOLD	KOMMENTARER
DC.Identifier		Slik det ser ut her er det filnavnet på bildet som er unik, dvs "Bil_m.jpg"
DC.Date.Issued	"19012000"	
DC.Creator.Agentrole	Fotograf	
DC.Creator.Agentname	Geir Bakke	
DC.Creator.AgentAffiliation	Adresseavisen	
DC.Contributor.Agentrole	Journalist	
DC.Contributor.Agentname	Geir Bakke	
DC.Contributor.AgentAffiliation	Adresseavisen	
DC.Subject.Classification	10 Utstillinger, messer	
DC.Subject	"Bilutstilling - Detroit"	Frie emneord.
DC.Type	"Digital bilde"/.jpg	
DC.Description	"Bildet av hele dashbordet i nærprofil - ratt t.v"	

Tabell 8.7. Eksempel på Mapping til DC for formatet for digitale bilder i Fotostation

DC METADATA	INNHOLD	KOMMENTARER
DC.Description.Billedtekst	"Hele dashbordet som TV-skjerm, med direkte adgang til Internett, TV, Mobiltelefon, navigasjonssystemer, alt stemme styrt slik at du kan holde hendene på rattet"	
Side	37	
DC.Title	Internett i bilen	
DC.Relation.IsPartOf	Adresseavisen	
DC.Relation.Sak.1	Bilutstilling i Detroit	Første gang bildet brukes.
DC.Source	Digitalt kamera	
Element som anbefales brukt i tillegg		
DC.Type.Quality	Middels	Her må faste verdier bestemmes.
osv. osv....		
Repeterende element		
DC.Relation.IsPartOf		
DC.Relation.Sak.n		**n** indikerer hvor mange gang er bildet er blitt brukt. F.eks DC.Relation.Sak.2, DC.Relation.Sak.3 osv.
Side		
DC.Date.Issued		

8.3.5. Mapping av Ad-hoc formatet for Internettpublisering

Adresseavisen ønsker å bruke XML for å motta stoff fra eksterne kilder. En DTD for XML basert på metadata i tabell 6.5, (s98) og 6.6, (s99). er allerede definert, og dermed kan denne omdefineres dersom det viser seg at DC kan tilfredsstille behovet like bra som eksisterende metadata.

Tabell 8.8. Mappingoversikt av Internettpubliseringsformatet til DC for artikkel

DC METADATA	METADATA FOR INTERNETTFORMATET	KOMMENTARER MED HENSYN PÅ SKJEMA OG VERDIER SOM BENYTTES
DC.Identifier	Artikkelref	f.eks **http://www.adressa.no/artikkel.awml?artikkelref=19012000AM1U4JA.** Skjema=Artikkel formatet.
DC.Title	Tittel	
DC.Title.Alternate	Undertittel	
DC.Subject.Classification	kategori	F.eks: DC.Subject.Classification=bil skjema=Hovedkategori, og DC.Subject.Classification=bil og motor skjema=Temakategori
	temakategori	
DC.Date.Issued	Publisert dato	
DC.Date.Sperredato	Sperredato	Administrativ metadata
DC.Creator.Agentname DC.Creator.Agentrole DC.Creator.Agent-Affiliation	Byline	Dette er vedkommende som har skrevet artikkelen. DC.Creator.Agentname dekker dette. Her anbefales allikevel å ta med Agentrole og AgentAffiliation.
DC.Description.Ingress	Ingress	
DC.Title.Alternate eller DC.Title.Forsidetittel	Forsidetittel	Her kan en dekke denne type tittel i Alternate feltet som anbefales brukt av DCMI for alle andre titler enn hovedtittel. En kan også innføre sitt eget sub-element som nettopp er forsidetittel, for å vise hvilken type tittel det gjelder i denne sammenheng.
DC.Description.Forside-tekst	Forsidetekst	
DC.Subject	Emneord	Frie Emneord.
- - -	Status	Administrativ metadata.
- - -	Brødtekst	For å få samsvar med SIFT formatet kan bare **/tekst/** benyttes. **/tekst/** er metadatanavnet som også benyttes i CCI Word for selve artikkelteksten.

8.3.5.1. Drøfting av mappingen

Artikkelref Dette er primærnøkkelen i relasjonsdatabasen og er altså unik. Mappes av DC.Identifier. Her defineres skjema til å være artikkel, for å til hvilket informasjonsobjekt som indekseres etter dette formatet.

Tittel	Dette er tittel og mapper DC.Title
Undertittel	Dette kan dekkes av sub-elementet til tittel i DC. Vi får da DC.Title.Alternate her. Dette feltet er ikke så mye brukt av Adresse--avisen.
Kategori	Dette er hovedkategori eller seksjonskategori og er valgt ut fra en kontrollert vokabular fra en liste over faste kategorier. Mapper DC.Subject.Classification. *F.eks er kategori=bil for artikkelen i figur 8.6, (s182). når den legges ut på Internett.*
Temakategori	Dette er en underkategori av Kategori elementet. *F.eks er temakategori=bil og motor for artikkelen i figur 8.6, (s182). når den legges ut på Internett.* Her kan en også bruke DC.Subject.Classification, og en kan egentlig sette det inn under samme kategori, dvs at du får DC.Subject.Classification=bil, bil og motor. Det jeg tenker her er at både temakategori og kategori (som er hovedkategori) kan dekkes av DC.Subject.Classification fordi begge er kategorisering etter emne. Her kan en legge inn skjema=hovedkategori og skjema=underkategori.
Publisert dato	Den dagen artikkelen legges ut på internett. Mapper DC.Date.Issued.
Kommentar	Mapper ikke DC. Dette elementet brukes som Clipboard internt og kan fortsette med det, men elementet har ingen interesse ved distribusjon av artikkelen. Tidligere har arbeidsgruppen for beskrivelse og tema (dvs subject and description) foreslått et sub-element til beskrivelseselementet som heter note. Det kan fungere i dette tilfellet dersom det er noe som omfatter selve bildet. Dette elementet dekker det samme be-hovet for tilleggsopplysninger som **/merknad/** elementet i SIFT og Fotostation, og kommentarer kan her erstattes av et **/merknad/** felt.
Sperredato	Slik dette elementet brukes er det slik at dersom det legges ut i basen før det er publisert på Internett, vil dato for når den skal legges ut stå her. I så møte er den ikke tilgjengelig før den dagen den står i Internettutgaven. DC har ikke et sub-element under DC.Date som matcher dette og vi innfører derfor et for sperredato, d.v.s DC.Date.Sperredato.
Status	Har ikke funnet noe element som mapper DC her. Her er det en kort kommentar som legges inn f.eks "Åpen inntil sperredato". Jeg ser ikke

det helt store behovet for et Status felt her som ikke kan dekkes i f.eks et kommentar eller merknadselement. Det er ikke behov for å søke på dette elementet, men kan være greit å beholde for å angi statusen til artikkelen.

Byline Dette er den som har skrevet artikkelen og her kan det også defineres hvilken rolle vedkommende har. Mapper DC.Creator.Agentrole= Journalist og DC.Creator.Agentname som er navnet på journalisten. Journalist kan være ekstern og en må derfor også ta med organisasjonen som opphavspersonen tilhører, som dekkes av DC.Creator.AgentAffilation.

Ingress Dette er en beskrivelse eller en slags innledning til artikkelen og er dermed å regne som en beskrivelse og vil dermed mappe DC.Description. Siden vi har flere beskrivelser her kan det være aktuell å utvide formatet her til DC.Description.ingress for å definere hvilken type beskrivelse dette her. Ingress er en beskrivelse som brukes mye i avissammenheng og dermed er ikke en slik utvidelse en løsning her.

Brødtekst Dette er resten av teksten i hele artikkelen. I den DTD'en i XML som Adresseavisen har definert er elementet tekst brukt. Dette kan videreføres og settes som DC.Description.tekst, men det er ikke en beskrivelse av innholdet siden dette er fulltekstdokumentet og derfor ikke data om data. Brødtekst kan være et eget element utover DC som kan spesifiseres i XML f.eks <Artikkeltekst>, som viser at det er teksten til artikkelen. Allikevel er dette et fulltekst dokument og for å skille mellom andre beskrivelser, når metadata er integret i selve dokumentet, kan en bruke DC.Description.fulltekst.

Forsidetittel Dette er enda en tittel som er brukt og den brukes på førstesiden til Internettutgaven og er dermed å regne som kort notis som refererer deg videre til artikkelen. og kan dermed tekkes av Tittel feltet til DC, i f.eks DC.Title.Alternative, eller en kunne også ha definert et DC.Title.Forsidetittel.

Forsidetekst Dette er et element som følger forsidetittel, og er den teksten som står i førstesidenotisen i Internettutgaven. I SIFT blir en artikkel som også har stått på forsiden i papirutgaven lagret som to enheter. I Internettfor-matet blir førstesideoppslaget (som er en kort notis) av

Internettutgaven arkivert sammen med hovedartikkelen. Forsidetekst er på denne måten en type sammendrag av artikkelen som kommer lengre bak. Vi setter det under DC.Description.Forsidetekst.

Emneord Dette er en liste med frie emneord og kan legges i DC.Subject.

Vedlegg Referanse til vedlegget som er et bilde. Vi får her DC.Relation.Image, som da viser til bildet sin unike ID. F.eks DC.Relation.Image=Vedleggsnummer. Vedleggsnummer er metadatanavnet til den unike ID til vedlegget som Internettredaksjonen bruker. Dette er vist i tabell 8.9, (s180). Bildevedlegget har sine egne metadata som følger det, siden det eksisterer som eget objekt.

Tabell 8.9, (s180). viser de metadata som Internettredaksjonen påfører sine bildevedlegg, og hvordan disse mappes mot DC element. Selv om vedlegget eksisterer som eget objekt følger det artikkelen. /**Billedtekst/** påføres i CCI WORD og denne overføres også i råtekst til SIFT som en del av /**tekst/** elementet.

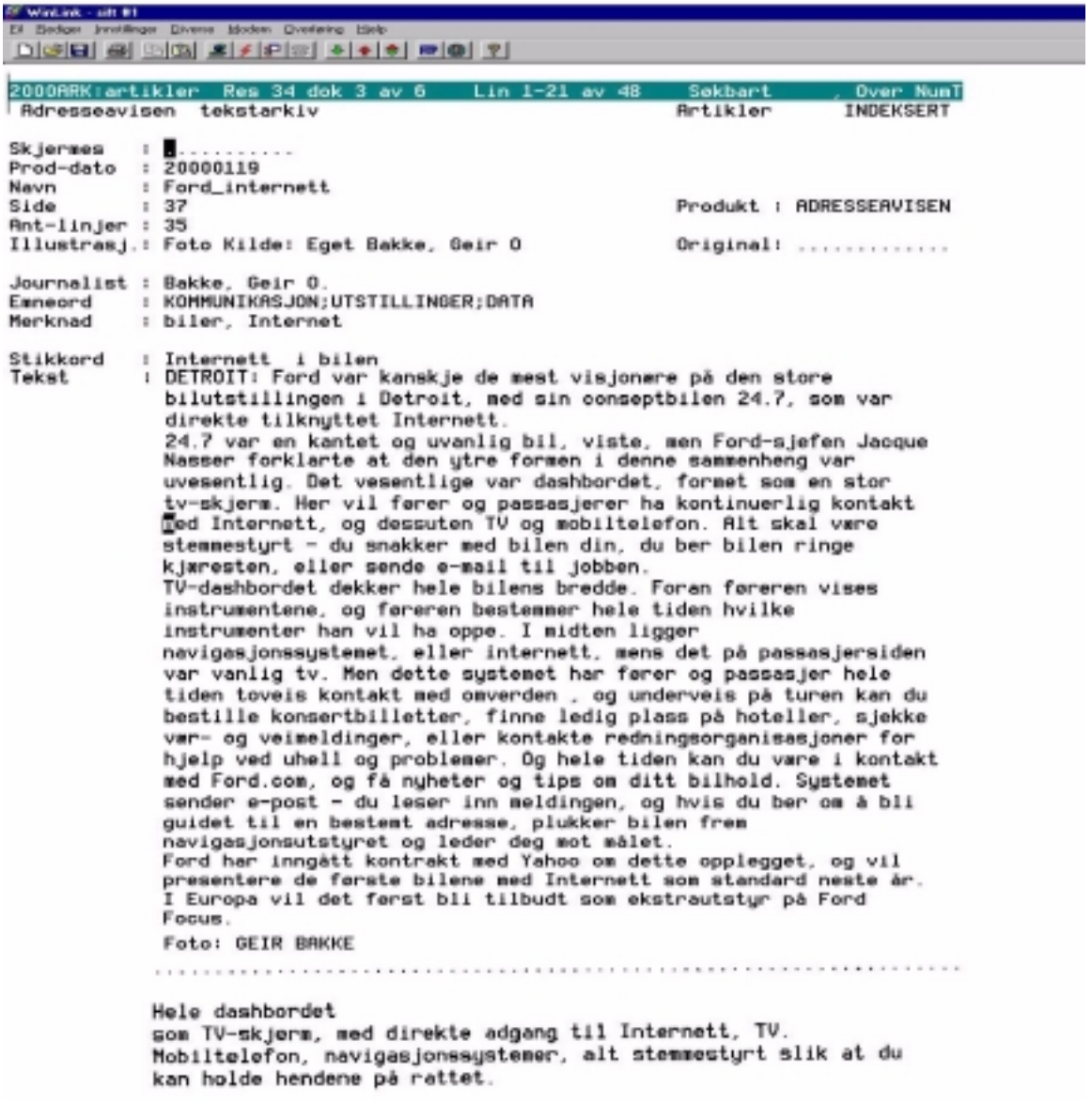
Tabell 8.9. Mappingoversikt av Internettpubliseringsformatet til DC for vedlegg

DC METADATA	**METADATANAVN VEDLEGGET**	**KOMMENTARER MED HENSYN PÅ SKJEMA OG VERDIER SOM BENYTTES**
DC.Format	Filnavn	Bilde/.jpg.
DC.Identifier	Vedleggsnummer	Dette er den unike ID til bildet her.
DC.Format.Extent	Bredde	Er nødvendig å vite bildet sitt størrelse, derom ikke fast størrelse er bestemt på forhånd. Det er nødvendig å skille ut bredde og høyde for plassering i skjermen av Internettutgaven, og de har derfor fått hver sitt sub-element. DC.Coverage.Place med bruk av x og y koordinater kan muligens brukes her, men er vel mer tenkt brukes for geografisk utstrekning.
	Høyde	

Tabell 8.9. Mappingoversikt av Internettpubliseringsformatet til DC for vedlegg

DC METADATA	METADATANAVN VEDLEGGET	KOMMENTARER MED HENSYN PÅ SKJEMA OG VERDIER SOM BENYTTES
DC.Creator.Agentname DC.Creator.Agentrole DC.Creator.Agentaffiliation	Byline	Opphavsperson, dvs fotografen.
DC.Description.Billedtekst	Tekst (Billeddeksten)	Tekst til bildet som skal stå på trykk i Internettutgaven.

Med utgangspunkt i artikkelen i figur 8.6, (s182). kan mapping i DC tilfredsstilles som vist i tabell 8.10, (s183).



FIGUR 8.6. Artikkel indeksert i SIFT metadataformat for stoff/artikler

Tabell 8.10. Eksempel på indeksering i DC av artikkel i figur 8.6, (s182). med utgangspunkt i Internettformatet

DC ELEMENT	SCEME/SKJEMA	CONTENT/INNHOLD	METADATA I SIFT FR ARTIKLER/ STOFF
DC.Title		Internett i bilen	**/Stikkord/**
DC.Creator.Agentrole		Journalist	- - -
DC.Creator.Agentname		Geir O. Bakke	**/Journalist/**
DC.Subject.Classification	"Egent kontrollert liste"	Kommunikasjon, utstillinger, DATA	**/Emneord/**
DC.Subject	"Fri emneord"	biler, internett	**/Merknad/**
DC.Description.ingress		Detroit: Ford var kanskje den mest visjonare på bilutstillingen i Detroit, men sin conseptbil 24.7, som var direkte tilknyttet Internett.	**/Tekst/**
DC.Description.Billedtekst		Hele dashbordet som Tv-skjerm, med direkte tilgang til Internett, TV, mobiltelefon, navigasjonssystemer, alt stemmestyrt slik at du kan holde hendene på rattet	
DC.Publisher		Adresseavisen	
DC.Date.Issued	"ANSI X3.30-1985"	"20000119"	**/Prod-dato/**
DC.Type		Artikkel	- - -
DC.Identifier	"Navnesystem"	Ford_Internett	**/Navn/**, (**/Art-nr/**)
DC.Language	"ISO 693"	NO	- - -
DC. Relation.IsPartOf		Adresseavisen	**/Produkt/**
DC.Coverage.Place		Midt-Norge eller annet behov for dekning knyttet til tid og sted.	- - -
DC.Contributor.Agentrole		Fotograf	- - -
DC.Contributor.Agentname		Geir O. Bakke	**/kilde/**og **/tekst/**
DC.Contributor.AgentAffiliation		Adresseavisen (kunne også inneholdt AP/NTB m.m)	I **/Kilde/** sammen med fotografnavnet
DC.Rights		Adresseavisen	- - -
DC.Rights		Geir O. Bakke	- - -
Tilleggselement som DC ikke har			
Tekst		"Hele innholdet i artikkelen"	**/Tekst/**

8.4. Oppsummering av tilleggsbehov for bilder

Det er drøftet under noen av de ulike mappingsavsnittene sub-element som kan være aktuell å tilføre til bilder. Dette er blant annet basert på sub-element som også A.M.V. [37] benytter, men også egne sub-element er foreslått.

De tilleggselement som er drøftet er:

- DC.Type.Quality
- DC.Description.Genre
- DC.Format.Extent
- DC.Format.Min
- DC.Format.Max
- DC.Format.Orientation.

Jeg har valgt å imøtekomme brukernes ønske om et bilde for å beskrive genre til bildet. Dette er knyttet til beskrivelse av motivet, og er naturlig å innføre under beskrivelseselementet til DC. Anbefales innført for alle formatene.

DC.Type.Quality sier noe om kvaliteten på ressursen som beskrives. For filmformatet vedkommene vil ressursen si filmmappen, men det er kvaliteten på selve bildet som er av interesse. Det hadde da vært naturlig å innføre sub-elementet **Quality** til DC.Source, siden det er negativnummeret som her viser til hvert fotografimotiv. Siden bilder som også ligger i filmarkivet, også er scannet inn i Fotostation, og her er lagret som egne objekt vil selve bildet være ressursen. Sub-elementet Quality kan da legges under DC.Type.Quality. I illustrasjonsformatet gjelder det samme. Når det gjelder film-formatet kan en også benytte DC.Type.Quality ved at dette refereres til hvert bilde som beskrives for filmmappen. DC.Type.Quality vil derfor være et repeterende element for filmformatet, men ikke for de andre billedformatene.

DC.Format.Extent gjelder bildet sin utstrekning i bredde og høyde. For film der motivet stammer fra en filmstripe, vil alle negativ være like store, da de har en fast størrelse. Dette elementet er derfor ikke aktuell for film. For illustrasjoner, som ofte foreligger i papirform, vil størrelse på motivet ha noe å si for eventuell kopiering og scanning av bildet for gjenbruk. Elementet er aktuell for illustrasjonsformatet. For Fotostation blir alle bilder scannet inn i en fast størrelse. Allikevel kan en bruke format elementet til å si noe om manipuleringsmulighetene en har til bildet, f.eks min og max størrelse som bildet kan ha uten at kvaliteten forringes. Dette er aktuelt alt etter hvor stor plass en har til bildet i tilknytning til en artikkel, både for papirutgaven og

Internettutgaven av avisen. Format som her da kan benyttes er DC.Format.Min og DC.Format.Max, hvor A.M.V. har brukt DC.Format.full og DC.Format.small. DC.Format.Extent kan altså ligge som informasjon i Illustrasjonsformatet, og DC.Format.min og DC.Format.Max innføres i formatet for Fotostation. Dersom Fotostation på sikt erstattes av SIFT vil DC.Format.Min og DC.Format.Max være de som står igjen for å angi full str. og minste størrelse til bildet.

DC.Format.Orientation tilfredsstiller behovet for å angi om bildet er et høydebilde, eller breddebilde, (dvs vertikalt eller horisontalt). Dette er aktuelt å benytte for alle tre formatene.

Vedlegget til Intranett formatet får egne metadata lagt på av Internettredaksjone. Vedlegget er lagret som eget objekt, og hentes ut fra CCI. Slik det er nå overføres ikke metadata fra CCI til bildet, slik det er tenkt og metadata må derfor påføres på nytt i Internettredaksjonen. Bilder som ligger i Fotostation får påført metadata der, og videreføres derfra til CCI systemet. Hadde kommunikasjonen her fungert optimalt og metadata hadde blitt med, ville metadata påført i Fotostation, blitt med når bildevedlegg ble hentet ut fra CCI for å relateres til artikkelen. På bakgrunn av dette blir ikke de metadata som legges på vedlegget for utlegging i Internettutgaven tatt med i oppsummering av metadata for bilder.

For alle formatene for bilder, vil metadata for utgiver og rettigheter være aktuell å legge på ved publisering via Intranett. Alle informasjonsobjekt i Adresseavisens arkiv vil da ha utgiver "Adresseavisen", og rettigheter til informasjonsressursen vil enten være knyttet til opphavsmann eller Adresseavisen alt etter avtaler. Adresseavisen fungerer her på samme måte som et forlag, og forlag som har avtale med opphavsmann har også visse rettigheter til ressursen. DC har elementene DC.Rights og DC.Publisher for beskrivelse av rettigheter og utgiver.

8.5. Oppsummering av mapping

Av mappingen ser vi at de ulike formatene har flere like behov, men også medieavhengige variasjoner. Indeksering av filmmapper skiller seg ut fra de øvrige informasjonsobjekt ved indeksering, siden flere billedmotiv er beskrevet pr. mappe. Resultater av mappingen er oppsummert i tabell 8.11, (s187) og 8.12, (s188).

De metadata som tapes i mappingen og som utgår er metadataene:

- /**antall-linjer/** for formatet for stoff/artikler i SIFT
- **/antall/** for illustrasjonsformatet
- /**retur-adr/** for illustrasjon, film og Fotostation formatet

Ingen av de metadataene som her tapes benyttes ved søk av brukere.

Artiklene i SIFT og i Internettutgaven inneholder samme informasjon, men Internettformatet skiller ut flere metadata enn SIFT formatet. Det er heller ikke samsvar i navngivningen.

Dersom samlinger skal gjøres tilgjengelig for allmenheten over Intranett er det også behov for i tillegg å legge til noen metadata som viser hvem som har rettighetene til informasjonsressursene og hvem som er utgiver. Dette tilfredsstilles av DC.Rights og DC.Publisher. DC.Publisher vil ha verdien "Adresseavisen" for alle informasjonsobjekt som her er beskrevet.

Som vi har sett av mappingen for de ulike formater i dette kapittelet, er metadatene for bilder i Internettformatet veldig beskjedent. Det som er hensikten her er at metadata påført i Fotostation blir med i import til CCI. Når da Internettredaksjone ønsker å legge ut en artikkel med bilde i Internettutgaven, er hensikten at bildet og artikkel kan hentes ut fra CCI og at metadata skal følge med. Det skal være tilstrekkelig å foreta indeksering i ett ledd.

Fotostation er det verktøyet som fungerer best her for indeksering av bilder. Dette er fordi du har direkte tilgang til ressursen, og ressursen kan gjenbrukes direkte uten å "lete opp" ressursen. I tillegg følger metadata hvert enkelt bilde, og er lagret som eget objekt. Det hadde også vært mulig å indeksert hvert enkelt motiv i SIFT, men her er det uansett bare metadataposten som lagres.

Hvilke verktøy som her bør brukes ser jeg ikke på som min oppgave å belyse.

Det som skiller Internettformatet fra SIFT formatet for artikler er:

- Internettformatet lagrer metadata for notisen til hovedartikkelen som har stått på første side i Internettutgaven. SIFT lagrer denne som egen enhet/post, og er dermed å ses på som et eget dokument. Dette er grunnen til at Internettformatet har metadata for **/forsidetittel/**, og **/forsidetekst/**.

Det skulle ikke være noe i veien for at også SIFT innførte metadataene /**forsidetittel/** og **/forsidetekst/**. Dette er valgfritt.

Bruk av DC elementet for hvilket språk ressursen er tilgjengelig på er aktuell, dersom basen skal åpne for søk i et informasjonssystem som også har informasjonsressuser på andre språk. En ser

viktigheten av å informere om hvilket språk informasjonsressursen er tilgjengelig på når du f.eks søker i en base som har artikler på både engelsk, fransk, japansk osv. Alle kan ikke lese alle språk. Når ressursen mangler denne opplysningen kan det føre til at f.eks du bestiller en artikkel som kun er tilgjengelig på japansk. Denne misforståelsen oppstår ofte fordi originaltittel ofte gjengis på engelsk i internasjonale. Jeg ser ikke det store behovet for å legge språkinformasjon til ressursene på nåværende tidspunkt, siden alle tekstdokumentene som gjøres tilgjengelig vil være på norsk.

Til slutt vil jeg også anbefale å bruke DC.Coverage.Place for alle format, som da viser til geografisk sted som bildet dekker.

Dersom alle informasjonsobjekt skal gjøres tilgjengelig via Internett er det også nødvendig med bruk av DC.Publisher for å klargjøre hvem som er utgiver av bilder, artikler m.m.

Tabell 8.11. Forslag til metadataformat for artikler ved Adresseavisen

FELLES METADATA FOR BEGGE FORMATET	**MEDIA-AVHENGIGE VARIASJONER**	
	METADATA INTERNETT FORMATET	***METADATA SIFT FORMATET***
DC.Identifier DC.Date.Issued DC.Creator.Agentname DC.Creator.Agentrole DC.Creator.AgentAffiliation DC.Contributor.Agentrole DC.Contributor.Agentname DC.Contributor.AgentAffiliation DC.TitleDC.Relation.IsPartOf DC.Subject.Classification DC.Subject DC.Relation.Illustrasjon DC.Relation.Bilde DC.Relation.Film DC.Relation.Tegning DC.Description.Ingress DC.Description.Billedtekst DC.Description..Rights DC.Description.Publisher	DC.Title.Alternate DC.Title.Forsidetittel DC.Description.Forsidetekst status DC.Date.Sperredato	Ingen
tekst side merknad		

Tabell 8.12. Forslag til metadataformat for bilder ved Adresseavisen

FELLES METADATA	**MEDIA-AVHENGIGE VARIASJONER**		
DC.Identifier DC.Date.Issued DC.Creator.Agentname DC.Creator.Agentrole DC.Creator.AgentAffiliation DC.Contributor.Agentname DC.Contributor.Agetnrole DC.Contributor.AgentAffiliation DC.Subject.Classification DC.Subject DC.Title DC.Relation.Sak DC.Relation.IsPartOf DC.Description DC.Description.Genre DC.Rights DC.Coverage.Place DC.Publisher Side Merknad Retur DC.Format DC.Type.Quality	***FILM***	***ILLUSTRASJONER***	***FOTOSTATION***
	DC.Identifier.Jobb DC.Identifier.Film DC.Date.Created DC.Relation.Filmmappe DC.Source.Type DC.Source.Bildenummer DC.Format	DC.Identifier DC.Format DC.Format.Extent	DC.Identifier DC.Date.Created DC.Source.Type DC.Source.Bildenummer DC.Format.Min DC.Format.Max

Det er gjennom hele kapittelet gjort grundige drøftinger for mapping av hvert element. Derfor har jeg bare en sluttkommentar: Tabellene taler for seg.

Kapittel 9

Evaluering av hovedfagsavhandlingen

9.1. Evaluering av resultater

I denne oppgaven har jeg fremlagt forslag til metadatasett for indeksering av bilder, film og artikler ved Adresseavisen, med noen medieavhengige variasjoner. Dette forslaget har sin basis i Dublin Core. Jeg mener resultatet som er fremlagt vil fungere bra som et utgangspunkt for Adresseavisen. Ved implementering av dette forslaget vil det bidra til bedre samsvar mellom de metadataformater som Adresseavisen i dag bruker. I mitt forslag til format er det tatt hensyn til de variasjoner som oppstår, alt etter om artikler og bilder er lagret i SIFT, Fotostation eller i Internettdatabasen. Hensikten må allikevel være at et verktøy blir brukt til indeksering av artikler og bilder, slik at en unngår dobbeltlagring for informasjonsressursene. Alle bilder som det er aktuelt å bruke i avisen blir scannet inn i Fotostation og lagret i mapper der. Dersom alle bilder ble indeksert i Fotostation, er det ikke behov for i det hele tatt å indeksere filmmapper og illustrasjoner i SIFT som gjøres idag. [dette er illustrert i figur 6.8, s (93)]. Fordelen med Fotostation er at bildene er lagret digitalt som egne objekt, og at bildet kan gjenbrukes uten å lete frem kilden til bildet. Dette vil også lette arbeidet til Arkivansatte som da kan foreta indeksering en gang, istedet for både å indeksere bildet i Fotostation og filmmappen i SIFT.

Det er allikevel viktig å være oppmerksom på at forslaget ikke er noen fasit, og at løsningen er basert på min oppfattelse av brukskontekst og de brukerbehov som resultater fra undersøkelsen gjenspeiler. Det er mitt ønske at andre hovedfagstudenter vil videreføre dette arbeidet som her er gjort, og ta fatt på relaterte oppgaver som jeg har forslått. [se "Forslag til videre arbeid" i avsnitt 9.5, s (193)].

9.2. Kompleksiteten i avhandlingen

Detaljeringsgraden i denne opgaven er høy. Dette gjelder blant annet beskrivelse av arbeidsoppgaver ved Arkivet og hvordan de ulike metadaformater som benyttes brukes her.

Dette har jeg sett som nødvendige, da brukskonteksten rundt de ulike metadataformatene er viktig når en skal vurdere om dissse behovene kan tilfresstilles bedre med et annet format som utgangspunkt. Når jeg da skal legge frem forslag til løsning, er det viktig for leseren å vite detaljerringsgraden i indekseringsarbeidet, nettopp fordi bedrifter og organsiasjoner har sin egen praksis på dette.

Når det i tillegg er vesentlige ulikheter mellom de ulike formater fra prosess til prosess internt i bedriften, og de har flere navnekonflikter, har det til tider vært frusternede å skille de fra hverandre.

Nettopp på grunn av denne oppgavens kompleksitet har jeg sett det nødvendig å gjøre stor bruk av figurer og tabeller, for å gjøre fremstillingen så oversiktlig som overhodet mulig.

9.3. Feilvurderinger

I all forskning vil det ligge en del feilkilder. Som nevnt i innledningen av denne avhandlingen er det derimot viktig å være kritisk til eget datamateriell og vite den begrensing dataene har. Viktige ting kan ikke sies for ofte, og jeg må få fremheve igjen at det tynne grunnlaget for analyse av undersøkelsen som her foreligger. Når jeg opererer med så få som 15 brukere, er dataene veldig følsom for variasjoner. Denne følsomheten for endringer er årsaken til at jeg ikke kan konkludere med noen bastante konklusjoner fra undersøkelsen, og at resultatene må tas med "to klyper salt".

"Etterpåklokskap" oppstår ofte i ettertid av en undersøkelse er gjort, og resultater av dette fremlagt. Jeg ser flere områder når det gjelder brukerundersøkelsen hvor jeg kunne ha gjort ting annerledes.

Etter all arbeidet det var med å utforme spørreskjema, var det svært skuffende en lav svarprosent, når jeg i utgangspunktet hadde forventet minst 50%. Dette fikk jeg meg iallfall til å reflektere over ting som kunne vært gjort annerledes. Jeg ønsket i utgangspunket å se ulikheter mellom M-brukere og IM-brukere, og få avkreftet eller bekreftet min hypotese om at de fleste

søkte i fritekst. Selv om mye tyder på at de fleste søkte i fritekst fikk jeg dette delvis bekreftet, men hadde ikke nok grunnlag til å kjøre statistikk på dette. Undersøkelsen fikk derfor en mindre rolle i oppgaven enn jeg i utgangspunktet hadde tenkt. Til tross for dette, og på bakgrunn av den tid det tok å analysere det jeg fikk inn av spørreskjema, ser jeg at det ikke bare var ulemper med denne lave svarprosenten. På grunn av spørreskjemaets omfattelse vil en større svarprosent betydd en betydellig mer arbeidsmengde.

I ettertid ville jeg nok ha utformet skjemaet mer mot ett tema, istedet for å "gape" over alt på en gang. Det er mye interessant å undersøke, men en må vite å begrense seg. I og med at utformingen av spørreskjemaet ble gjort på en såpass tidlig stadium i avhandlingen min, er nok årsaken til at det ble så omfattende. I forhold til mappingsforslaget som jeg skulle legge frem til slutt ser jeg at fokus på M-brukerne burde vært målgruppen for undersøkelsen. Brukerbehov hos M-brukerne om hvilke metadata de benyttet ved søk ville jeg ha dratt direkte nytte av i mappingskapittelet. I mappingen er det tatt hensyn til de 5 brukerbehov til M-brukerne som her var representert. I tillegg ønsket jeg å finne ut av hvordfor IM-brukerne ikke brukte metadata ved søk. Dette ville kreve såpass mye innsyn i arbeidsoppgaver og informasjonsbehov at det ville blitt en avhandling i seg selv. Dette vil også gått litt på siden av min hovedmålsetnig for denne avhandlingen, nemlig:

> Hvordan de metadataformater Adresseavsien behov kunne tilfredsstilles i DC som kjerneformat med tanke på publisering av sine informasjonsressurser.

Når allikevel undersøkelsen ble som den ble har jeg valgt å oppsummere alle de resultater den ga, for deretter å fokusere på brukerbehovene til M-brukere som jeg kunne dra nytte av i mitt forslag til metadataformat. De øvrige resultater kan sette i gang en tankeprosess som jeg mener kan være nyttig å bruke som grunnlag for videre brukerstudier innen dette området. De tendenser jeg her har påvist kan da danne hypoteser for videre brukerstudier.

Etter at jeg hadde foretatt undersøkelsen vurderte jeg i tillegg om jeg skulle følge opp med dybdeintervju av neon av brukerne, eventuelt foreta en loggføring av brukernes søk mot SIFT databasen. Begge delene antok jeg for tidkrevende, og ikke mulig å gjennomføre innen for den tidsrammen som jeg hadde satt meg for min hovedfagsavhandling.

Det beste rådet jeg kan gi til andre studenter her er at du tenker nøye igjennom hva du vil at undersøkelsen du foretar deg skal gi svar på, og hva du kan bruke svarene til. Undersøkelser krever mye arbeid, og det er ikke alltid at en får den respons og det resultat en ønsker.

Jeg har erfart at det ikke er den enkleste sak å få brukere til å delta i undersøkelsen, og at disse ikke ser den samme viktigheten av dette som du selv legger i det. That is the HARD facts.

9.4. Litteraturreferanser

Når det gjelder resultater fra undersøkelsen har jeg ikke drøftet dette opp mot resultater fra andre undersøkelser. Årsaken til dette er at jeg ikke har funnet andre undersøkelser som jeg kunne relatere til min problemstilling. Resultater fra to andre prosjekter er brukt i denne avhandlingen, som er prosjektet til Anne Marie Vercoustre [37] og en brukerundersøkelse av Margaret E. Graham [17] som begge har fokus på indeksering av bilder.

Jeg har også henvist til arbeidsutkast fra DCMI, som er grunnlaget for skriving av kapittel 9. Disse arbeidsutkast er henvist til ved URL'er, og kan derfor bli fjernet etterhvert som DCMI sitt arbeid oppdateres.

9.5. Forslag til videre arbeid

I dette prosjektet har jeg fremlagt forslag til hvordan Adresseavisen sitt behov kan tilfredsstilles i DC, men ikke foretatt noen implementering av metadata for de faktiske informasjonsobjekt. Jeg har også nevnt at XML er en måte å implementere DC metadata på. Bruk av XML er veldig aktuelt for tiden, og det er også en trend i avismiljøet for å ta i bruk denne teknologien.

Forslag til videre arbeid kan settes opp punktvis:

1. Implementering metadataforslaget for artikler og bilder i avis som her foreslås i XML. Jeg ser for meg at en utvikler en prototype for å se hvordan dette forslaget kan fungere i praksis.
2. Ved Institutt for datateknikk og informasjonsvitenskap (IDI), innen retningen informasjonsforvaltning er DIGLIB digital biblitek utarbeidet som pr. dato består av en billed-database, IDI sin base for hovedfagsavhandlinger og en billeddatabase som består av flyfotografier , prospekter og manuskriptkart. Det hadde vært ønskelig å få til et samarbeid med Adresseavisen for integrering av deres arkiv i DIGLIB prosjektet for tilgang for IDI sine studenter. Dette er en oppgave som er aktuell for hovedfagsavhandlinger ved informasjonsforvaltning ved Institutt for datateknikk og informasjonsvitenskap ved NTNU.
3. Med utgangspunkt i det kjerneformatet for metadata som her foreslås kan en sammenligne med andre aviser sine behov for å fremlegge et metadataformat som kan fungere for indeksering av informasjonsressurser i et avismedium. Dagbladet og Stavanger Aftenblad bruker SIFT systemet og samme format for indeksering av artikler i SIFT. En oppgave her er også å legge frem Document Type Definitions i XML for det forslag som en kommer frem til her. Som nevnt er XML en teknologi som avismiljøet er åpen for, og som flere har tatt i bruk.

4. I brukerundersøkelsen i denne avhandlingen ble det fokusert på i hvilken grad metadata ble brukt i informasjonssøkingsprosesessen og hvilke metadataelement som ble brukt ved søk etter informasjon. Det er nødvendig med flere undersøkelser av søkeatferd for å finne ut mer om hvilke metadata som er riktige å bruke tilpasset ulike informasjonsbehov og arbeidsoppgaver. Med den riktige kunnskap om metadata her kan en knytte metadata til ulike arbeidsoppgaver og informasjonsbehov og gi den enkelte bruker sitt eget "view" eller perspektiv presentert i informasjonsrommet i et Intranett.

Litteraturliste

1. Bakken, Steinar. Kunnskapsorganisering ved hjelp av Edb. Kap.9: Innføring av MARC formatet i databasen MARCBASE, s87-94. Kap 10: ISBD, AACR2, MARC, s95-97. 1994. *ISBN:* 82-90790-06-6.
2. Belkin, Nicholas J & Croft, Bruce W. : ***Retrieval Techniques. Annual review of Information Science and Technology*** (ARIST), vol. 22, s110-145.1987.
3. BIBLINK Prosjektet. Http://hosted.ukoln.ac.uk/biblink/
4. BIBSYS Digitalt bibliotek. Husby, Ole. ***Mapping av Dublin Core til BIBSYS-MARC.*** 19.10.1998. Http://www.bibsys.no/meta/d2m/norXwalk.htm.
5. BIBSYS Digitalt bibliotek.Husby, Ole: ***Mapping av BIBSYS-MARC til Dublin Core***. 9.10.1998. Http://www.bibsys.no/meta/d2m/marc2dc.html.
6. Buchland, Micheal: ***What is a digital document***. Http://sims.berkeley.edu/~buckland/digdoc.html.
7. Carven, Tim: ***Thesaurus construction. Faculty of information and media studies***. Univ. of western Ontario. Http://instruct.uwo.ca/gplis/677/thesaur/main00.htm
8. Chapmann, Ann/Day,Micheal/Hiom, Debra: ***Metadata: Cataloguing practice and Internet subject-based information gateways.*** Ariadne. Volum 18. Desember 1998. ISSN:1361-3200.
9. Chepesiuk, Ron. ***Organizing the Internet: The 'Core' of the Challenge. American Libraries***, Januar 1999. Volum 30. Nr. 1. s60-66. ISSN 00029769. "Academic Search Elite" database.
10. ***Discovering online resources across the humanities : a practical implementation of the Dublin Core*** / [edited by Paul Miller and Daniel Greenstein on behalf of the Arts and Humanities Data Service (AHDS) and the UK Office for Library and Information Networking (UKOLN). Trykt: Bath : UKOLN, 1997. Sidetall: 95 s. : fig. ISBN: 0-9516856-4-3 Eiere: NBO UBIT)
11. Dublin Core Element Set, Version 1.1: Reference Description.Dublin Core Metadata Initiative http.//purl.org/dc/documents/rec-des-19990702.htm
12. Fabritius, Hannele.: ***Information Seeking Behaviour of Journalists***. The Second International Congress on Electronic Media and Citizenship in the Information Society. Http://www.uta.fi/~lihafa/jan99sem.html
13. Frakes, William B. & Baexa-Yates, Ricardo: ***Information Retrieval. Data Structures & Algorithms***. *1992.* Prentice Hall.
14. Fotostation 3.5 Userquide. Http://www.fotoware.com/documentation/pdf/fs35.pdf.

15. FotoWare Norge AS sin hjemmeside. http:www.fotoware.com.
16. Gradmann, Stefan: ***Metadata: Old wine in new bottles.*** 64th IFLA General Conference August 16 - August 21, 1998. Http://www.ifla.org/IV/ifla64/007-126e.htm
17. Graham, Margareth E. The Description and indexing of images. Report of a survey of ARLIS members 1998/99. Institute for Image DATA research. University of Northmbria at Newcastle. D-lib Magazine. Volum 5 nr 10. Oktober 1999. ISSN 1082-9873
18. Hills, Philip J.: ***Information management systems. Implication of the human - computer interface***.
19. Gundersen, Roy/Brandshaug, Rune: BIBSYS- 25 years of library automation and co-operation in Norway. Trondheim. 9s.
20. Husby, Ole.: ***Metadata***. Foredrag ved kunnskapsorganisasjonsdagene 1997. HIO. Http://www.bibsys.no/meta/korg97.html
21. Husby, Ole. ***Metadata***. Forelesning ved RBTs Videreutdanningsprogram i IT for ansatte i fag og forskningsbibliotek. November 1999.
22. Husby, Ole: ***Dublin Core Metadata Element Set:*** Norsk referansedokument. Http://www.bibsys.no/meta/dc/dcref.html
23. Johnson, Robert. ***Elementary statistics. (s.39).*** 1996. 7.utgave.
24. Lagoze, Carl. ***The Warwick Framework. A Container Architecture for Diverse Sets of Metadata.*****D-Lib Magazine**, July/August 1996. ISSN 1082-9873. Http://www.dlib.org/dlib/july96/lagoze/07lagoze.html#WarwickFull.
25. Levy, David M & Marshall, Catherine C.: ***Going Digital: A look at Assumptions underlying digital librarires.*** Communication of the ACM. April 1995. Vol 38. No4. s77-84.
26. Marchionini, Gary: ***Information seeking in the electronic age***. Kap 3. S27-60. Cambridge University Press, c1995
27. Megill, Kenneth: ***The corporate memory. Information management in the electronic age***. Bowker-Saur publishing. 1997.Kap
28. ***Nordic Metadata Project.*** Http://linnea.helsinki.fi/meta/
29. Ohren, Oddrund/Natvig,Marit/Brevik, Ole: SINTEF Rapport. ***Avansert Interanett samarbeid (AIS)***. Des. 1999. Sintef Tele og data.
30. Pollock, Annabel/Hockley, Andrew: ***Whats wrong with Internet searching.*** D-lib Magazine. Mars 1997.
31. Seltzer, Richard/Ray, Eric J./Ray, Deborahs. ***The Altavista search revolution. How to find anything on the Internet.*** *Kap 5. s84-105.* Osborne-McGraw Hill. 1997.
32. Sheth, Amit/Wolfgang, Klas: ***Multimedia data management. Using metadata to integrate and apply digital media.*** McGraw-Hill. 1998.
33. Statens Datasentral a.s. ***SIFT Introduksjon til SIFT Plus***.

34. Tranøy, Knut Erik.:***Vitenskapen - samfunnsmakt og livsform***. 1996. Oslo. Universitetsforlaget.
35. Uschold, M. and King, M: ***Towards a methodology for building Ontologies, Workshop on Basic Ontological Issues in Knowledge Sharing***. International joint Conference on Artificial Intelligence. 1995
36. Universitetsbiblioteket i Trondheim. ***Hva er BIBSYS***. Http://www.ub.ntnu.no/veiledning/bibsys/l1a.htm.
37. Vercoustre, Anne-Marie /Paradis, Franqois: ***Metadata for photographs: From Digital Library to multimedia application***. Lecture notes in Computer Science. Research and advanced technology for digital libraries. Third European Conference, ECDL 99, Paris, sept. 1999.
38. Wale, Thorbjørn. ***Innføring i journalistikk***. 5 utgave. 1997. ISBN8271471643
39. Weibel, Stuart: ***The State of the Dublin Core Metadata Initiative***. April 1999. D-Lib Magazine. Volum 5. Nr. 4. 16s. Http://www.dlib.org/dlib/april99/04weibel.html
40. Weibel m. fl. 1998: ***Dublin Core Metadata for resource discovery.*** ftp://ftp.isi.edu/in-notes/rfc2413.txt
41. Aalberg, Trond: ***Integrert gjenfinning i heterogene dokumentbaser***. Hovedfagsoppgave i informasjonsforvaltning. Institutt for datateknikk og informasjonsvitenskap ved NTNU. 1998.
42. The Dublin Core metadata initiativ. Http://purl.org/dc

Arbeidsutkast fra DCMI sine arbeidsgrupper:

43. DC Title Working Group. ***Proposal for Title Qualifier.*** Http://purl.org/DC/groups/qualifierproposal-title.htm. 2000-02-10.
44. ***Dublin Core Qualifiers: Title and Identifier.*** DC Title Working Group. Augusut 1999. Http://purl.org/DC/groups/title-qualifierreview.html
45. ***Dublin Core Sub-elements. Summary of work prior to and during DC, and an outline of issues still to be resolved.*** Oktober 1997.
46. ***The 7th Dublin Core Metadata Workshop*** .October 25-27, 1999. 7. DC Konferanse. Die Deutsche Bibliothek Frankfurt am Main, Germany. Http://www.ddb.de/partner/dc7conference/results.htm
47. **Coverage Working Group.** Http://www.mailbase.ac.uk/lists/dc-coverage/files/wd-coveragequal.htm

48. ***DC Agents Working Group - Qualifier Proposal Date***. 1999-11-12 . Http://www.mailbase.ac.uk/lists/dc-agents/files/qualifier-final.txt
49. DC Relation/Source Working Group - ***Review of Relation Qualifier Usage***. 1999-08-04. Http://purl.org/dc/groups/relation-qualifierreview.htm
50. ***DC Subdesc Working Group. Proposed subject, Description and language qualifiers***. 30Des. 1999. Http://www.mailbase.ac.uk/lists/dc-subdesc/files/dcsubdesc-final.html
51. ***DC Type Qualifiers***. DCMI Working Draft. 10 Desember 1999. Http://www.loc.gov/marc/dc/typequalif-19991210.html
52. Date working group.***DC Date Qualifiers***. DC (Final) Working Draft - 14 December 1999**.** Http://www.mailbase.ac.uk/lists/dc-date/files/prop-19991214.html
53. Title working group. ***Proposal for Title Qualifier.*** 2000-02-10. Http://purl.org/DC/groups/qualifierproposal-title.htm
54. DC Coverage Working Group. ***Review of Coverage Qualifier Usage*** .1999-08-02. Http://www.mailbase.ac.uk/lists/dc-coverage/files/qualifiers.html
55. Network working group. ***Dublin Core Metadata for Resource Discovery*** .Arbeidsnotat. September 1998. ftp://ftp.isi.edu/in-notes/rfc2413.txt
56. Subelement working group. ***Subelement Working Draft.*** 2.11.1998. Http://purl.org/DC/documents/wd-subelements-current.htm
57. DC-Workshops. The 7th Dublin Core Metadata workshop. ***Results on the workinggroups.*** 1998-02-11. Http://www.ddb.de/partner/dc7conference/results.htm
58. ***Approval of initial Dublin Core Interoperability Qualifiers***. DCMI. 17 April 2000. Http://www.mailbase.ac.uk/lists/dc-general/2000-04/0010.html
59. ***CEN Workshop Agreement - Metadata for multimedia information***. 27 Mars 2000. Http://www.cenorm.be/news/press_notices/metadata.htm

Søkemotorer, verktøy, prosjekt (URL referanser):

60. Ultraseek. Http://ultraseek.com/
61. Swish-E. Http://sunsite.berkeley.edu/SWISH-E/
62. Microsoft's Index Server. Http://www.microsoft.com/
63. Autonomy Knowledge Server . Http://www.autonomy.com/knowledge/ksintro.htm
64. Blue Angel Technologies MetaStar. Http://www.blueangeltech.com/
65. Verity Search 97 Information Server. Http://www.verity.com/products/infoserv/index.html
66. Altavista. Web-søkemotor. Http://www.altavista.com/
67. Altavista sine hjelpesider. Http://www.altavista.com/
68. Northern Light. Web-søkemotor Http://www.northernlight.com/
69. FAST Search. Http://www.fast.no/
70. UK Office for Library and Information Networking. (UKOLN). Http://www.ukoln.ac.uk/
71. DC Dot. UK Office for Library and Information Networking (UKOLN). Http://www.ukoln.ac.uk/cgi-bin/dcdot.pl
72. Dublin Core Metadata Template. Nordic metadata project. Http://www.lub.lu.se/cgi-bin/nmdc.pl
73. ***SOSIG: Social Science Information Gateway***. Http://www.ukoln.ac.uk/services/elib/projects/sosig/

A .1. Presentasjon av Dublin Core Sub-element

DCMI har enda ikke utarbeidet forslag til formaliserte sub-element, fordi det fremdeles er store uenigheter omkring dette i de arbeidsgrupper som forsker kontinuerlig på dette. I tabellen nedenfor har jeg tatt utgangspunkt i de ferskeste arbeidsnotarer fra DCMI web-side, og hva de har anbefalt for bruk av sub-element på det tidspunkt denne oppgaven skrives. Der hvor ikke andre arbeidsutkast refereres til har jeg tatt utgangspunkt i siste forslag fra DCMI på sub-elementer datert 17 april [58]. Det er allikevel viktig å være klar over at disse sub-element som du finner nedenfor ikke er identisk med de ulike prosjekter som benytter seg av DC som utgangspunkt, da flere prosjekter har eksperimentert med sub-element ut i fra de ulike behov. Disse prosjektenes ulike behov har også vært gjenstand for drøfting og utarbeidelse av forslag til sub-element. Oversikten i denne tabellen er sett på som forslag for nåværende tidspunkt og ikke er noe som er vedtatt av DCMI, og er heller ikke å se på noe endelig forslag og kan dermed bli forkastet ved nærmere diskusjon senere. Eventuelle endringer under dette punkt vil hele tiden finnes på Http://purl.org/DC under de ulike arbeidsgruppers virkeområde. Etter tabellen følger også en videre beskrivelse av hvert element og videre status rundt arbeidet med sub-elementer. Disse sub-elementene som presenteres her er også gjenstand for drøfting under mappingen i kapittel 8. For de som er interessert å se nærmere på arbeidsutkastene fra de ulike arbeidsgruppene har jeg i tabellen også referert til disse. Alle de sub-element som her nevnes kan brukes siden sub-element skal tjene for utvidelse etter behov og hver og en kan definere sine egne sub-element eller legge til noen ekstra

.

Tabell A 1. Oversikt over de DC Sub-elementer som foreligger pr. dato fra DCMI

Element	**Sub-element**	**Element beskrivelse**
DC.Title [53]	DC.Title.Alternative	En alternativ tittel til hovedtittelen for informasjonsressursen.
DC.Creator DC.Publisher DC.Contributor [48]	DC.Creator.AgentType	Indikerer rollen til navnet Agent
	DC.Creator.AgentName	Det formaliserte/vanlige navnet som Agenten her innehar
	DC.AgentAffiliation	Her er det organisasjonen som Agenten er assosiert med
	DC.AgentJurisdiction	Navnet på agentens innflytelsesområde/rettsområde
	DC.AgentDateRange	Datoer for Agentens levetid, fra født til død.
	DC.AgentRole	Indikerer hvilken rolle Agenten innehar
	DC.AgentLink	En referanse til metadata som beskriver den navngitte Agent.
DC. Subject [50]	Classification	Her defineres hvilket klassifikasjonssystem emneord er tildelt etter. f.es UDK, MESH emneord osv.
DC.Description [50]	Abstract	Et sammendrag av ressursen.
	TableOfContent	En innholdsfortegnelse som viser innholdet i ressursen.
	Note	Annen tilleggsinformasjon om innholdet i ressursen
	Release	En identifikasjon av utgaven, en release eller versjon av ressursen
DC.Date [52]	DC.Date.Created	Den dato informasjonsressursen er ferdig produsert av opphavspersonen, f.eks den dato du avslutter artikkelen/boken og anser den som ferdig.
	DC.Date.Modified	Den dato informasjonsressursen blir endret på
	DC.Date.Issued	Den dato informasjonsressursen blir formelt utgitt
	DC.Date.Available	Datoangivelse for når informasjonsressuren er tilgjengelig. Dette angis ofte med tidsintervall, f.eks at et dokument er tilgjengelig i en periode og deretter blir sperret for innsyn.
	DC.Date.Valid	Datoangivelse som sier noe som informasjonsressursen gyldighet i forhold til tid. Som oftest er dette et tidsintervall/tidsepoke.
DC.Type [51]	DC.Type.DCT1	Et kontrollert vokabular som indikerer ressursens natur eller genre.
DC.Format	Extent	Her er det tenkt å benyttes IMT kodingsskjema som står for "Internet Media Type". Extent betyr utstrekning, altså størrelse. Medium betyr type medium det referes til.
	Medium	
DC.Identifier	Ingen sub-element	

Tabell A 1. Oversikt over de DC Sub-elementer som foreligger pr. dato fra DCMI

Element	Sub-element	Element beskrivelse
DC.Source	Ingen sub-element	Drøftes under Relation working group som stiller spørsmålstegn ved om dette egentlig er en variasjon av RELATION elementet
DC.Language	Ingen sub-element	Kodingsskjema som her foreslås er IS0639-2, og RFC1766.
DC.Relation	Is version of	Er versjon av
	Has version	Har versjon
	Is replaced by	Er erstattet med
	Is required by	Er krevet av
	Requires	krever
	Is required by	Er krevd av
	Is part of	Er del av
	Is referenced by	Er referert av
	References	Referanser
	Is format of	Er format av
	Has format	Har format
DC.Coverage	Place	Dekningsområde knyttet til sted
	Time	Dekningsområde knyttet til tid
DC.Rights	Ingen subelement	

A 1.1. DC.Tittel:

I følge siste rapport her [43] har arbeidsgruppen pr. dato blitt enig om å foreslå DC.Title.Alternative, som er ment å dekke både forkortelser for tittel og oversettelse av tittel til andre språk. Når en bruker feltet som oversettelse til andre språk må en i tillegg benytte attributtet LANG for å definere språk.

På den 7 DC konferansen (DC-7) i Frankfurt i oktober 1999 oppsummerte arbeidsgruppe for tittel-elementet ulike alternativ som er brukt av de ulike prosjekter tilknyttet DCMI. Disse er: DC.Title.Main, DC.Title.Subtitle, DC.Title.Alternative/DC.Title.Alternate, DC.Title.Abbrivated/ DC.Title.Short, DC.Title.Abbreviation, og DC.Title.Release. Når det gjelder bruk av Release elementet har det vært diskutert om dette eventuelt skal gå under tittel med DC.Title.Release=2. edition, eller om det skal dekkes under description elementet som DC.Description.Release og DC.Description.Note. På DC-7 ble det også enighet om å anbefale bruk av DC.Title for hovedtit-

tel, og for titler som var blitt endret eller omskrevet ble det anbefalt at DC.Title skulle bli repetert.

A 1.2. DC.Creator/DC.Publisher/DC.Contributor:

Fra siste arbeidsnotat fra DC Agent Working Group [48] er som spesifisert i tabellen ovenfor. Elementene beskrives videre her

:

Tabell A 2. DC Sub-element (DC-qualifiers)

Navn	**Anbefalt bruk**	**Attributt verdier**
Agent Type	Verdier for attributtet hentes fra en DCMI vedlikeholdt liste for termer som kan brukes.	Fra DCMI liste for øyeblikket er disse verdier definert: "Person" - et menneske, "Corporate" - en organisasjon, "Conference" - en hendelse/møte, "Instrument"- et verktøy, objekt eller tjeneste.
Agent Name	Dette vil være det navnet som brukes for å referere til Agenten	*F.eks "Mary Jane Smith", "Acme Corprotation", "7th Dublin Core Workshop"*
Agent Affiliation	Agenten som er nevnt kan videre bli assosierte med den organisasjonen vedkommende er knyttet til, dvs hvilken organisasjon er "Agent Name" tilknyttet.	Eksempler på verdier her er "Acme Corporation". Mary Smith er **Agent Name** av rapporten, men hennes arbeid er gjort som ansatt hos "Acme Corporation" som blir spesifisert i sub-elementet **Agent Affiliation**
Agent Jurisdiction	Her defineres det hvilken innflytelse eller rettsdømme du som person innehar. Her anbefales det å velge verdier fra et kontrollert vokabular liste over mulige verdier.	Eks. på attributtverdier her kan være "Sør - Australia", "United Kingdom", "Trondheim" osv.
Agent Date Range	For mennesker vil dato intervall her være når vedkommende ble født og når vedkommende eventuelt døde. For organisasjoner vil det være den dato Organisasjonen ble opprettet og når den eventuelt ble avviklet.	Eksempel på bruk her vil være bruk av tidsintervall, f.eks født-død angitt som 12.3.1972-1.6.2000. Det er agentens levetid en er ute etter her, med start og sluttdato. Attributtverdier kan være hvilken som helst gyldig dato.

Tabell A 2. DC Sub-element (DC-qualifiers)

Navn	Anbefalt bruk	Attributt verdier
Agent Role	Den rollen som du i ditt navn innehar når du skaper ressursen. Her anbefales det også at det velges fra en liste av roller som du kan inneha. Listen over ferdigdefinerte roller som anbefales å bruker her er: MARC RELTOR skjemaet.Andre vokabular som brukes her bør være klart definert og identifisert ved unike skjemanavn -	Verdier for "MARC RELATOR "skjemaet finnes her (http://www.loc.govUmarc/relators/re9802r1.html). Denne MARC RELATOR skjemaet, og andre lignende vokabular kan inneholde verdier som ikke gir mening når de er brukt i sammenheng med en eller flere av AGENT elementene Creator, Publisher og Contributor. I MARC RELATOR Skjemaet skal ikke verdien "Publisher" bli brukt sammen med Creator og Contributor elementet. MARC RELATOR skjemaet er ikke anbefalt for bruk for Publisher elementet.
Agent Link	Her er det anbefalt at sub-elementet brukes slik at det refererer til et tall eller tekst som tilhører et formalisert identifikasjonsystem	Verdier her kan være en Uniform Resource (URI), hvor URI peker til en post eller navnet autoritetsfil som beskriver Agenten mer detaljert.

A 1.3. DC.Date

Alle feltene nevnt i tabellen over er sub-elementer foreslått til drøfting for den 7. DC Workshop som fant sted i Oktober 99 Resultatene av denne workshopen foreligger ikke enda i form av noen rapport, men sammendrag fra resultatene etter workshopene som ligger på weben [46] viser at sub-elementene created, issued, modified, available, valid ble godkjent som anbefaling. I tillegg ble from og to diskutert som mulige sub-elementer for å definere tidsintervall, men ikke oppnådd enighet om. Hensikten med sub-elementene under Dato er å vise de ulike livssyklusene til dokumentet, fra det blir skapt, utgitt, gjort tilgjengelig osv. Fra DC-7 er det problemområdet rundt ivaretaelse av sammensatte datoer som 1971-1999, geologiske perioder m.m [57]. De sub-element som er foreslått for dato er tiltenkt delt inn i tre grupper:

1. Skapelse av verket som inkluderer Created, Datagathered og Valid
2. Utgivelse av verket, som inkluderer Issued og Available
3. Overførelse av verket, som inkluderte Accepted og Acquired.

Etter at disse grupperinger av sub-element ble foreslått har elementene Datagathered, Accepted og Adquired blitt forkastet, så hele gruppe 3 faller vekk. Siden en nytt element, modified er kommet til, er min mening at det siste elementet kunne settes inn under gruppe "Endring av verket", slik at vi får denne grupperingen:

1. Skapelse av verket som inkluderer Created og Valid.

2. Utgivelse av verket, som inkluderer Issued og Available.
3. Endring av verket, som inkluderer Modified.

A 1.4. DC.Coverage

I tidligere arbeidsutkast var elementene placename, box, point, periodname, date, datestart, dateEnd.

foreslått som egne sub-element av DC.Coverage som vist i tabell . Nå har en blitt enig om to sub-elementnavn, som er **place** og **time**, og tilfredsstilt dette ved hjelp av forslag til 6 kodingsskjema, hvor 4 er til **place** elementet og 2 er til **time** elementet. For **place** er det kodingskjemaene DCMI point, ISO3166 for stedsnavn, DCMI Box og TGN (The Getty Thesaurus of Geogrphic Names). For sub-elementet **time** er DCMI Period og W3C-DTF foreslått.

Tabell A 3. Sub-element for DC.Coverage

DC.Coverage.Placename	Navnet på en dekningsområde for sted, f.eks billedsamling fra "Fransisco Bay"
DC.Coverage.Box	Den geografiske utstrekning, deklarert ut i fra et koordinatsystem
DC.Coverage.Point	En enkelt punkt i rommet, definert ut i fra et koordinatsystem
DC.Coverage.PeriodName	Navnet på en tidsperiode som informasjonsressursen går under, f.eks "Romertiden"
DC.Coverage.Date	En numerisk dato, deklarert ut i fra et definert datosystem
DC.Coverage.DateStart	En numerisk dato, definert ut i fra et definert dato system, og som uttrykker starten på denne datoperioden
DC.Coverage.DateEnd	En numerisk dato, definert ut i fra et definert dato system, og som uttrykker slutten på denne tidsperioden

Alle de som forslag til sub-element som er gjort under dette elementer er basert på analyse av de ulike sub-element som allerede er i bruk for dette elementet av ulike prosjekter [54]. Når det gjelder bruk av elementet Place som er foreslått anbefales det her at det brukes en liste med kontrollerte verdier for stedsnavn, som ISO 3166. Kodingsskjema Box og Point brukes disse dersom dekningsområde trenges å defineres ut i fra et koordinatsystem, f.eks max størrelse på et bilde i x og y koordinater. DCMI har utarbeidet forslag til hvordan disse skal brukes f.eks ble det tidligere forslaget omkring **Periodname** forslått brukt "DC.coverage.periodName" scheme = "historic" content = "Ming Dynasty". Bruk av feltet **time** som skal tilfredsstille tidligere element **Date**, **DateStart** og **DateEnd** anbefales brukt ved at det referer til standardiserte datofor-

mat som ISO8602 eller følger World Wide Web Consortium (W3C) sine retningslinger for hvordan uttrykke Dato og Tid. Siden **place** og **time** nylig er blitt foreslått for å erstatte de tidligere sub-element er det enda ikke kommet noe utkast med eksempel til hvordan **place** og **time** konkret skal brukes.

A 1.5. DC.Relation

Fra aller siste arbeidsutkast er det blitt enighet om 12 ulike sub-element til relationelementet [58], som vist i tabellen. Det er ikke utarbeidet noen eksempel på hvordan disse skal brukes i praksis, men flere prosjekter [49] tilknyttet DCMI bruker flere av disse elementene som hver bruker dem etter sitt behov. Kodingsskjema som her er foreslått er URI.

A 1.6. DC.Subject / DC.Description

I aller siste utkast er det elementene Abstract og tableOfContents som sub-element til DC.Description [58]. Tidligere her har også DC.Description.Note og DC.Description.Release vært foreslått i tidligere utkast [50], og her blir også sub-element Classification foreslått for DC.Subject, og disse har jeg valgt å ta med i tabell 1.

A 1.7. DC.Type. DCT1

For dette sub-elementet er det et kontrollert vokabular som forslås, og hvor disse termer kan inngå i vokabularet: interaktiv ressurs, datasett, hendelse, bilde, lyd, tjeneste, software, samling, tekst.

A 1.8. DC.Format

Det er ønskelig at de to sub-elementene Extent og Medium skal godtas som anbefaling, men diskusjonen om dette er enda ikke ferdig ifølge Andy Powell. Det kodingsskjema som her er tenkt brukt er IMT som står for "Internet media type", og viser til nettopp ferdigdefinerte mediatyper for Internett som en kan velge mellom. Eksempler for hvordan Extent og Medium skal brukes er ikke publisert på DCMI sine sider enda, og det foreligger heller ingen forklarelse til elementene.

Tabell B 1. Metadataelement M-brukere mener er uegnet for søk

Bruker	Illustrasjoner	Film	Artikler/stoff
1	/antall/ill-id/ill-type/	/jobb-id/film-nr/film-id/retur/bilde-nr/	/art-nr/ant-linjer/original/
12	/antall/ill-id/ill-type/	/jobb-id/film-nr/film-id/film-dato/	/art-nr/ant-linjer/original/merknad/
13	/antall/ill-id/opphav/ merknad/side/	/jobb-id/film-nr/film-id/ merknad/retur/bilde-nr/ side/	/art-nr/side/ant-linjer/original/merknad/

Tabell B 2. Metadataelement IM-brukere mener er uegnet for søk

Bruker	Illustrasjoner	Film	Artikler/stoff
3	/merknad/retur/	/merknad/retur/bilde-nr/	/art-nr/ant-linjer/original/stikkord/
8	/antall/ill-id/gruppe/ produkt/		/art-nr/produkt/side/ant-linjer/kilde/original/ merknad/
9			/art-nr/prod-dato/produkt/side/ant-linjer/illustr./ kilde/original/merknad/
11	/antall/ill-type/	/jobb-id/film-nr/film-id/ merknad/retur/bilde-nr/ produkt/side/	/art-nr/prod-dato/side/ant-linjer/illustr./original/ merkand/
14			/art-nr/produkt/side/ant-linjer/
15	alle element	alle element	/art-nr/prod-dato/

I

Tabell B 3. IM- brukernes opplæring i SIFT i forhold til bruk av metadata

IM-Brukere	Opplæring	Fornøyd med opplæringen
2	D	JA
3	D	JA
5	B	Vet ikke
7	B	JA
8	A	JA
9	D	NEI
10	B	JA
11	D	NEI
14	B	NEI
15	B	JA

Tabell B 4. M- brukernes opplæring i SIFT i forhold til bruk av metadata

M- brukere	Opplæring	Fornøyd med opplæringen
1	A	JA
4	B	NEI
6	A	Vet ikke
12	A	NEI
13	E	JA

Tabell B 5. M-brukeres opplevelse av søkesituasjoner

Bruker	Få eller ingen treff	For mange treff	For mange treff som ikke er relevante
1	Av og til	Av og til	Av og til
4	Av og til	Aldri	Aldri
6	Ofte	Sjelden	Sjelden
12	Vet ikke	Av og til	Ofte
13	Av og til	Sjelden	Aldri

Tabell B 6. IM-brukeres benyttelse av andre kilder enn SIFT

Brukere	Gir arkivet jobben (A) (fotnote 1)	Databaser på nettet (B)	Spør kollegaer ekstern og internt (C)	Oppslagsverk på huset (D)	Andre kilder (E)
1[a]		OFTE	AV OG TIL	SJELDEN	ALDRI
4	OFTE	OFTE	OFTE	OFTE	ALDRI
6	SJELDEN	OFTE	AV OG TIL	OFTE	OFTE
12 [a]		SJELDEN	AV OG TIL	OFTE	SJELDEN
13	ALDRI	OFTE	ALDRI	SJELDEN	ALDRI

a. Arkivansatte - svarer ikke på alternativ A.

Tabell B 7. Hvilke metadata M-brukere bruker ved søk etter Illustrasjoner

Brukere	Brukshyppighet	Metadatafelter
Bruker 1 (Arkivet)	Alltid	/sak/
	Ofte	/ill-id/prod-dato/opphav/brukt-dato/
	Av og til	/side/
	Sjelden	/illustr/ gruppe/merknad/
	Aldri	/antall/ill-type/retur/produkt/
Bruker 4 (Kultur)	Av og til	/prod-dato/sak/brukt-dato/produkt/
	Sjelden	/illustr/side/
	Aldri	/antall/ill-id/ill-type/opphav/gruppe/retur/ retur-adr/merknad/
Bruker 12 (Arkivet)	Alltid	/sak/brukt-dato/
	Ofte	/prod-dato/opphav/merknad/side/
	Av og til	/illustr/gruppe/
	Sjelden	/retur/
	Aldri	/antall/ill-id/ill-type/

Tabell B 8. Hvilke metadata M-brukerne bruker ved søk etter etter film.

M-bruker	Brukshyppighet	Metadataelement
Bruker 1	Ofte	/film-id/fotograf/gruppe/sak/brukt-dato/side/motiv/
	Av og til	/prod-dato/
	Sjelden	/merknad/retur/produkt/
	Aldri	/jobb-id/film-nr/film-dato/bilde-nr/
Bruker 12	Alltid	/sak/brukt-dato/motiv/
	Ofte	/prod-dato/fotograf/produkt/
	Av og til	/gruppe/merknad/side/
	Sjelden	/jobb-id/film-nr/film-id/film-dato/retur/
	Aldri	/bilde-nr/

Tabell B 9. Hvile metadata IM-brukerne bruker ved søke etter stoff/artikler

Bruker	Avdeling	Metadatelementer
Bruker 1	Alltid	/emneord/tekst/
	Ofte	/prod-dato/produkt/journalist/merknad/
	Av og til	/side/illustr/kilde/
	Sjelden	/ant-linjer /
	Aldri	/art-nr/original/stikkord/
Bruker 4	Alltid	/tekst/
	Ofte	/journalist/Emneord/
	Av og til	/side/illustr/kilde/
	Sjelden	/ant-linjer/
	Aldri	/art-nr/original/stikkord/
Bruker 6	Av og til	/prod-dato/journalist/emneord/merknad/
	Aldri	de øvrige elementene
Bruker 12	Alltid	/emneord/merknad/tekst/
	Ofte	/prod-dato/produkt/side/illustr./journalist/
	Av og til	/kilde/
	Aldri	de øvrige elementene

Tabell B 9. Hvile metadata IM-brukerne bruker ved søke etter stoff/artikler

Bruker	Avdeling	Metadatelementer
Bruker 13	Ofte	/emneord/merknad/tekst/
	Av og til	/prod-dato/produkt/side/illustr./journalist/
	Sjelden	/original/stikkord/
	Aldri	de øvrige elementene

Tabell B 10. Metadata som M-brukere synes er vanskelig å forstå etter brukere

Bruker	Illustrasjoner	Film	Artikler/Stoff
1	/ill-type/	/jobb-id/ film-id/	/kilde/original/
6			/art-nr/produkt/side/ant-linjer/illustr./kilde/original/stikkord/ tekst/
12	/ill-id/	/jobb-id/ film-id/	/original/
13	/opphav/	/jobb-id/ film-id/ gruppe/	/produkt/stikkord/

Tabell B 11. Metadata som IM-brukere synes er vanskelig å forstå etter brukere

Bruker	Illustrasjoner	Film	Artikler
8	/ill-id/illustr./gruppe/	/gruppe/	/produkt/
9	/antall/opphav/gruppe/ merknad/retur/	/jobb-id/film-nr/film-id/film-dato/prod-dato/gruppe/ merknad/retur/bilde-nr/side/	/art-nr/prod-dato/produkt/side/ ant-linjer/illustr./kilde/original/ merknad/
11	/ill-id/ill-type/gruppe/	/film-id/film-dato/	/original/
14	/antall/ill-id/illustr./ill-type/ produkt/	/jobb-id/gruppe/produkt/	
15	alle element	alle element	/art-nr/prod-dato/

Tabell B 12. Metadataelement M-brukere mener er uegnet for søk

Bruker	Illustrasjoner	Film	Artikler/stoff
1	/antall/ill-id/ill-type/	/jobb-id/film-nr/film-id/retur/bilde-nr/	/art-nr/ant-linjer/original/
12	/antall/ill-id/ill-type/	/jobb-id/film-nr/film-id/film-dato/	/art-nr/ant-linjer/original/merknad/
13	/antall/ill-id/opphav/ merknad/side/	/jobb-id/film-nr/film-id/ merknad/retur/bilde-nr/ side/	/art-nr/side/ant-linjer/original/merknad/

Tabell B 13. Metadataelement IM-brukere mener er uegnet for søk

Bruker	Illustrasjoner	Film	Artikler/stoff
3	/merknad/retur/	/merknad/retur/bilde-nr/	/art-nr/ant-linjer/original/stikkord/
8	/antall/ill-id/gruppe/ produkt/		/art-nr/produkt/side/ant-linjer/kilde/original/ merknad/
9			/art-nr/prod-dato/produkt/side/ant-linjer/illustr./ kilde/original/merknad/
11	/antall/ill-type/	/jobb-id/film-nr/film-id/ merknad/retur/bilde-nr/ produkt/side/	/art-nr/prod-dato/side/ant-linjer/illustr./original/ merkand/
14			/art-nr/produkt/side/ant-linjer/
15	alle element	alle element	/art-nr/prod-dato/

Tabell B 14. IM brukeres benyttelse av andre kilder enn SIFT

Brukere	**Gir arkivet jobben (A)**	**Databaser på nettet (B)**	**Spør kollegaer ekstern og internt (C)**	**Oppslagsverk på huset (D)**	**Andre kilder (E)**
2	ALLTID	OFTE	AV OG TIL	SJELDEN	ALDRI
3	AV OG TIL	VET IKKE	AV OG TIL	OFTE	AV OG TIL
5	AV OG TIL	-	-	-	-
7	ALDRI	OFTE	OFTE	OFTE	SJELDEN
8	AV OG TIL	AV OG TIL	AV OG TIL	SJELDEN	OFTE
9	AV OG TIL	ALDRI	AV OG TIL	AV OG TIL	ALDRI
10	ALDRI	ALLTID	ALLTID	ALDRI	ALDRI
11	SJELDEN	OFTE	AV OG TIL	SJELDEN	-
14	ALDRI	OFTE	OFTE	OFTE	AV OG TIL
15	SJELDEN	OFTE	-	OFTE	AV OG TIL

Tabell B 15. Arbeidsoppgaver i forhold til brukere og avdeling

Arbeidsoppgaver	Bruker nr.	Avdeling
Indeksere stoff i SIFT	1, 12	Arkivet
Indeksere artikler/stoff til utlegging på Internett	11	Internetredaksjonen
Søke etter informasjon og bilder i SIFT og Fotostation	1, 12	Arkivet
Redigering	2, 4	Desken
	11	Internett
	13	Kultur
Layout	2	Desken
Kvalitetssikring	2	Desken
Generelt Journalistisk arbeid, produsere stoff til avisen	6, 10, 13	Kultur
	7, 9, 14, 15	Nyhet/sekretariatet
	11	Internettredaksjonen
Journalistisk arbeid innen forskning og utdanning	5	Nyhet/sekreatriatet
Kulturreportasjer, Uke-Adressa	3	Kultur
Reportasjeledelse	8	Nyhet/sekreatriatet
Utvalg av bilder	4	Desken

Tabell B 16. Informasjonsbehov i SIFT etter bruker og avdeling

Informasjonsbehov	Bruker nr.	Avdeling
Fakta/Bakgrunnsopplysninger om saker	1, 12	Arkivet
	3, 6, 10, 13	Kultur
	5, 7, 8, 9, 14, 15	Nyhet/Sekretariatet
	2	Desken
	11	Internettredaksjonen
Sjekke skrivemåter, kilder, årstall, datoer, navn etc.	1, 12	Arkivet
	6	Kultur
	5, 9, 14, 15	Nyhet/Sekretariatet
	2	Desken
	11	Internettredaksjonen
Finne frem bilder fra SIFT og FOTOSTATION	1, 12	Arkivet

Tabell B 17. Gi-opp situasjoner i informasjonsinnhenting etter bruker, avdeling og opplæring

Gi opp situasjon	Bruker nr.	Avdeling	Type Opplæring
Ikke vet hvordan begrense søket	11	Internett	B
Mangel på tid	4		B
	9, 15		9=D, 15=B
Vet at informasjonen ikke finnes i SIFT	1, 12	Arkivet	A
	10	Kultur	B
Tungvint i bruk	8, 5	Nyhet/Sekretariatet	5= B, 8=A
Det som finnes er utilstrekkelig	7	Nyhet/Sekretariatet	B
	13	Kultur	E
For mange treff	6	Kultur	A
	11	Internettredaksjonen	D
Ustabilt system	5	Nyhet/Sekretariatet	B
Antar at informasjonen har falt ut av SIFT, (dvs at det antas at det ikke er lagt inn eller er blitt slettet.)	3	Desken	D
	6	Nyhet/Sekretariatet	A
Antar seg selv som lite flink til å søke i SIFT	15	Nyhet/Sekretariatet	B
Systemet gir ikke hjelp ved feil syntaks, vanskelig å treffe	5	Nyhet/Sekretariatet	B

Tabell B 18. M-brukeres brukshyppighet av SIFT

Bruker	Antall år brukt SIFT	Antall dagers bruk pr. uke	Dagshyppighet
1	1993 - >	4	kontinuerlige hele dagen
6	1993 ->	5	< 3 ganger pr .dag
13	1993 ->	4	> 3 ganger pr .dag
4	3 - 5 år		
12	< 1 år	4	kontinuerlige hele dagen

Tabell B 19. IM-brukeres brukshyppighet av SIFT

Bruker	Antall år brukt SIFT	Antall dagers bruk pr. uke	Dagshyppighet
2	1993->	4	>3 ganger pr .dag
3	1993->	5	< 3 ganger pr .dag
10	1993->	annet	< 3 ganger pr .dag
8	1993->	5	< 3 ganger pr .dag
10	1993->	annet	< 3 ganger pr .dag
9	3 - 5 år	5	< 3 ganger pr .dag

Tabell B 19. IM-brukeres brukshyppighet av SIFT

Bruker	Antall år brukt SIFT	Antall dagers bruk pr. uke	Dagshyppighet
7	3 - 5 år	5	kontinuerlig
14	3 - 5 år	5	> 10 ganger pr. dag
15	1-3 år	5	> 10 ganger pr. dag
11	< 1 år	5	< 3 ganger pr .dag
5	< 1 år	5	> 3 ganger pr .dag

Anita Iren Oppedal
Brinken 7
7043 TRONDHEIM
E-MAIL: iren@ifi.ntnu.no
Web : http://www.ifi.ntnu.no/~iren
Tlf: 73 51 00 55

04.02.99

Adresseavisen ASA
V/ ansvarlig redaktør Gunnar Flikke
Postboks 6070
7003 TRONDHEIM

SØKNAD OM SPØRREUNDERSØKELSE PÅ ADRESSEAVISEN

Jeg viser til telefonsamtale med Arne Blix angående en spørreundersøkelse på Adresseavisen i forbindelse med min hovedfagsoppgave ved institutt for datateknikk og informasjonsvitenskap ved NTNU.

Tema for oppgaven min er informasjonsgjenfinning uavhengig av media (bilder, lyd, tekst osv) og fysisk lagringssted. Spesielt ønsker jeg å se på forskjeller i fritekstsøk og søk ved bruk av metadata. Med metadata mener jeg her faste beskrivelsesfelt som tittel, journalist, fotograf, produksjonsdato etc.

I hovedfagsoppgaven min tenker jeg meg et "informasjonsrom" som inneholder informasjon av ulike media fra Universitetsbibliotekene i Norge, Internett og en Avis, (som i dette tilfelle er Adresseavisen). Et slikt informasjonsrom har som mål å tilfredsstille avisens totale informasjonsbehov. Problemer med ulike medier og ulike datakilder(arkiv, databaser o.l.) er at de er lagret i ulike format, har ulike søkerutiner slik at det er vanskelig med et felles søkevindu mot disse. Resultatet blir at brukerene ikke finner ønsket informasjon eller at de må søke i ulike datakilder på ulike måter før de finner det de leter etter.

Jeg setter hovedfokus på :

"Hvordan beskrive informasjon på en så god måte at den lar seg gjenfinne "

Bakgrunnen for mitt valg av oppgave og case er mitt arbeid ved arkivet i Adresseavisen sommerene 97 og 98 samt deltid samme sted 97 og 98. I tillegg jobber jeg ved Universitetsbiblioteket ved NTNU, 2 dager i uken.

For å finne en mulig løsning på integrasjon av de tre nevnte datakilder (avis, Internett og Universitetsbiblioteket) er det viktig å innhente opplysninger om den erfaring brukere som til daglig jobber med informasjonsbehandling har,

i og med at det er brukerenes behov som ønskes tilfredsstilt.

Resultatet av undersøkelsen i Adresseavisen vil bli sammenholdt med relevante forskningsprosjekt eller andre undersøkelser foretatt i sammenlignbare bedrifter. Målet med dette vil være å se på ulikheter og likheter og sammen med resultatet fra flere undersøkelser gi et bedre konklusjonsgrunnlag for eventuelle løsninger.

Målgruppen er sporadiske og hyppige brukere av SIFT basen. Brukere som **aldri** har brukt SIFT er ikke interessante i denne undersøkelsen. Brukergruppene er delt inn i 8 grupper som vist i spørsmål 6 i undersøkelsen. For å få kunne trekke konklusjoner ut av besvarelsene er det nødvendig med besvarelser fra 5-6 personer p.r. gruppering, totalt 35-40 spørreskjema.

Forutsatt at De godkjenner gjennomføring av spørreundersøkelsen, må de ansatte bli informert om dette. Hvis det er ønskelig deltar jeg gjerne i forberedelse og gjennomføring av et informasjonsmøte. Det er ønskelig at så mange som mulig er villig til å delta.

Jeg opplyser om at alle som deltar i undersøkelsen garanteres full anonymitet. Min faglige veileder, professor Ingeborg Sølvberg og jeg er underlagt taushetsplikt. Data vil bli behandlet konfidensielt i henhold til retningslinjer fra Datatilsynet.

Jeg ønsker å gi deltakerne så god tid som mulig til undersøkelsen og har satt siste frist for innlevering til fredag 12 mars.

Til sommeren/høsten kan det bli aktuelt å følge opp undersøkelsen med noen få dybdeintervju. Hvis dette blir aktuelt vil jeg ta kontakt på nytt for valg av relevante intervjuobjekt.

Dersom det skulle være uklarheter i forbindelse med dette vennligst ta kontakt med undertegnede eller min veileder ved:

E-post : Ingeborg Sølvberg ingeborg@cs.ucsb.edu, ingeborg.solvberg@ifi.ntnu.no

Vedlagt følger undersøkelsen i sin helhet, samt problembeskrivelsen for oppgaven min.

Håper på positiv og snarlig tilbakemelding.

På forhånd takk.

Vennlig hilsen

Anita Iren Oppedal

2 Vedlegg