Indeksering af Internetressourcer Er »Metadata« løsningen?

Af Bo Gerner Nielsen og Erik Thorlund Jepsen

Indledning

Den eksplosive vækst i både antallet af dokumenter tilgængeliggjort via Internettet og benyttelsen af Internettet har i stigende omfang medført en fokusering på, hvorledes information tilgængeliggjort via Internettet registreres og formidles til nettets brugere.

Registrering af Internetressourcer besværliggøres bl.a. af Internettets dynamiske karakter, hvor dokumenter ændrer indhold og udseende samt i nogle tilfælde skifter navn eller adresse. I ligeså høj grad besværliggøres registreringen dog af, manglen på ekspliciteret genfindingsrelevant information i dokumenterne (*Hvem er ophav til dokumentet? Hvad er titlen?, Hvornår er dokumentet publiceret? m.m.*).

"Dublin Metadata Core Element Set" er et forsøg på igennem ophavsgenererede indekseringer, at etablere et bedre datagrundlag for både automatisk og manuel registrering og formidling af Internetressourcer. Ideen er, at ophavet til en given ressource hæfter oplysninger indenfor en række forhåndsbestemte områder (som titel, forfatter, emne, udgiver m.m.) til ressourcen. Disse i ressourcen iboende oplysninger kan benyttes ved indekseringen hos de forskellige institutioner, der udarbejder søgeværktøjer (søgemaskiner, emnelister og kataloger) i forhold til Internetressourcer.

Såfremt Dublin Core vinder indpas hos både ophavene til Internetressourcer og udbyderne af søgeværktøjer, vil forbedringen af datagrundlaget for indekseringen kunne betyde, at brugerne af søgeværktøjerne præsenteres for fyldigere og mere ensartede beskrivelser af dokumenterne. Denne artikel problematiserer dog, hvorvidt Dublin Core vil medføre de store indekseringsmæssige lettelser eller genfindingsmæssige forbedringer, såfremt man ikke udarbejder retningslinjer for syntaks og semantisk kontrol til hjælp for ophavenes indeksering af egne dokumenter, da indeksør eller søger i så fald selv skal medtænke alle mulige former af navne, titler, emneord m.m. for at sikre et rimeligt søgeresultat.

Artiklen indledes med en karakteristik af ressourcer på Internettet samt en typologisering og beskrivelse af de forskellige søgeværktøjer til Internetressourcer. En kort omtale af problemerne ved MARC-katalogisering af Internetressourcer efterfølges af en forklaring på "Metadata"-begrebet. Udfra en gennemgang af formålet med og indholdet i "Dublin Metadata Core Element Set" diskuteres, hvilke implikationer Dublin Core kan få for indekseringen til og søgningen i henholdsvis søge-

maskiner, strukturerede emneindekser og bibliotekskataloger.

Karakteristik af Internetressourcer

Internetressourcer adskiller sig på flere områder fra de dokumenttyper, vi traditionelt arbejder med at registrere.

Manglende stabilitet
Internetressourcer er elektroniske dokumenter, som i mange tilfælde er foranderlige. Indholdet af - og strukturen i - de enkelt dokumenter kan forandres, ligesom dokumenterne kan skifte adresse (URL) eller forsvinde. Et hyppigt problem består i, at forældede udgaver ikke fjernes fra en server, hvilket mindsker sandsynligheden for at identificere en ny udgave.

Sammensatte dokumenter
Samme Internetressource indeholder ofte flere forskellige datatyper: tekst, billeder, videoklip, animationer og lyd, samtidig med at ressourcerne kan bestå af flere materialeformer: præsentationer, artikler, vokabularer, artikler, links til andre ressourcer m.m.

Komplekse dokumenter
Mange Internetressourcer er konstrueret som et web (væv) af relaterede dokumenter, hvor de enkelte dokumenter kan have selvstændige adresser (URL's). Ofte er der en indgangs(start)side, der viser den overordnede struktur af ressourcen (se f.eks. Danmarks Biblioteksskoles hjemmeside: http://www.db.dk/), men en startside opremser sjældent alle dokumenter med selvstændige adresser (ofte kun den overordnede struktur af ressourcen).

Manglende bibliografisk information i dokumentet
Der er stor forskel på, hvor megen information de enkelte ressourcer formidler om forhold, vi i biblioteksverdenen normalt *betragter* som væsentlige for formidlingen af dokumenter. Desværre foreligger der ikke nogen nyere og grundige undersøgelser af disse forhold, men en OCLC-stikprøveundersøgelse af 100 dokumenter foretaget i 1993 (Assessing information on the Internet, 1993) viste at:

96 % af dokumenterne havde nogen bibliografisk information i starten af dokumentet
30 % af dokumenterne havde nogen bibliografisk information i slutningen af dokumentet
90 % af dokumenterne medtog titeloplysninger
73 % af dokumenterne medtog ophavsoplysninger
64 % af dokumenterne medtog oplysninger om publiceringstidspunkt

Den fysiske beskrivelse af Internetressourcer besværliggøres dels af web-konstruktionen, som medfører problemer angående hvilke dele af ressourcen, der skal beskrives, og dels af dokumenternes sammensatte struktur, hvor det ofte er elementer som billeder og grafik, der optager megen plads. Det er altså svært at lave en fyldestgørende fysisk beskrivelse samtidig med, at man må forvente, at den fysiske beskrivelse har betydning for brugernes relevansbedømmelse af dokumentet, specielt set i lyset af sammenhængen mellem dokumentets omfang (indhold af multimedie-elementer) og transmissionshastigheden.

Definitionsproblemer i forhold til bibliografiske elementers funktioner
Ligesom vi kan komme ud for at mangle informationer, kan vi i formidlingsmæssig sammenhæng få problemer med at forholde os til de informationer, der er indeholdt i Internetdokumenter. Sådanne problemer kan for eksempel opstå i forhold til ophavsfunktioner, hvor ophav som designer og web-master kan være angivet i stedet for eller på lige fod med de, vi normalt betragter som, intellektuelt og kunstnerisk ansvarlige for dokumenterne. En anden problematik vedrører aktualiteten af dokumentet. Hvor vi for andre materialetypers vedkommende ofte får oplysninger om udgave og version, gør dette sig sjældent gældende for dokumenter på Internettet, hvilket primært medfører problemer, når ældre versioner ikke er fjernet fra serveren. Oprettelses- og opdateringstidspunkt for Internetressourcen kan - hvis oplyst - anvendes som alternativ til udgaveinformation.

Genfindingsmuligheder i forhold til Internetressourcer

Der eksisterer for nuværende to væsensforskellige typer af søgemuligheder: *søgemaskiner* og *strukturerede emneindekser*

Søgemaskiner
Søgemaskiner er databaser, der afsøger nettet for URL'er, indekserer den tilknyttede ressource og lægger indekseringen over i databasen. Flertallet af disse søgemaskiner gør brug af automatisk indeksering og udtrækker termer fra bestemte dele af teksten (titel, undertitel, første 5 linier...). I enkelte søgemaskiner udføres dog manuel indeksering. De fleste søgemaskiner anvender en eller flere former for partiel match, men flere tillader også mulighed for kombinatorisk søgning.

Søgemaskinernes absolutte force er databasernes omfang, som kun muliggøres qua den automatiske indeksering.

Den største ulempe hos søgemaskinerne skyldes den forholdsvis tilfældige beskrivelse af dokumenterne, som er en følge af den automatiske udtrækning af termer. Disse dokumentbeskrivelser gør det svært for brugerne at identificere de fleste dokumenters relevans i forhold til brugerens informations- eller oplevelsesbehov. En anden væsentlig ulempe er de manglende muligheder for at definere, hvilke datatyper man søger på - at lave "feltspecifik søgning". Eksempelvis vil en søgning på termen "copyright" medføre et utal af hits, som emnemæssigt intet har med copyright-forhold at gøre, men udelukkende er genfundet da der i dokumenterne optræder oplysning om dokumentets copyright-forhold.

Strukturerede emneindekser
Strukturerede emneindekser er lister over udvalgte Internetressourcer. Disse lister er ofte hierarkisk opbyggede udfra kriterier som: Emne, form, sprog, nationalitet m.m. Eftersom listerne baseres på intellektuel udvælgelse og indplacering af ressourcer, er omfanget af indekserede ressourcer betydeligt mindre end i søgemaskinernes databaser. Strukturerede emneindekser benytter sig af hypertekstfunktionalitet, hvorfor de er nemme for slutbrugere at orientere sig i.

Til enkelte af emneindekserne udarbejdes egentlige beskrivelser af de enkelte dokumenter, men de fleste større emneindekser har ingen dokumentbeskrivelse udover titlen (ofte endda en konstrueret titel) på dokumentet. Flere emneindekser opererer dog - som søgemaskinerne - med automatisk udtrukne beskrivelser.

De strukturerede emneindekser lider altså under samme manglende mulighed for relevansbedømmelse som søgemaskinerne. Emneindekser indeholdende manuelt udarbejdede beskrivelser af dokumenterne vil være meget ressourcekrævende at vedligeholde, såfremt en tilfredsstillende dækning skal opnås.

Katalogisering af Internetressourcer

En stigende frustration over ulemperne ved søgemaskinerne og de strukturerede emnelister fik allerede i starten af 90'erne flere til at foreslå MARC-formatet som standard til registrering af Internetressourcer. Argumenterne for at benytte MARC-formatet var primært:

- MARC-formatet er det mest udbredte og afprøvede metadata-format
- MARC-formatet tillader beskrivelse på flere niveauer
- MARC-katalogisering integrerer via katalogen formidlingen af Internetressourcer med formidlingen af øvrige materialetyper i biblioteket
- Brugen af felt "856" muliggør direkte opkobling til ressourcen fra den bibliografiske post

(frit efter Sha, 1995, s.468-70)

MARC-katalogisering har dog vist sig at være problematisk grundet forskellige forhold:

- Kompleksiteten i MARC-formaterne betyder, at registreringen er meget ressourcekrævende.

Katalogiseringsprocessen er tidskrævende og stiller store krav til katalogisator.

- Som tidligere nævnt besværliggøres katalogiseringen af problemer med at identificere oplysninger vedrørende dokumentet samt problemer med at definere feltindhold. Problemerne med at definere feltindhold bør dog løses ved udarbejdelsen af adækvate regler og retningslinjer for katalogiseringen af Internetressourcer.
- Enkelte opfatter det som et problem, at katalogiseringen af Internetressourcer betyder et brud med den hidtidige opfattelse af katalogens funktion som en registrant over en samling. Den katalogiserende institution har sjældent samme kontrol over de katalogiserede Internetressourcer som over andre materialer, der indgår i katalogen. Dette er først og fremmest et problem, der opstår på grund af de tidligere omtalte stabilitetsproblemer ved Internetressourcerne. I tilfælde, hvor den katalogiserende institution abonnerer på - eller på anden måde har indgået aftaler om adgang til en Internetressource - kan man med lidt velvilje sagtens betragte Internetressourcerne på lige fod med bibliotekets øvrige materialer.

Hvad er metadata?

Selve metadatabegrebet betyder i sin simpleste form bare "data om data". Dette spænder over alle former for data, der benyttes til identifikation, beskrivelse og lokalisering af enhver form for dokumenter. Begrebet er blevet betragtet som en af hjørnestenene i beskrivelsen og tilgængeliggørelsen af informationer på Internettet.

Ganske logisk har der ikke været de store slagsmål om hvorvidt metadata er en god idé i forhold til beskrivelsen af Internettets mangeartede dokumenter. Alle har kunnet blive enige om at en eller anden form for beskrivelse er bedre end slet ingen beskrivelse. Det, der i høj grad har været diskuteret på diverse konferencer og workshops (f.eks. Weibel, 1995; Dempsey, 1996; Heery, 1997:II) er, hvilken form og struktur disse metadata skulle have.

Metadatakonceptet, som det bruges i både traditionelle og elektroniske biblioteker, refererer typisk til information, der

- giver forholdsvis kort karakteristik af individuelle dokumenter i bibliotekets samling,
- er lagret som en del af en bibliotekskatalog
- og er ment som hjælp for brugerne til genfinding af dokumenter af interesse (Smith, 1996, s.2).

Normalt ville man i biblioteksverdenen nok kalde metadata for katalogisering (Woodward 1996, s.207), men da der er mange andre grupper end bibliotekarer involveret i arbejdet med metadata, vil de associationer der knytter sig til katalogisering virke forvirrende.

Der eksisterer flere forskellige typer af forslag til metadata formater, hvor det mest kendte i bibliotekskredse er MARC-formatet. Derudover, eksisterer der dog mange andre formater, både generelle / universelle og mere fagspecifikke som TEI, der er tilknyttet humaniora og lingvistik og FGDC, der benyttes i geografisk / kartografiske miljøer (Hakala, 1997).

Et format der har mulighed for universel udbredelse er Dublin Core formatet.

Dublin Core

Det oprindelige navn var egentlig Dublin Metadata Core Element Set, men det bliver for nemheds skyld nu udelukkende refereret til som *Dublin Core* eller endnu kortere DC.

Dublin Core blev fremsat som det *minimum* af metadataelementer (core element set), der er nødvendige til beskrivelsen af dokumenter af enhver slags i et netværksmiljø som Internettet.

Formålet med Dublin Core er at skabe rammen for *ophavsgenererede beskrivelser* af Internetressourcer.
Det har derfor i første omgang primært været nødvendigt at skabe en fælles forståelse for og kort-

lægning af behov, styrker, svagheder og løsninger ift. de nuværende problemer.

Formålet er dels at skabe konsensus om et basalt sæt af metadataelementer til beskrivelse af netværks-/Internetressourcer, dels at skabe en teknisk simpelt implementerbar løsning i forhold til inddragelse af beskrivelseselementerne.

Dublin Core er primært udviklet gennem en række af workshops og konferencer. Mange af referaterne, arbejdspapirerne og resultaterne fra disse ligger naturligt nok tilgængelige på Internettet, hvor der også er flere diskussionsgrupper.

Overordnet vedligeholdes bestræbelserne på udviklingen af Dublin Core løst af OCLC (http://purl.oclc.org/metadata/dublin_core.html), der i det hele taget arbejder meget med forbedringen af adgangen til Internettets dokumenter. Bl.a. også PURL (Persistent Uniform Ressource Location), der kort fortalt baserer sig på idéen om, at et Internetdokument får tildelt én adresse, der vedblivende skal sikre adgang til dokumentet, selvom dette reelt skifter adresse.

For at opfylde kravet, om at indekseringen skal kunne foretages af ophavene selv, skal de enkelte metadataelementer være så entydige og så let intellektuelt forståelige som muligt.

Dublin Core elementerne og deres grundlæggende funktion.
Som baggrund for udvælgelsen af de elementer der skal indgå i Dublin Core, er der på de ovennævnte workshops og konferencer (Weibel, 1995; Dempsey, 1996; Weibel, 1996; Heery, 1997:II), blevet opstillet en række krav til den grundlæggende funktion af et Dublin Core element.

For det første skal antallet af metadataelementer være så lille som muligt, men skal samtidig betragtes som en basis der skal kunne udvides efter behov. Først og fremmest, fordi vi her arbejder med ophavsgenererede beskrivelser, må formatet gøres så overskueligt som muligt af hensyn til de uuddannede "indeksører", som skal foretage den intellektuelle beskrivelse og vurdering af deres egne dokumenter. Sekundært skal formatet give mulighed for at der rent faktisk kan laves indekseringer af dokumenterne på et ret højt niveau.

For det andet skal det tilstræbes, at metadataelementerne skal kunne beskrive dokumenter og dokumenttyper indenfor så mange emneområder som muligt. Dette falder meget naturligt, da naturen af Internettet er meget mere mangfoldigt både på forskelligheden af dokumenttyperne end almindeligvis set i biblioteker og på dokumenternes potentielle kontinuerlige udvikling.

For det tredie skal metadataelementerne beskrive det indre indhold i dokumenterne, både intellektuelt og formmæssigt. Herunder er syntaks-uafhængighed især blevet påpeget som værende vigtigt, idet formelle definitioner, (som f.eks. bestemmelser om kun at ophavene skulle beskrive emnet i navneform eller bestemmelser om brug af entals- og flertalsform), kan skabe problemer i forhold til princippet om intellektuel simplicitet i brugen af metadata. Derudover skal det være valgfrit om man udfylder alle punkterne på listen over metadataelementer, da nogle elementer måske ikke vil være relevante i forhold til alle typer af dokumenter. Samtidig skal det være mulige at gentage alle metadataelementer da der f.eks. kan være flere forfattere eller nødvendigt med flere emneord. Endelig skal metadataelementsættet indeholde en mulighed for tilpasning til forskellige miljøer gennem tildeling af kvalifikatorer til de enkelte elementer.

Generelt er kvalifikatorer ment som en hjælp til at specificere det semantiske indhold i et bestemt element (Hakala, 1997). Kvalifikatorerne er en udvidelse af Dublin Core og er ment som en hjælp til mere sofistikerede indekseringer, primært med det formål at man får nemmere ved at foretage en relevansvurdering af dokumenterne.

Kvalifikatorspørgsmålet har været en problematisk diskussionspunkt, idet nogle mente at det stred mod den oprindelige idé om simplicitet i Dublin

Core, at lave den slags udvidelser. Modargumentet har været at man gennem brugen af kvalifikatorer kunne specificere det semantiske indhold i et metadataelement. F.eks. kan man specificere, hvilken kontrolleret emneordsliste man har brugt til sine emneord, eller om man har brugt et ISBN eller ISSN nummer. Eller man kan specificere en forfatterangivelse ved at vedhæfte e-mailadresse, telefonnummer etc. (Hakala, 1997).

Det fjerde og sidste meget vigtige punkt er, at der udelukkende skal medtages elementer der kan bidrage til genfindingen af dokumenterne.

Metadataelementerne.
Antallet af metadataelementer, der tildeles de enkelte dokumenter vil, som det fremgår af ovenstående, variere en hel del. Alligevel er det jo nødvendigt at fastlægge, hvilke metadataelementer der egentligt skal kunne *tildeles* i beskrivelsen via Dublin Core.

Ser man lidt historisk på udviklingen af Dublin Core er antallet af elementer steget fra tretten elementer til de nuværende femten. I den oprindelige udgave var der ikke medtaget "beskrivelse" og "rettigheder".

1. Titel	Navnet på objektet. Hvis objektet ikke har noget egentligt navn kan man benytte en beskrivende tekststreng.
2. Forfatter/ophav	Den person eller organisation der er primært ansvarlig for det intellektuelle indhold.
3. Emne	Det videns- eller emneområde objektet tilhører.
4. Beskrivelse	En tekstuel beskrivelse i form af en indholds- eller abstractbeskrivelse.
5. Udgiver	Den forlægger eller det forlag der er ansvarlig for udgivelsen, offentliggørelsen eller tilgængeliggørelsen af objektet.
6. Andre ophav end forfattere eller "creators"	Personer eller organisationer der har haft betydelig indflydelse på objektet, såsom redaktører eller oversættere.
7. Dato	De datoer der er forbundet med objektet.
8. Ressourcetype	Genretypen af objektet (fx. roman, ordbog, digt).
9. Format	Angivelse af det fysiske format objektet forefindes i, såsom HTML eller Postscript.
10. Ressource identifier	En entydig beskrivelse af dokumentet (f.eks. i form af en http-adresse).
11. Kilde	Angivelse af en evt. trykt eller elektronisk kilde, hvorpå objektet bygger.
12. Sprog	Angivelse af hvilket sprog det intellektuelle indhold af objektet er på.
13. Relationer	Angivelse af hvilke relationer objektet har til andre objekter eller samlinger.
14. Dækning	Et specialfelt, der fortrinsvist knytter sig til at beskrive en geografisk eller tidsmæssig begrænsning af et bestemt dokument.
15. Rettigheder	Ophavsrettigheder tilknyttet objektet.

(frit efter Weibel, 1995 og Lagoze, 1996)

Teknisk implementering

Den praktiske succes afhænger i høj grad af hvordan softwareudviklerne af HTML-editorerne modtager Dublin Core, da det vil være en stor hjælp for den enkelte indeksør (webmaster, webdesigner, osv.), hvis muligheden for tildeling af metadataelementerne er indbyggede i editoren.

Derudover er man i søgesituationen afhængig af at indekseringsrobotterne fra søgemaskinerne tager hensyn til metadata som en mulighed, der hvor den findes.

I sidste ende er det også et spørgsmål, der får betydning for hvorvidt de folk der stiller dokumenterne til rådighed vil bruge tid på at indeksere dokumenterne. Incitamentet til at benytte Dublin Core vil være at det medfører større benyttelse af ressourcen.

Alle felter der opbygges efter Dublin Core modellen får tildelt "fornavnet" DC. Herefter følger selve beskrivelsen. På den måde vil en indekseringsrobot fra en søgemnaskine kunne identificere at f.eks. DC.author eller DC.title er metadataelementer, som skal registreres på en bestemt måde i databasen.

Dette giver så efterfølgende databaseværterne en mulighed for at stille feltspecifikke søgemuligheder til rådighed for brugerne af baserne.

Til brug for indeksørerne af dokumenterne arbejdes der med udviklingen af en række hjælpemidler. De vigtigste er en fastlæggelse af brug af kvalifikatorer og en indkorporering af Dublin Core elementerne i HTML-kodningen. Dette arbejde er på nuværende tidspunkt ikke fuldt afsluttet, da der er divergerende synspunkter på bl.a. selve nødvendigheden af kvalifikatorer og den nemmeste måde at implementere beskrivelsen i HTML-kodningen.

Grunden til at man har valgt at indkorporere Dublin Core elementene i selve HTML-kodningen er primært at det vigtigste element i Internettet i øjeblikket er det hypertekstbaserede WWW. Da man ikke ønsker at beskrivelsen skal figurere på skærmen, hver gang man kommer ind på et Internetdokument, er det nødvendigt at Dublin Core elementerne optræder som såkaldte META tags på lige fod med f.eks. baggrundsfarve, skriftstørrelse osv. På den måde vil søgemaskinernes indekseringsrobotter kunne identificere beskrivelsen uden at den optager plads på selve pc-skærmen (Dempsey, 1996; Heery, 1997:II; Weibel, 1996).

En anden hjælp til beskrivelsen af Internetdokumenter er at alle felter i Dublin Core kan dubleres. Dette kan være nødvendigt måske især i forhold til elementer som forfatter/ophav, andre ophav, emne og relationer.

Der findes en række eksempler på igangværende forsøg med brug af metadata til brug for Internetbeskrivelser bl.a. Nordisk metadata projekt (http://linnea.helsinki.fi/meta/index.html), The National Library of Australia and the National Library of New Zealand's National Document and Information Service (NDIS) Project (http://www.nla.gov.au/2/NDIS) og SOLINET's Monticello Electronic Library (http://www.solinet.net/monticello/monticel.htm)

Dublin Core og implikationer for genfinding af Internetressourcer

Som tidligere omtalt er et af de største problemer ved både automatisk indeksering - som den foretages i søgemaskiner - og manuel katalogisering af Internetressourcer identifikationen af relevante data i dokumenterne. Såfremt Dublin Core vinder indpas som standard for ophavsbaserede beskrivelser, vil dette medføre en væsentlig forbedring af datagrundlaget for både automatisk og manuel indeksering, da formatet både specificerer og separerer de 15 forskellige datatyper.

Implikationer for søgemaskiner

I søgemaskinerne er de beskrivelser, der ligger til grund for brugernes identifikation og relevansbedømmelse, baseret på automatisk udtrækning af

data. Fremfor at udtrække bestemte dele af selve teksten kan en overførsel af Dublin Core elementer danne basis for en beskrivelse. En sådan beskrivelse vil præsentere de samme datatyper, såfremt de er indekseret af ophavet, i den samme rækkefølge for alle dokumenter, hvilket vil lette brugerens bedømmelse af det enkelte dokument samt valg mellem flere dokumenter fremkommet ved samme søgning. Dublin Core kan også danne udgangspunkt for flere forskellige vis-formater.

Et af kravene til Dublin Core var, at alle elementer skal have søgerelevans, og det vil være indlysende at værterne for søgemaskinerne benytter separeringen mellem de forskellige datatyper til at tilbyde feltspecifik søgning i alle eller flere af de eksplicierede datatyper (f.eks. forfatter- eller titelsøgninger). Dette er dog på ingen måde uproblematisk. Kravet om syntaks-uafhængighed samt ophavenes manglende kendskab til kontrollerede vokabularer (og måske endda en ressourcebetinget manglende vilje til at benytte sådanne) medfører, at man ikke kan operere med vedtagne navneformer og kontrollerede emneord, ligesom man kun i begrænset udstrækning vil udfærdige regler og retningslinjer for, hvorledes ophavene skal udfærdige navne, titler, emneord m.m.. Dette medfører, at automatisk genbrug af Dublin Core elementerne på ingen måde sikrer samling og entydig identifikation af f.eks. dokumenter som skyldes samme ophav eller af dokumenter omhandlende samme emne. Brugen af kontrollerede vokabularer er ikke kun problematisk i indekseringsfasen, men betinger også at søger kender og benytter sådanne i formuleringen af forespørgsler til systemet.

Brugen af Dublin Core elementer ved den automatiske indeksering i søgemaskinerne kan altså medføre fyldigere og mere ensartede beskrivelser af dokumenterne. Det vil dog kun i begrænset omfang medføre forbedrede søgeresultater. Specielt i forhold til recall, da man ikke har nogen garanti for at samme ting kaldes ved samme navn, hvorfor slutbrugerne selv skal medtænke og indkorporere synonymer og alternative navneformer i deres søgninger.

Implikationer for strukturerede emneindekser
Som for søgemaskinernes vedkommende vil Dublin Core elementerne, såfremt de er indekseret af ophavet, kunne danne baggrund for en automatisk generering af fyldigere og mere ensartede beskrivelser i de strukturerede emneindekser, hvor de fleste p.t. kun angiver titlen.

Dublin Core indekseringen vil også lette identifikation og selektion af relevante dokumenter, der skal indgå i et emneindeks.

Ideelt kan Dublin Core elementerne "Emne" og "Beskrivelse" lette indekseringen - placeringen af dokumentet i indekset - for de, der vedligeholder et givet emneindeks, men i praksis vil emneindeksets målgruppe, indeksørens og ophavet's opfattelse af - eller tilgang til - emneområdet nok ofte afvige væsentligt fra hinanden, hvorfor en direkte indplacering udelukkende baseret på ophavets indeksering sjældent vil være tilstrækkelig. Både i forhold til materialevalg og indeksering vil det som oftest være nødvendigt at orientere sig i selve dokumentet.

Implikationer for bibliografiske databaser og kataloger
Dublin Core formatets felter kan bruges direkte ved opbygningen af databaser. Dette kan f.eks. blive tilfældet i søgemaskinernes databaser.
Hvis man vil indeksere Internetressourcer i eksisterende bibliografiske databaser - herunder også lokale kataloger, vil ophavsgenererede indekseringer som Dublin Core tilvejebringe et forbedret datagrundlag for indekseringen, men en tilpasning til databasens inddateringsformat er nødvendig.
Der er udarbejdet konverteringstabeller mellem Dublin Core og forskellige MARC-formater (Mapping the Dublin Core Metadata Elements to USMARC 1995; Day 1996), men en automatisk konvertering uden efterfølgende manuel tilretning er problematisk:

- Ophavets forståelse af felt-/elements-funktion kan, selvom intellektuel simpelhed er et af kravene til Dublin Core, afvige fra det tiltænkte.

F.eks. kan ophavet betragte en "webmaster" som væsentligste ophav, selvom dokumentet har en egentlig forfatter. Ligesom ophavet vil kunne sammenblande elementer som "emne" og " type".

- Kravet om syntaksuafhængighed medfører en nødvendighed for tilpasning af navneformer og emneord såfremt databasen eller katalogen i en eller anden udstrækning tilstræber kollokation: *Skal navne udformes direkte eller inverteret; hvordan forholder man sig til sammensatte efternavne; kontrol med titler - specielt periodicatitler; benyttes kontrolleret vokabular ved tildeling af emneord...m.m.*

Sidstnævnte problem kan delvis tackles, idet det er muligt for ophavet i forlængelse af et givet Dublin Core felt at angive , hvilket vokabular eller regelsæt der er anvendt ved indekseringen af feltet. Dette betinger selvfølgelig, at ophavet har benyttet et givet vokabular eller regelsæt, hvilket nok sjældent vil være tilfældet for andre end ophav tilknyttet BDI-sektoren.

Eksempel på anvendelse af kvalifikator: *Såfremt dokumentets emne er angivet ved en notation fra f.eks. UDK, kan oplysningen om at UDK er benyttet som vokabular medføre, at notationen konverteres fra DC-feltet "emne" til MARC-feltet "080", som anvendes til UDK-notationer*

Konklusion

Ideen med ophavsgenererede ensstrukturerede metadata vedhæftet Internetressourcer kan på flere områder lette indeksering og genfinding af ressourcer tilgængeliggjort via nettet. Dette betinger dog, at ideen vinder indpas hos ophavene til sådanne ressourcer.

En inkorporering af Dublin Core elementerne i de editorer, der benyttes til genereringen af Internetdokumenter, vil højne sandsynligheden for, at ophavene foretager den ønskede indeksering af egne dokumenter. Såfremt de institutioner, der udarbejder søgemaskiner, emnelister og andre hjælpemidler til genfinding af Internetressourcer, baserer deres udvælgelse af dokumenter på tilstedeværelsen og kvaliteten af metadata i dokumenterne, vil dette alt andet lige betyde en øget interesse hos ophavene for at generere og vedhæfte metadata..

Ophav, der er interesseret i en bedre formidling og øget fremfinding af deres dokumenter, vil naturligt have interesse i at vedhæfte metadata til deres dokumenter. Ligesom sådanne ophav på længere sigt nok vil være tilbøjelige til at stille krav til editor-producenter og udbyderne af genfindingsværktøjer om, at deres indsats resulterer i en øget benyttelse af deres ressourcer. Vi tror, at en sådan reaktion med tiden vil kunne føre til en øget forståelse hos ophav, formidlere og slutbrugere af Internetressourcer for, at den blotte tilstedeværelse af metadata ikke er tilstrækkelig, men at det også er væsentligt at sikre kvaliteten - og specielt ensartetheden - af disse metadata. En sikring af kvaliteten af metadata vil naturligt betinge implementeringen af regler og retningslinjer for udarbejdelsen af disse, hvilket absolut vil være en opgave som biblioteksinstitutioner og -personale kan medvirke til varetagelsen af.

Referencer

Caplan, Priscilla *1996*
Metadata for Internet Ressources: The Dublin Core Metadata Element Set and Its Mapping to USMARC / Priscilla Caplan and Rebecca Guenther. I : Cataloging and Classification Quarterly, vol. 22 (3/4) 1996. - s.43-58.

Assessing information on the Internet 1993
Assessing information on the Internet : toward providing library services for computer-mediated communication / Matt Dillon et al. - Dublin, Ohio : OCLC Online Computer Library Center, 1993.

Day, Michael *1996*
Mapping Metadata Formats to MARC / Michael Day. I: BIBLINK - LB 4034 :
D1.1 Metadata Formats / Rachel Heery. [citeret 7. august 1997]. Tilgængelig på Internet:
http://www.ukoln.ac.uk/metadata/BIBLINK/wp1/d1.1/doc0010.htm

Dempsey, Lorcan *1996*
The Warwick Metadata Workshop : A Framework for the Deployment of Resource Description / Lorcan Dempsey; Stuart L. Weibel I: D-Lib Magazine, July/August 1996 [citeret 9.juni 1997]. Tilgængelig på Internet:
http://hosted.ukoln.ac.uk/mirrored/lis-journals/dlib/dlib/dlib/july96/07weibel.html

Hakala, Juha *1997*
Dublin Core Metadata Element Set and it's applications / Juha Hakala. [citeret 9. juni 1997]. Tilgængelig på Internet:
http://linnea.helsinki.fi/meta/present.html

Heery, Rachel *1997:I*
Review of Metadata Formats / Rachel Heery [citeret 9. juni 1997]. Tilgængelig på Internet:
http://www.ukoln.ac.uk/metadata/review.html

Heery, Rachel *1997:II*
The 4th Dublin Core Workshop : (DC Down Under) 3-5 March 1997, Canberra, Australia / Rachel Heery et al. [citeret 9.juni] Tilgængelig på Internet:
http://www.ukoln.ac.uk/metadata/resources/dc4-notes.html

Lagoze, Carl *1996*
The Warwick Framework : A container Architecture for Aggregating Sets of Metadata / Carl Lagoze; Clifford A. Lynch; Ron Daniel Jr. [citeret 9.juni 1997]. Tilgængelig på Internet:
http://cs-tr.cs.cornell.edu:80/Dienst/Repository/2.0/Body/ncstrl.cornell%2fTR96-1593/html

Mapping the Dublin Core Metadata Elements to USMARC 1995
[citeret 7.august 1997]. Tilgængelig på Internet:
http://www.nlc-bnc.ca/ifla/documents/libraries/cataloging/metadata/dp86.txt

Sha, Vianne T. 1995
Cataloguing Internet resources: the library approach / Vianne T. Sha. I : The Electronic Library, Vol.13, no.5, 1995. - s.467-76.

Smith, Terrence R. *1996*
The Meta-Information Environment of Digital Libraries / Terrence R. Smith. I: D-Lib Magazine, July/August 1996. - [citeret 9. juni 1997]. Tilgængelig på Internet:
http://www.dlib.org/dlib/july96/new/07smith.html

Weibel, Stuart L. *1995*
OCLC/NCSA Metadata Workshop Report / Stuart L. Weibel et al. [citeret 9. juni 1997] Tilgængelig på Internet: http://www.oclc.org:5046/oclc/research/publications/weibel/metadata/dublin_core_report.html

Weibel, Stuart L. *1996*
A Proposed Convention for Embedding Metadata in HTML / Stuart L. Weibel. [citeret 9. juni 1997]. Tilgængelig på Internet: http://www.oclc.org:5046/~weibel/html-meta.htm

Woodward, Jeannette *1996*
Cataloging and Classifying Information Ressources on the Internet / Jeannette Woodward. I : Annual review of Information Science and Technology / ed. by Martha E. Willams. - New Jersey : Published on behalf of the American Society of Information Science by Information Today, 1996. - s. 189-220.