Kortlægning af standardiseringsmæssige tiltag og behov for samme

Price, Adrian; Løvschall, Kasper

Publication date:
2004

Document Version
Også kaldet Forlagets PDF

DEF Programområde: E-publicering

Projekt 1a:

Kortlægning af standardiseringsmæssige tiltag og behov for samme

En rapport udarbejdet for DEF af:

Adrian Price (DVJB) og
Kasper Løvschall (AUB)

November 2004
Version 1.1

Projekt 1a: Kortlægning af standardiseringsmæssige tiltag og behov for samme

Indholdsoversigt:

Summary:

I 2003-2004 gennemføres en række projekter i DEF programkomité for elektronisk publicering. Ét af projekterne udmøntes i denne rapport om kortlægning af standarder og best practices inden for elektronisk publicering.

Det er vigtigt, at de i DEF og andre tjenesters udviklede services sker i samspil med andre offentlige instanser. Herved sikres, at brugernes adgang sker via konsistente og sammenhængende grænseflader samtidig med at de bagvedliggende data er åbne og dokumenterede. Derfor er det vigtigt, at centrale standarder og best practices indarbejdes for at sikre denne interoperabilitet.

Det primære formål med rapporten er derfor, at skabe:

- En oversigt over standarder og "best practices" i forbindelse med elektronisk publicering
- En oversigt over standardiseringstiltag i udvalgte lande, som vurderes at kunne tjene til inspiration for en forbedret dansk praksis
- En undersøgelse af behovet for standardiseringstiltag i fælles DEF regi

Rapporten er bygget op i tre dele, hvor første del beskæftiger sig med den tekniske og organisatoriske infrastruktur herunder fastlægger dele af den terminologi og de forudsætninger, der anvendes gennem rapporten.

Anden del beskæftiger sig kort med informationsflowet for adgang herunder formidling samt bevaring i relation til formater, standarder, software og data- og informationsudveksling.

Tredje beskriver udvalgte indholdselementer for de i anden del fastlagte elementer i informationsflowet.

Anbefaling:

Det er vigtigt, at der på tværs af organisationer og instítutioner foregår en samordning af anvendte standarder og best practices på e-publiceringsområdet. Det skal selvfølgelig ske på de områder, hvor dette er hensigtsmæssigt. To meget oplagte områder vil naturligt være: forskningsregistrering samt etableringen af institutional repositories.

Her og nu adgang til universiteternes forskningsproduktion er ikke i sig selv tilstrækkeligt. Det er af stor nødvendighed, at langtidsbevaring og -tilgængelighed indarbejdes i de fælles initiativer. OAIS-modellen, som rapporten refererer til og selv anvender som referenceramme, kan anvendes som model for bevaringsaktiviteter, men på flere niveauer (detaljeringsgrad) afhængig af institutionens formål og funktion.

Institutionerne bør tage stilling til hvilke dokumenttyper, man vil anbefale til opbevaring med henblik på den anvendelse af materialet, der sker både her og nu samt i fremtiden. Egnetheden af en række dokumenttyper er tvivlsom eller uafklaret. Det er vigtigt at bemærke, at der må være tale om en beslutning på baggrund af en række kompromisser, da der ikke kan laves en entydig anbefaling.

Der skal sikres en kvalitet af opsamlede metadata, der tilgodeser både den lokale anvendelses krav samt muliggør en udveksling af disse data systemer imellem. Det betyder som oftest, at den største granularitet findes på det lokale niveau, men det skal samtidig være muligt at reducere dette til et mere overordnet niveau, som anvendes af andre systemer, uden at værdien går tabt.

Rapporten indeholder et eksempel på et muligt metadataformat (og herunder et sæt publikationstyper), der kan ses som et sæt kernefelter, der kan indgå i et fælles format.

XML er et anbefalelsesværdigt format til registrering af metadata på grundlæggende niveau, da dataudveksling forenkles og fordi formatet kan være selvbeskrivende. Ved opbygning af nye systemer bør XML altid overvejes. Det er blot vigtigt at bemærke, at anvendelse og konvertering til XML i sig selv ikke løser dataudvekslingsproblematikken. Det kræver en standardisering af de formater/skemaer, som institutionerne anvender. Dog er det langt enklere at konvertere fra et XML format til et andet (både XML eller helt andre formater) end det er fra mange af de andre metadataformater, vi anvender i dag. XML forenkler også muligheden for at tilgændeliggøre sine metadata gennem f.eks. webservices.

Som et vigtigt led i sikring af adgang til digitale ressourcer og genbrug af ressourcer på tværs af anvendelsesområder, er det afgørende, at der indføres et system, som vil sikre standardiseret navngivning af ressourcer uafhængig af ressourcernes lokalisering, for at opnå vedvarende og global adgang.

Standardiseret navngivning vil i sig selv ikke give et objekt status som "bevaring sikret for eftertiden" eller "adgang sikret for eftertiden". Det eneste, som kan garantere denne status, er den instans som til enhver tid har erklæret sig som "objektets kustode". Kravene må være, at et objekt kan (1) identificeres entydigt og det ligger i navngiving og metadata, (2) lokaliseres til enhver tid og (3) der findes en bevaringsstrategi for en ressource. I rapporten findes en række konklutioner og anbefalinger, der vedrører en mere konkretiseret stillingtagen til objektidentifikation.

Det er vigtigt, at der på institutionelt niveau efterleves de af staten stillede krav om tilgængelighed til information. Kravene er så veldokumenterede, at der ikke burde være problemer med at de services, der implementeres, overholder de givne krav.

Forord:

Med øget opmærksomhed på forskning og universiteternes ”ejerskab” af forskernes publicering af forskningsresultater, er der ved at opstå ”nye” adgangsveje til elektroniske forskningspublikationer med de såkaldte ”institutional repositories” og via protokoller som f.eks. OAI–PMH, som både samler og spreder digitale objekter og deres metadata. Samtidig er der øget opmærksomhed på bibliotekernes bevaringsforpligtelser over for digitalt materiale, som kræver indførelse af nye metoder – både teknisk og organisatorisk.

Denne rapport har lagt hovedvægten på de kernestandarder og best practices, der er forbundet med e-publicering. Det er forsøgt at placere disse standarder og best practices inden for en systematisk ramme, hvor langtidsbevaringsaspektet er en integreret del: adgang til digitale ressourcer forudsætter hensyntagen til bevaring. Men kun de overordnede aspekter af langtidsbevaring af digitale objekter er beskrevet, og de specifikke problemstillinger som vedrører bevaring af digitalt materiale, er **ikke** behandlet. De vil blive behandlet i fremtidige DEF projekter.

A: Indledning:

A.1: Teknisk og organisatorisk infrastruktur:

Udvikling af en national teknisk og organisatorisk infrastruktur for elektronisk publicering kræver fastlæggelse og overholdelse af en række standarder. Men overholdelse af standarder vil ikke i sig selv sikre brugbare løsninger, og her er det vigtigt, at man inddrager erfaringer høstet fra andre lignende tjenester. Erfaringer fra disse "best practices" kan være tekniske, organisatoriske eller handle om hvordan resultaterne formidles.

Fastlæggelse og overholdelse af standarder og best practices er afgørende for at sikre (1) adgang til information og (2) langtidsbevaring af information. Standarder og best practices findes i hele informationsflowet fra produktion, gennem indlemmelse og genfinding i kataloger og arkiver og til foranstaltninger som sikrer langtidsbevaring.

I januar 2002 udkom Open Archival Information System (OAIS) referencemodel, en anbefaling udarbejdet af Consultative Committee for Space Data Systems (CCSDS) [OAIS]. Referencemodellen fastlægger det organisatoriske og/eller systemmæssige grundlag, som er nødvendigt for langtidsbevaring af ”informationsbærende objekter”. Når vi betragter langtidsbevaring som en integreret del af biblioteksvirksomhed, som foregår parallelt med formidling, kan OAIS-modellen bruges til at klarlægge terminologi og standarder og til at fastlægge et hensigtsmæssigt work- og informationsflow.

I afsnit B fig. 1 er et informationsflow illustreret og danner grundlag for behandling af standarder og best practices i denne rapport. Der er ”lånt” fra OAIS-modellen hvor nødvendigt, men det understreges, at aktiviteter i forbindelse med langtidsbevaring behandles ikke i detaljer i denne rapport: OAIS-model er blevet brugt til at give rapporten en systematik. Fokus er ikke på modellen men på standarder og best practices mm. men forhåbentlig kan modellen anvendes og udvides i fremtidige projekter, hvor det er relevant.

Samtidig med publicering af OAIS-modellen blev begrebet ”trusted digital repository” (TDR) [TDR] lanceret. TDR er et vigtigt begreb i arkivernes og bibliotekernes formidlings- og bevaringsvirksomhed, da det sætter fokus på det organisatoriske og administrative element. Eller sagt på en anden måde, begrebet gør opmærksom på det faktum, at det ikke kun er ”teknik” eller ”systemer”, som skal inddrages, og som løser de opgaver, der er forbundet med formidling og bevaring af digitale objekter. Det er også et spørgsmål om, at institutioner skal anvende metoder og procedurer, som overholder kvantitative og kvalitative mål til sikring af adgang til digital information, og at disse metoder og procedurer skal dokumenteres. OAIS-modellen og TDR-begrebet går hånd i hånd. Igen er fokus i denne rapport ikke på indhold i TDR-konceptet, men der henvises til det, når det organisatoriske element i en bestemt løsning eller område er vigtig.

OAIS-modellen og TDR-begrebet giver os et grundlag for en fælles terminologi, en ramme for at beskrive funktioner og aktiviteter, en sammenhæng mellem informationsflow og organisation og et parallelt fokus på både formidling og bevaring af digitalt materiale. De kan med fordel anvendes systematisk i fremtidige projekter.

A.2: Terminologi:

Som på mange andre områder er det svært at fastlægge en entydig terminologi for e-publicering og vi har ikke forsøgt i rapporten at anvende en ”vandtæt” terminologi.

Digitale objekter: Et implicit fokus i denne rapport er "e-prints", dvs. hovedsagelig tekstdokumenter, som svarer nogenlunde til trykte dokumenter, som er blevet overført til et digitalt medie, f.eks. når et bibliotek etablerer et arkiv for en forskergruppes akademiske afhandlinger, artikler mm. I virkeligheden er der dog tale om ”digitale objekter”, da disse ”objekter” – om de så består af tekst, billede, lyd eller er sammensatte – ikke behandles særskilt i rapporten. Formidling og langtidsbevaring af de forskellige objekttyper foregår på et andet, lavere niveau i informationsflowet. Fokus i rapporten er på ”metadata niveauet” og hovedsagelig på metadata og aktiviteter forbundet med genfinding af objekterne.

Men når vi taler om universiteternes produktion af afhandlinger, artikler mm. er ”e-prints” stadig dominerende i institutional repositories og andre kataloger, men det vil være misvisende at bruge ordet ”e-print”. I stedet for vil betegnelser som ”digitalt objekt”, ”objekt”, ”digitalt materiale”, ”ressourcer”, ”e-ressourcer” osv. blive anvendt.

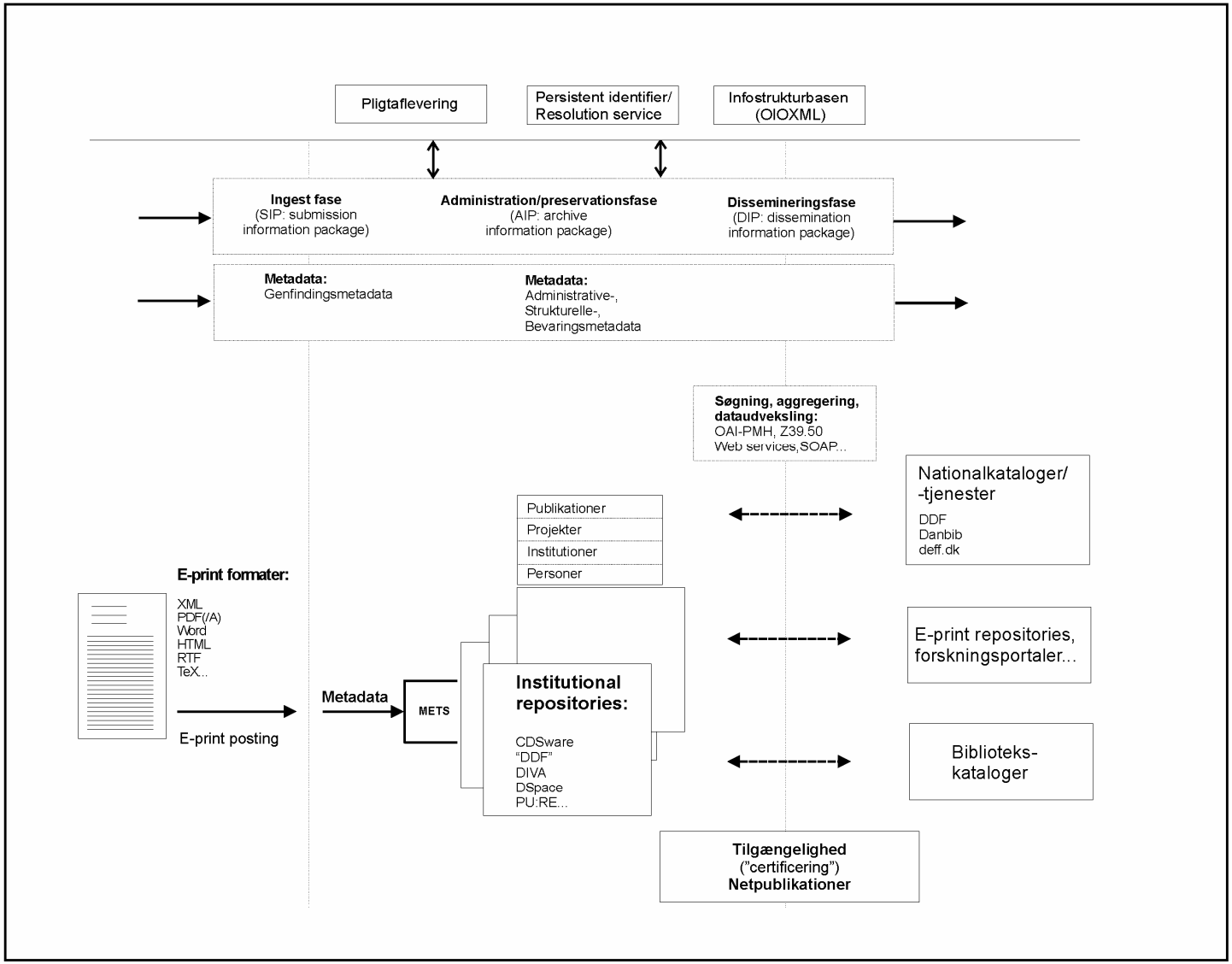
Repositorier: I de sidste par år er betegnelsen ”institutional repositories” blevet lanceret som et nyt begreb, også for at sætte fokus på en bevægelse mod en ny måde at publicere på: Det er institutionerne selv, f.eks. universiteterne, som først og fremmest skal råde over forskernes produktion af forskningsresultater. Der er blevet foreslået flere danske alternativer til begrebet ”institutional repositories”, f.eks. ”digitale arkiver”, ”digitale forskningsarkiver” eller ”institutionelle repositorier”. I sidste ende må det være et politisk/formidlings spørgsmål om en ”institutional repository” er andet end et ”digitalt bibliotek”, en ”katalog” eller om ordet ”arkiv” er vigtigt. I rapporten tages ikke stilling til terminologien og der anvendes flere betegnelser i flæng.

B: Informationsflow:

Fig.1 viser en række elementer i et informationsflow, som strækker sig fra produktion af digitale objekter, indlemmelse i et repositorie, tilgængeliggørelse via søgning og aggregering, formidling til brugerne og frem til pligtaflevering og andre langtidsbevaringsaktiviteter. Figuren er ikke et udtømmende billede af informationsflowet og bruges dels som en slags "indholdsfortegnelse" til dette dokument og dels som et forsøg på at give et systematisk grundlag at bygge videre på.

Når fokus er på repositoriet, kan man se elementer fra OAIS funktionelmodel [OAIS,4-1] i det landskab som tegnes i fig.1.:

Fig. 1:


Pligtaflevering
Persistent identifier/
Resolution service
Infostrukturbasen
(OIOXML)
Ingest fase
(SIP: submission
information package)
Administration/preservationsfase
(AIP: archive
information package)
Dissemineringsfase
(DIP: dissemination
information package)
Metadata:
Genfindingsmetadata
Metadata:
Administrative-,
Strukturelle-,
Bevaringsmetadata
Søgning, aggregering,
dataudveksling:
OAI-PMH, Z39.50
Web services,SOAP...
Nationalkataloger/
-tjenester
DDF
Danbib
deff.dk
Publikationer
Projekter
Institutioner
Personer
E-print formater:
XML
PDF(/A)
Word
HTML
RTF
TeX..
E-print posting
Metadata
METS
Institutional
repositories:
CDSware
"DDF"
DIVA
DSpace
PU:RE...
E-print repositories,
forskningsportaler...
Biblioteks-
kataloger
Tilgængelighed
("certificering")
Netpublikationer

En koncis beskrivelse af flowet kan illustrere de elementer og aktiviteter, som er involverede:

Et digitalt objekt (som kan være en artikel, afhandling, bog…) produceres i et bestemt format eller er et sammensat objekt. Objektet afleveres og indlemmes i et institutional repository (jeg synes ikke du skal blande UK og DK stavemåder i de sammensatte ord – altså enten: 'institutionelt repositorie' eller 'institutional repository'), i arkivet. Til objektet knyttes de nødvendige og tilstrækkelige metadata, som sikrer objektets administration, genfinding og bevaring. Repositorier, som kan være lokale, nationale eller internationale kan optræde som data og/eller servicerepositorier, dvs. objekterne kan udveksles med eller transporteres til og/eller fra andre repositorier. Denne udveksling og transport af dataobjekter foregår ved brug af diverse protokoller og aftaler. Formidling af objekterne sker med hensyntagen til standarder for netpublikationer og tilgængelighed, og grænseflader udvikles med hensyntagen til best practices inden for informationsarkitektur. Der skal sikres adgang til objekterne i repositorierne, også for fremtiden, ved at sikre at objekterne tildeles en vedvarende adresse, og at der sker en løbende test og opretholdelse af objektets integritet.

Informationsflow for adgang og bevaring
Etablering af en infrastruktur for et repositorie sikrer ikke kun adgang her og nu. Som nævnt, er foranstaltninger for langtidsbevaring ved at komme i fokus, og det kan anbefales at opbygge en organisatorisk og teknisk infrastruktur, hvor der tages højde for begge aspekter. Der kan være forhold (f.eks. systemmæssige begrænsninger, forskellige arkivforpligtelser osv.), som betyder, at der endnu kun fokuseres på adgang til digitale ressourcer, men der er elementer, som har indflydelse på både adgang og bevaring.

Forbindelser mellem adgang og bevaring nævnes i teksten, hvor det er relevant. Det er ofte "to sider af samme sag". Klarlægning og etablering af f.eks. et repositories bevaringsaktiviteter (som vil variere fra repositorie til repositorie) kræver specialiseret viden og behandles ikke i denne rapport.

C: Standarder:

C.1: Formater:

Et objekts format bestemmes som regel af en lokal brugergruppes arbejdsbetingelser, traditioner samt behov og ikke nødvendigvis af hensyn til hvilke formater, der vil være mest hensigtsmæssige at anvende set fra en håndterings- eller bevaringssynsvinkel. Selvom produktion af et digitalt objekt kan være uden for et repositories direkte kontrol, er det en vigtig fase, da filformater har stor betydning for adgang og bevaring. Det er vigtigt for et repositorie at forsøge at få indflydelse på valg af formater hos brugerne, da formaterne skal håndteres af repositoriet, når objekterne skal indlemmes samt at der skal tages hensyn til anvendelsen af indholdet – både af dagens og fremtidens brugere.

Der anvendes som regel et miks af formater og disse skal kunne håndteres af repositoriet.

En række gængse filformater er:

- MS Word (og andre Microsoft dokumentformater)
- RTF
- HTML og XHTML
- XML
- PDF

- Postscript (PS)
- TeX og LaTeX
- JPEG og TIFF

Der findes flere filformater for både tekst, lyd og billeder.

Det er i høj grad hensynet til langtidsbevaring, der bestemmer hvilke filformater repositoriet skal tillade. Objekterne som indlemmes nu er formentlig tilgængelige med dagens teknologi, men spørgsmålet er om denne teknologi også er tilgængelig fremover. Og jo flere formater man skal bevare, jo vanskeligere bliver opgaven. Repositorier vil oftest specificere hvilke formater, der kan indlemmes og denne "positivliste" skal i høj grad fastlægges af hensyn til opretholdelse af den langsigtede adgang til objekterne, som repositorier forpligtiger sig til.

Disse hensyn kan være i modstrid med andre faktorer, f.eks. en brugergruppes anvendelse af et bestemt format, som igen kan kræve specielle hensyn mht. adgang og bevaring.

Generelt kan det anbefales at vælge tilladte formater ud fra flg. kriterier:

1) Så få forskellige tilladte formater som muligt.
2) Formater skal helst have en specifikation som er offentlig tilgængelig.
3) De skal helst ikke være proprietære, dvs. ejes af kommercielle virksomheder hvor specifikationer er lukkede med risiko for at de forsvinder fra markedet.
4) Formatet skal helst være udbredt. Hvis brugergruppen er stor, er der større chancer for at formatet supporteres og udvikles for fremtidig brug.
5) Formatet skal i sig selv være uafhængig af andre programmer, hardware eller anden teknik.
6) De tilstrækkelige og nødvendige metadata skal følge objektet så adgang til og bevaring af objektet sikres.

Det tyske DissOnline projekt [DISSONLINE] for indsamling, registrering og bevaring af tyske disputatser mm. har fastlagt hvilke filformater, de foretrækker at modtage [DISSFORMAT]. Deres foretrukne formater er, i prioriteret rækkefølge: 1. XML/SGML 2. HTML 3. PDF 4. PS og 5. andre formater f.eks. RTF, MS Word, TeX, DVI[1] samt ASCII tekst osv. Disse formater er en anbefaling og ikke et krav.

I en opgørelse fra 1998 – 2001 af 300 disputatser afleveret til Humboldt Universitet i tilknytning til projektet Digitale Dissertationen, var fordeling blandt filformater: MS Word 75% og LaTeX 22%.

Der er plusser og minusser samt en række faktorer, som skal tages i betragtning for hvert format. Nedenfor gennemgås et udvalg af de gængse formater, med de forhold som skal tages i betragtning, for både indlemmelse i et arkiv og for langtidsbevaring.[2] Det er kun muligt her at gennemgå et udvalg af tekstformater. Andre formater, især hvad angår lyd og billedfiler, kræver specialiseret viden og er uden for denne rapports rammer.

[1] Device independent (DVI): et slutformat fra TeX og LaTeX og lignende tekstopsætningssystemer
[2] For en kort gennemgang af udbredte e-print filformater se (JISC, 27-31).

C.1.1: MS Word (og andre MS Office programmer)

Microsoft Word er i dag det mest anvendte dokumentformat. Word programmet og format er proprietært, og der findes få systemer, som kan producere formatet. Til gengæld findes der flere programmer, som kan konvertere fra Word til andre formater, f.eks. til PDF, XML, RTF. Formatet udkommer jævnligt i nye versioner, og nye formater kan give problemer for konvertering/bevaring.

Det faktum at et så udbredt format som Word ikke er egnet som arkivformat for langtidsbevaring, kan lige nu opvejes af, at formatet kan konverteres forholdsvis let til andre formater, og af at der findes en del programmer, som kan klare denne konvertering.

Ovenstående kommentarer gælder også for de andre programmer som udgør "Office pakken", og som også er meget udbredte (Excel, Power Point osv.)

Det er for tidligt at vurdere[3], hvilken betydning Microsoft Office 2003 XML Reference Schemas (som omfatter Wordprocessing ML, Spreadsheet ML, og FormTemplate Schemas) vil få for Microsoft produkters anvendelse i udvikling af eller deltagelse i arbejde med åbne standarder. Åbne standarder er vigtigt, set fra et adgangs- og bevaringssynspunkt. Disse skemaer er stadig forbundet med en licens, selvom licens er "royalty-free" (se: http://www.microsoft.com/mscorp/ip/format/xmlpatentlicense.asp). I Microsoft Office Word 2003 er det muligt at anvende eksterne XML skemaer, men der mangler endnu erfaring med dem.

C.1.2: PostScript - (PS)

I 1985 blev PostScript (PS) sprog udviklet af firmaet Adobe Systems Inc. PostScript sprog er et "page description" sprog. Selvom sproget er proprietært, er det et åbent sprog, hvor dokumentation er tilgængelig. En Encapsulated PostScript (EPS) fil er beskrivelsen af et billedelement i et "PostScript dokument", men mange programmer kan håndtere dette format separat. Udskrifter af PS filer kræver en printer med PostScript understøttelse, og fra en PS fil kan man fremstille en PDF fil. For visning af en PS fil på en skærm, kræves et program, f.eks. Ghostscript, som findes til (næsten) enhver platform. PS som arkivformat er ikke så udbredt[4], og der mangler erfaring med konvertering af PS til andre formater. [JISC,28-29]

C.1.3: Portable Document Format - (PDF)

PostScript er grundlaget for Portable Document Format (PDF) formatet, som også er udviklet af Adobe. PDF anvendes til at gengive det oprindelige dokument og er platform-uafhængigt. Ligesom PS er PDF et proprietært format, men det er åbent, da dokumentation er tilgængelig. PDF er et format af internettiden og udviklet med henblik på at distribuere dokumenter fra hjemmesider.

PDF er meget udbredt og er derfor velegnet som "formidlingsformat", men er mindre egnet i dag som arkivformat. PDF-formatet er proprietært og udkommer med jævne mellemrum i nye versioner, og disse ejerforhold og ændringer gør formatet mindre egnet som arkivformat.

[3] De første kommentarer til offentliggørelsen af formatet, har været, at det næppe vil få den store gennemslagskraft, da skemaet er så kompliceret, at de færreste vil kunne anvende det til noget brugbart.
[4] Forskningsmiljøer indenfor matematik og fysik udveksler og publicerer i større grad PostScript dokumenter

Herudover kan PDF-filer være afhængige af eksterne forhold, f.eks. kan et dokument være afhængigt af fonttyper, som hentes udefra.

Der findes konverteringsværktøjer fra PDF til f.eks. XML på markedet, og der arbejdes på udvikling af dem.

På grund af formatets enorme udbredelse og et akut behov for et format som vil sikre langtidsbevaring, arbejdes der på at udvikle et PDF-format, som kan anvendes til langtidsbevaring. I 2002 publiceredes et udkast til en standard for et PDF-arkivformat af en komité bestående af medlemmer fra Association for Information and Image Management (AIIM) og Association for Suppliers of Printing, Publishing and Converting Technologies (NPES). Dette udkast til en standard fik betegnelsen PDF/A (PDF-Archive) og er nu overtaget af ISO, hvor der arbejdes videre på formatet. [PDF/AISO] Man regner med, at en ISO PDF/A standard kan fastlægges i 2005.

I udkast til en standard for anvendelsen af PDF som arkivformat, blev flg. krav formuleret:

- Audio and video content are forbidden
- Javascript and executable file launches are prohibited
- All fonts must be embedded and also must be legally embeddable for unlimited, universal rendering
- Colorspaces must be specified in a device-independent manner
- Encryption is forbidden [PDF/ARLG]

I ISO Draft standard gøres brug af XMP - eXtensible Metadata Platform, udviklet af Adobe [XMPADOBE], for tilføjelse af metadata til bl.a. PDF-filer[5]. Metadata som følger PDF-filer er afgørende for preservation af et objekt gennem hele livscyklus. Fra Adobe Acrobat version 5.0 er der support for det nye XMP format [XMP], som forventes at blive taget op af W3C. XMP bruges til at indlejre metadata i applikationsfiler, f.eks. PDF, og vil betyde, at data om filerne (som kunne være relevante for preservation) vil følge som en del af ”pakken” i et arkiv. XMP er kompatibelt med XML/RDF.

En kort artikel om bevaring af PDF-filer findes. [PDFPRES]

C.1.4: Rich Text Format - (RTF)

RTF er udviklet af Microsoft som et platform- og applikationsuafhængigt format for både visning og udskrivning af tekst og illustrationer. Formatet er åbent og er meget udbredt i forskellige applikationer. Formatet anses for at være velegnet som arkivformat. [JISC,29]

C.1.5: Taggede formater – SGML, HTML, XML samt XHTML

Taggede (opmærkede) formater, som SGML, HTML og XML, er blevet ”flavour of the month” som foretrukne dokumentformaterings metode. Alle formater er afhængige af, at der findes en tilhørende Document Type Definition (DTD) eller som man anvender i dag: XML schema (skal det ikke være scheme?). Filerne er ren tekst og er derfor et åbent format, men der er problemer forbundet med formaterne. Anvendelse af tags kan f.eks. variere over tid, og man har forsøgt at definere stabile versioner med henblik på langtidsbevaring. Det kan også være svært at håndtere taggede dokumenter, da de som regel er sammensatte, med links f.eks.

[5] XMP anvendes allerede i flere af Adobes produkter.

til tekster/illustrationer som også er selvstændige objekter, i modsætning til f.eks. PDF dokumenter, som består af én fil.

HTML lider også under, at der ikke er krav til stringens i formatet. De forskellige fremvisere (typisk browsere) korrigerer selv for eventuelle fejl i opmærkning og struktur, som forfatteren har overset. Derfor er der ingen garanti for korrekt visning. Dette er der blevet forsøgt rettet op på i HTML's formodede arvtager XHTML, som er en sammensmeltning mellem HTML og XML.

C.1.6: TeX og LaTeX-formater

Både TeX og overbygningen LaTeX er makrosprog i familie med SGML, der anvendes til at beskrive typografi i tekster. Da makrosproget kan programmeres og redefineres står det enhver forfatter åbent at udvikle nye makroer og definitioner eller anvende andres. Formatet har vundet en stor udbredelse indenfor teknisk-naturvidenskabelig publicering, da det er det sprog, der har den bedste understøttelse for formelskrivning mm.

TeX bærer et nært slægtskab med programmeringssprog og man skal da også oversætte sine dokumenter i en "compiler" (oversætter) førend man kan skrive det ud på f.eks. en printer. Resultatet er en DVI (Device Independant) fil, der kan fødes videre til f.eks. en speciel printer- eller skærmdriver. Betegnelsen enhedsuafhængig skal dog tages med et gran salt, da DVI filerne ikke indeholder skrifttyper, hvilket gør, at filerne ikke nødvendigvis kan stå alene.

Selve tekstdokumenterne i TeX og LaTeX er dårlige udvekslingsformater, da de i sig selv er meget svært læselige og kan være afhængige af specielle makropakker og skrifttyper, som ikke nødvendigvis følger med dokumentet. Samtidig kan DVI filerne indeholde referencer til skrifttyper, som ikke er tilgængelige.

Den bedste håndtering vil være at konvertere DVI filerne til PostScript (det foretrækkes at forfatteren til dokumentet foretager dette for at sikre sig mod de tidligere nævnte problemer) eller PDF. Specielt ældre DVI dokumenter konverteret til PDF ser "uldne" ud på skærmen pga. anvendelse af bitmapskrifttyper og ikke mere moderne vektorskrifttyper.

C.1.7: Billedfiler – JPEG, TIFF

Både Joint Photographic Experts Group (JPEG) og Tagged Image File Format (TIFF) formater er nu åbne standarder, selvom de begge har en "proprietær fortid". Begge anses som anvendelige for langtidsbevaring, hvor TIFF betragtes mange steder som *de facto* standard, også f.eks. af Statens Arkiver.

C.2: Arkiv ingest og objektadministration:

Metadata:
Et digitalt objekt indlemmes i repositoriet i den fase som kaldes "ingest" i OAIS-terminologien (se fig. 1). Det kan være producenten selv, der, via "selv-arkivering", sørger for at objektet indgår, eller arkivet som indlemmer objektet. Hvis vi holder os til OAIS-terminologi, vil objektet i denne proces blive til en "information package". Informationspakken består af objektet, som er det indhold som skal gøres tilgængelig og bevares, og "meta-information" om objektet. "Meta-information" er metadata som tilknyttes objektet.

Det er hensigtsmæssigt at opdele metadataskemaer[6] (og -elementer) i forskellige typer, afhængig af det formål det pågældende element har i forhold til objektet. Der er pt. ingen konsensus på opdeling af metadata, men en opdeling i flg. 4 typer[7] anvendes ofte:

- **Genfindingsmetadata** (”resource discovery” metadata): metadata om objekter som sikrer, at de kan identificeres, også som relevante ift. en brugers konkrete behov, og gøres tilgængelige. F.eks. et objekts titel, ophav, emne osv. Denne type metadata betegnes ofte som deskriptive metadata.

- **Administrative metadata:** metadata som er nødvendige for administration af ressourcen, f.eks. dato for objektets indlemmelse eller opdatering, objektets indhold eller forhold til andre objekter osv.

- **Strukurelle metadata:** metadata som fastlægger hvordan sammensatte digitale objekter ”hænger sammen”, hvordan f.eks. kapitler, illustrationer og videosekvenser, som kan være separate filer, er relateret til hinanden og udgør objektet.

- **Bevaringsmetadata:** tekniske og administrative metadata som sikrer langtidsbevaring af objekter, f.eks. filformat, dato for næste filintegritetscheck osv.

Om et bestemt metadataelement opfattes som tilhørende en af de 4 metadatatyper, er som regel afhængig af anvendelse eller formål med elementet. F.eks. kan et objekts filformat både være relevant for adgang/benyttelse (deskriptive metadata) og for sikring af at filen kan benyttes fremover (bevaringsmetadata), og dette gør at det er umuligt at lave en vandtæt strukturering af metadata.

Ligesom metadata kan inddeles i forskellige typer, kan man også kigge på metadata, hvor anvendelsen opfylder bestemte formål. Set som en ”intra-arkiv” funktion, er formålet med anvendelsen af et metadataskema en *registrering* af objekterne til genfinding, selektion, bevaring osv. Set fra en ”inter-arkiv” funktion, kan formålet med anvendelse af et bestemt metadataskema f.eks. være at gøre objekterne tilgængelige gennem andre tjenester via søgning osv. Her anvendes metadata til *eksport/import* eller *høstning* fra/til ét arkiv/tjeneste til et andet arkiv/tjeneste.

De metadataskemaer, som anvendes til disse forskellige formål, kan (og i mange tilfælde vil) være forskellige. Et registreringsformat vil som regel skulle anvendes til flere funktioner, som skal støttes af metadata, som kommer fra måske alle 4 typer nævnt foroven, hvorimod et eksport-/import-/høstningsformat måske ”kun” har det formål at tillade genfinding af en ressource, f.eks. i en anden tjeneste.

En anden faktor som har indflydelse på udarbejdelse af et egnet metadataformat, er hvilke dokument-/ressourcetyper et repositorie forventes at skulle indeholde. Udover en kerne af metadataelementer, som vil være fælles for forskellige dokument-/ressourcetyper, har forskellige dokument-/ressourcetyper behov for forskellige metadataelementer til registrering, med henblik på genfinding, udveksling, bevaring osv.

[6] Et ”metadataskema” er en overordnet betegnelse for: (1) et sæt semantiske regler, dvs. elementerne og deres betydning og fastlæggelse af anvendelsen (2) et sæt regler, som fastlægger hvordan elementerne udformes og (3) syntaksen, hvordan elementerne nedfældes (bliver ”encoded”). [CAPLAN,6-7]

[7] I [JISC] rapporten om bevaring af e-prints opdeler metadata i: ”resource discovery” (= genfindings) metadata, ”technical preservation metadata”, og ”administrative metadata”.

Et metadataformat skal udarbejdes med arkivets formål i sigte og de dokument-/ressourcetyper, som arkivet forventes at indeholde. Arkivets formål er (som regel) registrering mhp. genfinding, bevaring og udveksling/aggregering med andre tjenester. Mht. dokument-/ressourcetyper, er det ikke muligt her at fastlægge et metadataformat for samtlige dokument-/ressourcetyper, men vi vil fokusere på dokumenttyper, som kendetegner et "e-print" arkiv. Dvs. de hovedsagelig tekstbaserede publikationstyper. Ved at tage hensyn til disse to faktorer, vil vi anbefale et metadataformat, som DEF repositorieprojekter kan anvende.

C.2.1: Fastlæggelse af et metadataformat:

Tabellen nedenfor er et forslag til et metadataformat, som kan anvendes af institutionelle repositorier, og som tager hensyn til et arkivs formål og dokumenttyper (se forrige afsnit)[8].

Udgangspunktet tages i formatet Dublin Core (DC), der er et format vedtaget af det åbne forum: The Dublin Core Metadata Initiative (DCMI) [http://dublincore.org], som arbejder for at udbrede og implementere interoperable metadata standarder og udarbejde specialiserede metadata vokabularier til beskrivelse af ressourcer med det formål at muliggøre mere intelligente informationssystemer. DC metadata standarden er et simpelt, men ganske effektivt sæt af elementer til beskrivelse af primært netbaserede ressourcer. Standarden i sin grundform består af femten elementer som er vedtaget på baggrund af international konsensus både indenfor og udenfor biblioteksverdenen. For en liste over de femten elementer og definitioner se [DCMI].

Selvom et af kernepunkterne i DC standarden har været enkelhed, har man erkendt, at formatet i sin grundform ikke egner sig som generisk registreringsformat. Det har bl.a. resulteret i en udvidelse af standarden til Qualified DC, som muliggør kraftig udvidelse af de femten standardfelter bl.a. ved "element forfining" , der kan gøre et DC element mere specifikt. Dertil kommer også muligheden for diverse "kodningsskemaer" herunder kontrollerede emneord eller at indhold i et felt skal opfylde en given ISO standard.

DC har vundet større udbredelse i forbindelse med udveksling af data mellem forskellige formater og dataservices. Her kan f.eks. nævnes Open Archives Initiativet (OAI) [http://www.openarchives.org], der arbejder for udbredelsen af interoperable standarder, der er målrettet mod facilitering og effektiv dissiminering af indhold fra f.eks. metadatabeskrevne arkiver. Oprindeligt er OAI grundlagt for at forbedre adgangen til e-print arkiver for at understøtte tilgængeligheden af uddannelsesinstitutioners vidensproduktion.

Da DC hovedsagelig bruges som udvekslings- og genfindingsformat, og da det ikke er velegnet som registreringsformat, er det nødvendigt med nogle udvidelser. Disse udvidelser er angivet som "DEF Core" (DEF-C) i tabellen og er samtidig et forslag til et fælles DEF "minimumsformat". Udover DC og DEF-C kan arkiverne lokalt udvide DEF-C format til at dække et lokalt behov for metadata.

DEF-C er en "flad" liste over metadataelementer og tager ikke stilling til metadatarepræsentation i et bestemt system.

Det kan være nyttigt at holde for øje, hvad metadata primært skal bruges til, når man går i gang med at fastlægge et format. Formål og anvendelse af DC, DEF-C og lokale felter kan opsummeres med nogle stikord:

[8] Formatet sigter kun mod registrering af forskningsresultater og ikke forskningsformidling, som har et bredere perspektiv.

DC elementer: er primært til identificering/genfinding (f.eks. DC Title) og lokalisering (DC Identifier) af objektet. Altså primært til registrerings- og udvekslings/genfindings-formål. DC bruges også til sikring af OAI kompatibilitet.

DC og DEF-C elementer: er til identificering/genfinding, lokalisering, adgang og anvendelse, f.eks. udvidet med emneord, typebetegnelser osv. af objektet.

DC og DEF-C og Lokale elementer: er til identificering/genfinding, lokalisering, adgang, anvendelse, formidling og analyse f.eks. udvidet med nødvendige institutionelle metadata, felter til bibliometriske analyser osv. af objektet.

Forslag til et fælles DEF metadataformat for institutionelle repositorier

No.	DEF Core (DEF-C)	Note	
1	Titel	Objektets/ressourcens navn	
2	Ophav	Person/organisation ansvarlig for indhold	
3	Emne	Overordnede, DEF26	
4	Beskrivelse	Beskrivelse af objektets/ressourcens indhold	
5	Udgiver	Primær udgivende enhed	
6	Bidragyder	Andre personer/institutioner som har bidraget til indholdet	
7	Dato	http://www.w3.org/TR/NOTE-datetime	
8	Ressourcetype	http://dublincore.org/documents/dcmi-type-vocabulary/	
9	Format	http://www.isi.edu/in-notes/iana/assignments/media-types/media-types	
10	Identifikator	"Primær" resolveable URI	
11	Sprog	http://www.ietf.org/rfc/rfc3066.txt http://www.loc.gov/standards/iso639-2/langhome.html	
12	Rettigheder	Erklæring vedr. rettigheder tilknyttet objekt/ressource	
13	Peer review	Peer review/ikke-peer review (af objektet i DC Title)	
14	Emneord	Detaljerede, (lokale) niveauer under DEF26	
15	Alternativ identifikator	Anden adgang til objektet angivet i "DC Title" end angivet i DC Identifier, f.eks. forlag, anden server	
16	Institution	Tilknyttede institutioner	
17	Primær enhed	f.eks. fakultet	
18	Sekundær enhed	f.eks. institut, center	
19	Tertiær enhed	f.eks. gruppe, sektion	
20	Publikations titel	Hvor DC Title publiceret, Journal/report/Conf.proceedings… title	
21	Publikationens identifikator	f.eks. ISSN, ISBN	
22	Bind		
23	Nummer		
24	Sider		
25	Redaktør		
26	Konference titel		
27	Konference dato		
28	Konference sted		
29	Lokal forskningskategori	Andre anvendte lokale kategorier end angivet i 8 Ressourcetype	
30	Status	Publiceret/ikke-publiceret	

Ressourcetyper (felt 8):

For at sikre interoperabilitet mellem repositorier samt mulighed for sammenligning mellem institutioner, er det vigtigt, at man anvender de samme ressourcetypebetegnelser. Forslag til ressourcetypebetegnelser (publikations- / dokumenttyper):

Bog
Kapitel i en bog
Tidsskrift
Artikel
Preprint
Rapport
Kapitel i en rapport
Avisbidrag
Konference
Konferencebidrag
Master afhandling
Phd afhandling
Doktorafhandling

Metadata og forhold til Den Danske Forskningsdatabase (DDF):

Et endeligt DEF metadataformat for institutionelle repositorier og DDF formatet fastlægges i et separat projekt i efteråret 2004.

C.2.2. Bevaringsmetadata:

Hvis langtidsbevaring af digitale objekter er et formål, som en institution skal opfylde, er et sæt bevaringsmetadata nødvendigt. Det er uden for denne rapports rammer at behandle bevaringsmetadata i detaljer, da det er et kompliceret og meget specialiseret område, som derfor skal behandles for sig. Vi vil i denne rapport nøjes med at give et overblik og henvise til det meget omfattende arbejde, i form af dokumenter og formater, som allerede eksisterer.

I det omfattende arbejde der er i gang mht. indførelse af metoder til bevaring af digitale objekter, anvendes OAIS referencemodel som regel som den grundlæggende model. OAIS-modellen giver infrastrukturen til bevaringsaktiviteter, som er nødvendige for styring af de involverede processer, og en del af de metadatamodeller, som er blevet fremsat, knytter sig i mange tilfælde også tæt til OAIS-modellen.

Cedars projektet [CEDARS], et eLIB finansieret projekt med deltagelse af 4 universitetsbiblioteker og som løb over 4 år (1998-2001), havde det formål at identificere, dokumentere og formidle metoder til langtidsbevaring af digitale objekter/samlinger. Der var 3 deltagende universitetsbiblioteker: Leeds, Cambridge og Oxford. Udover at undersøge og publicere anbefalede metoder og guidelines, har projektet udarbejdet et sæt bevaringsmetadata [CEDARSMETA], samt udviklet et ”demonstrator digital archive”, som en praktisk demonstration af de processer, som er nødvendige i preservation af digitalt materiale. OAIS-modellen er blevet anvendt til både metadata specifikation samt i implementering af arkivsystemet. Arkivet er en prototype på et distribueret system, og et af projektets formål var at undersøge, om en distribueret model var en farbar vej at gå. Formålet var også at undersøge

nogle koncepter og problemstillinger vedr. bevaring af digitale objekter, og på denne måde illustrere og klarlægge de tekniske, organisatoriske, og managementmæssige problemer, som er involveret i langtidsbevaring.

Et vigtigt resultat af Cedars-projekt har været en række rapporter, som illustrerer nogle af de kerneproblemstillinger, som er forbundet med langtidsbevaring af digitale objekter. Disse rapporter findes på Cedars' website [CEDARS]. Projektets slutrapport [CEDARSRAP] er et vigtigt dokument, som klarlægger både den teoretiske baggrund og de praktiske problemer og metoder, som er involveret i bevaring af digitale materialer.

Cedars projektet, efter publikation af metadataspecifikationen, har også deltaget i OCLC/ Research Library Group (RLG) Working Group on Preservation Metadata. De publicerede deres metadata framework i 2002 [OCLC/RLG] med baggrund i et white paper [OCLC/RLGWP]. OCLC/RLG har valgt at producere en "metadata ramme", som en syntese af 4 eksisterende skemaer: Cedars, National Library of Australia [NLAMETA], Networked European Deposit Library [NEDLIBMETA] og OCLC [OCLCMETA]. Denne "metadata-ramme" bygger også på OAIS-modellen.

OCLC/RLG working group fortsætter deres arbejde med bevaringsmetadata i PREMIS – PREservation Metadata: Implementation Strategies [PREMIS], hvor fokus er på udpegning af et sæt "core" metadataelementer og problemstillinger forbundet med implementering og management af metadata i et digitalt preservationssystem.

C.2.3. METS: en metadata ramme for digitale objekter:

Metadata Encoding and Transmission Standard (METS) [METS] er fastlæggelse af en "ramme" hvori deskriptive, administrative og strukturelle metadata, som udgør digitale objekter, kan implementeres i XML: " Without structural metadata, the page image or text files comprising the digital work are of little use, and without technical metadata regarding the digitization process, scholars may be unsure of how accurate a reflection of the original the digital version provides. For internal management purposes, a library must have access to appropriate technical metadata in order to periodically refresh and migrate the data, ensuring the durability of valuable resources." [METSOV] METS udvikles af Library of Congress og er hurtigt blevet adopteret af mange institutioner og systemer, som har behov for et format beregnet til både management af digitale objekter og formidling af disse objekter til brugerne.

Et METS XML dokument indeholder metadata for et objekt delt i 7 sektioner: 1. en METS header, som er information om selve METS dokumentet, 2. deskriptive metadata, 3. administrative metadata som er metadata om objektets komponenter og deres tilblivelse, 4. en filsektion, som er en liste over alle filer som udgør objektet, 5. en sektion som kortlægger objektets hierarkiske struktur ("structural map"), med linkning til filer og metadata i sektionen, 6. en "structural link" sektion med hyperlink mellem elementer i "structural map" sektionen, og til sidst 7. en "behavior section", hvor indholdselementerne i objektet kan associeres med en eksekverbare kode, f.eks. kan en billedfil vises i forskellige formater eller opløselighed.

METS kan også relateres til anvendelsen af OAIS som referencemodel: "Depending on its use, a METS document could be used in the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP) within the Open Archival Information System (OAIS) Reference Model." [METSOV] METS er forholdsvis nyt og er i fremgang, da det netop giver en ramme, som kan udnyttes i hele

informationsflowet, fra ”ingest” til ”out-gest”. Der udvikles på en række værktøjer, som understøtter disse muligheder. [METSREG]

C. 3: Søgning, aggregering, dataudveksling:

Der findes en række standarder til håndtering af søgning, aggregering og dataudveksling. En række vigtige nøglebegreber er her beskrevet:

XML (eXtensible Markup Language)
Er en forenklet dialekt af SGML og er en fleksibel metode til at gemme information i et struktureret format, som muliggør enkel lagring og udveksling af data.

Webservice
En webservice er et stykke software, der stiller sig selv tilgængeligt over internettet identificeret via en URI og anvender standardiseret XML til at kommunikere med. Den både modtager søgeforespørgsler og returnerer resultater i et forud defineret XML format.

SOAP (Simple Object Access Protocol) [SOAP]
Er en letvægtsprotokol og standard for udveksling af information med en web service og er baseret på XML. SOAP fungerer over http-protokollen, hvilket gør den meget enkel at arbejde med. Ofte anvendt i forbindelse med dynamisk information, der trækkes fra flere forskellige kilder; typisk databaser. Kan anses som indpakning for udveksling af informationer.

OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) [OAI]
Er en letvægtsprotokol, der definerer en mekanisme for høstning af XML-formateret metadata fra arkiver, der understøtter OAI. Protokollen indeholder ingen mekanismer til associering mellem metadata og det relaterede indhold, hvilket betyder, at dette skal være gjort i de udstillede metadata. Dublin Core metadataformatet er påkrævet som udvekslingsformat.

Z39.50
En international standard, der definerer en protokol for informationsudveksling. Protokollen muliggør søgning og fremfinding af informationer i et søgesystem uafhængig af hardware og software. Har fundet sin udbredelse og anvendelse i bibliotekssystemer, men er noget indviklet for udenforstående parter at anvende.

SRW/SRU[9]
SRW er en XML baseret protokol til søgning over internettet. Den anvender eksisterende teknologier til formålet så som: XML, URL, SOAP, HTTP og Xpath. Dataudveksling foregår i XML og kan typisk foregå via Dublin Core. Anvendes SOAP som transportlag kaldes protokollen for SRW (Search/Retrieve Web Service), mens den ved kommunikation direkte via URL’er kaldes for SRU (Search and Retrieve URL Service). Anvender Z39.50 semantik.

C. 4: Adgang og bevaring:

Som nævnt i indledningen er det et vigtigt salgsargument for etablering af e-print repositorier, at man kan tilbyde vedvarende adgang til forskeres digitale ressourcer. "Vedvarende adgang" betyder, at en ressource altid skal kunne lokaliseres til en gyldig adresse (URL), at det kan genfindes via metadata, og at der sker en løbende bevaring af ressourcens "digitale integritet".

[9] For en oversigtsartikel om SRW/SRU se: http://www.ariadne.ac.uk/issue40/

C.4.1: Permanente identifikatorer ("persistent identifiers"):

For at sikre at objekter i et arkiv bliver ved med at være tilgængelige, er det afgørende, at man kan garantere en entydig og vedvarende identifikation **og** adressering af objektet. En Uniform Ressource Locator (URL) er den mest udbredte metode i dag, hvor en ressources URL bruges til både at identificere (navngive) ressourcen, angive hvor ressourcen befinder sig, objektets adresse, samt den protokol som skal anvendes til at få adgang til ressourcen (f.eks. http, ftp osv.).

At en URL kombinerer både identifikation og adressering, anses for at være URL'ens svaghed: Hvis et objekt flyttes til en ny adresse, f.eks. en anden server, vil objektet tildeles en ny URL, som samtidig er en ny identifikation, og den vedvarende adgang til objektet vil blive svækket. Der er behov for et system, som klart og vedvarende adskiller identifikation/navngivning og adressering/lokalisering. Et objekt skal tildeles en entydig identifikator, et navn, som objektet beholder "hele livet". Til denne identifikator kan kobles objektets aktuelle adresse og i tilfælde af at objektet flyttes, kan den nye adresse kobles til den uændrede identifikator, og dermed fastholdes adgang over tid. Som erstatning for det nuværende URL-baserede lokaliseringssystem arbejdes der på at udvikle og indføre metoder og systemer, som bedre vil sikre fremtidig adgang til netressourcer. Der udvikles på metoder og systemer til indførelsen af det som oftest betegnes "persistent identifiers" på engelsk, permanente identifikatorer (PI).

Indførelse af et system som bedre vil sikre adgang til objekter over tid, er ikke kun vigtigt for "isolerede objekter", men er også af betydning for linkning mellem objekter: f.eks. fra en reference i en artikel til selve artiklen, styring af adgang til ressourcer i forbindelse med betaling for benyttelse og brug af objekter i flere forskellige sammenhænge, f.eks. i e-lærings miljøer[10]. En sikring af adgang til netressourcer er således et afgørende element i opbygning af en "digital infrastruktur".

C.4.1.1: Metoder og systemer til permanente identifikatorer:

I løbet af de sidste ca.10 år er der blevet arbejdet på forskellige metoder og systemer, som har til formål at sikre vedvarende adgang til e-ressourcer. Udgangspunkt for de fleste er erkendelsen af, at det er afgørende, at navngivning og lokalisering holdes skarpt adskilt. Dvs. at uafhængig af hinanden skal det fastlægges hvordan et objekt skal identificeres ved et entydigt og standardiseret navn og dernæst skal det fastlægges hvordan dette navn skal anvendes i lokalisering af objektet. At navngivning og lokalisering er adskilt og uafhængig, betyder, at man til enhver tid kan ændre lokalisering af objektet uden at svække adgang.

Det er også vigtigt at holde navngivning og adressering adskilt, da infrastrukturen for vedvarende adressering (lokalt og globalt) endnu ikke er fuldt udbygget eller afgjort. Til indførelse af PI'er er der navngivningskonventioner til udformning af identifikator (en meget udbredt konvention er f.eks. Universal Ressource Name (URN), se næste afsnit) og der findes softwaresystemer, som anvender standardiserede navngivningskonventioner og som kan implementeres til adressering af digitale objekter (f.eks. Handle systemet, se næste afsnit). Der findes metoder og systemer til etablering af infrastrukturen.

Resolution:

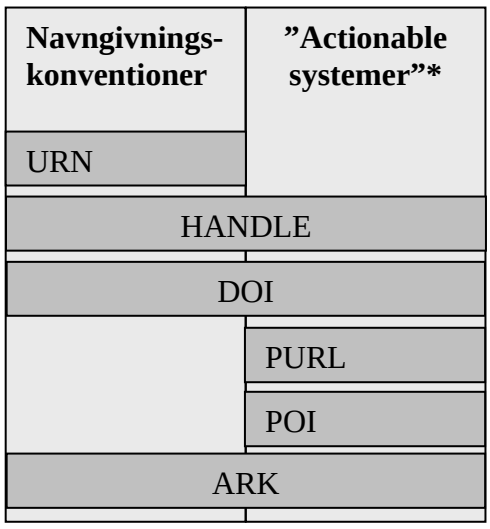
[10] For en speciel rapport om permanente identifikatorer og e-læring se [TSO].

Pga. adskillelsen mellem identifikation og adressering, er der behov for en mekanisme, som giver adgang til ressourcen via dens identifikator. Denne mekanisme betegnes på engelsk ”resolution”. DNS (Domain Name System) er sådan en ”resolution” mekanisme, som via et domænenavn, f.eks. www.bibliotek.dk, giver adgang (”resolves”) til en bestemt vært med et bestemt IP nummer.

Der er endnu ikke implementeret en global *resolution service* for andre protokoller end til almindelige URL’er (hostnames), så PI’er i dag skal kobles til en almindelig URL, som så anvender DNS-systemet til at give adgang til objektet. Fordi DNS systemet er så udbredt, bruger de fleste PI systemer i dag http-protokollens *redirection*, til at omstille til objektets aktuelle lokále. Denne situation forventes ikke at ændre sig på kort sigt og det vil også blive svært at ændre på længere sigt, da DNS systemet er så udbredt.

Der skal også nævnes et andet svagt punkt, dog af mindre betydning: almindelige internetbrowsere , hvor kun ”almindelige” URLs kan anvendes i dag. Dvs. for at anvende et PI i en browser, skal identifikatoren kobles til en almindelig URL, hvor der så anvendes http-protokol til at ”resolve” objektet til en adresse. At ændre denne browserbegrænsning, hvor der vil være mulighed for at anvende andre resolutionsprotokoller, vil kræve enighed om en ændring i browsere, som skal gennemføres af softwareproducenterne. Det forudsætter, at man er blevet enig om hvordan en ”resolutionsprotokol” skal fungere – og formentlig på globalt plan. (En midlertidig løsning vha. en browser-plug-in, hvor plug-in sørger for at oversætte til http, er muligt og er blevet afprøvet, men denne løsning er uholdbar i længden).

PI området deler sig i konventioner for navngivning af objekter alene og konventioner for navngivning **og** systemer for resolution af disse objekt-- ID’er til deres lokalisering. I det følgende beskrives kort de muligheder som findes i dag og i Fig. 2 nedenfor er en oversigt over disse konventioner og systemer. Disse konventioner og systemer beskrives i de efterfølgende afsnit.



*I et ”actionabelt system” er resolving mekanismen (lokalisering) specificeret som en integreret del af systemet

Fig. 2: Navngivningskonventioner og PI systemer

C.4.1.1.1: Uniform Resource Name – URN [URN]

Uniform Resource Name – URN – er en navngivningskonvention udviklet af Internet Engineering Task Force Network Working Group og fremsat i 1992. URN er nok den mest udbredte *konvention* for navngivning af ressourcer og anvendes af en række andre systemer.

Sammenhæng for URN er det "klassiske" men på nuværende tidspunkt ukomplette (eller snarere fragmenterede) udgangspunkt:

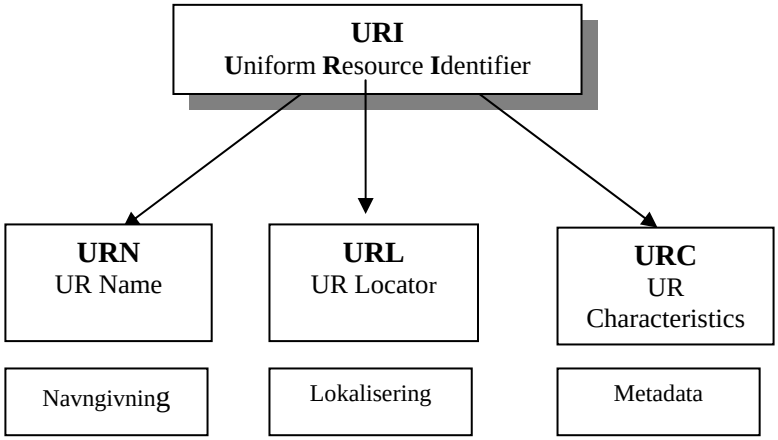


Fig. 3: URI og dens grundlæggende karakteristika

Figuren præsenteres her, fordi "grundsystematikken" bag URI stadig er vigtig at holde for øje. Navngivning, lokalisering og metadata (de 3 nederste elementer i figuren) udgør stadig grundpillerne i PI'er, selvom billedet i dag er mere fragmenteret og ukomplet.

En Uniform Resource Identifier (URI) [URI] specificerer en syntaks og semantik for identificering og for lokalisering af en ressource. En ressource tildeles en formel identifikation via et URN [URN] og ressourcen lokaliseres (adresseres) via en specificeret mekanisme (eller protokol) i en URL [URL]. (De mest udbredte protokoller er f.eks. http, ftp osv.)

Fordi et URN kun identificerer (navngiver) et objekt, skal der kobles en lokaliseringsmekanisme til, dvs. der kræves en såkaldt "resolver", som kobler objektets URN til ressourcens adresse. Der har været flere forsøg på at udvikle et *resolution system* til URN, sidst i 2001 med Dynamic Delegation Discovery System (DDDS), hvor det eksisterende Domain Name Service (DNS) foreslås anvendt til at lokalisere "resolvers" med information om ressourcerne. DDDS systemet er ikke slået igennem.

URN syntaks:

Hvis man ser på opbygning af et URN, så består et URN af følgende elementer: **urn:nid:nss** hvor "nid" er den registrerede Namespace Identifier og "nss" er Namespace Specific String, objektets identifikator, som tildeles af NID. NID registreres hos Internet Assigned Numbers Authority [IANA].

Et eksempel på et URN hvor NID er det registrerede namespace "NBN":

urn:nbn:fi-fe19981001

„nbn“ – National Bibliography Number – er den registrerede Namespace Identifier (NID) som definerer syntaksen: urn:nbn: <ISO country code>-<assigned string>. I ovenstående eksempel er: fe19981001 den Namespace Specific String (NSS), som tildeles ressourcen, i dette tilfælde af det finske nationale bibliotek. Det er også muligt at definere ”sub-namespaces” efter NID og i et URN er de adskilt med et ”:”.

NBN namespace blev oprindelig udviklet og registreret af det finske nationalbibliotek og siden accepteret af CDNL – Committee of Directors of National Libraries – som et URN-baseret system som kunne anvendes af nationale biblioteker.

URN specifikation fastlægger syntaksen i opbygningen af et URN, men tager f.eks. ikke stilling til hvordan de specifikke objektidentifikatorer (NSS) udformes: det bestemmes af NID. Der arbejdes på metoder til f.eks. at anvende ISBN- og ISSN-identifikatorer i et URN.

Relevante URN/NBN aktiviteter:

Die Deutsche Bibliothek (DDB), som siden 1998 har koordineret en central indsamling og formidling af tyske online afhandlinger har, som en del af ”Dissertation Online” projektet [DISSONLINE], etableret et nationalt ”URN management system”, hvor tildeling og vedligeholdelse af permanente identifikatorer er en af hovedingredienserne [DISSPI]. DDB anvender URNs og har siden september 2001 haft ansvar for URNs tildelt under namespace **nbn:de**. Til dato er der registreret ca. 14.000 URNs for afhandlinger på vegne af 29 deltagende biblioteker.

DDB’s ”URN management system” har mange interessante elementer. De deltagende biblioteker har adgang til et websystem til tildeling og vedligeholdelse af URNs for de enkelte ressourcer. Ressourcerne tilmeldes DDB systemet, som sørger for en efterfølgende høstning af ressourcen. Kernen er selvfølgelig selve URN⇒URL *resolving* tjeneste, som giver adgang til ressourcen, men der udføres også regelmæssig linkcheck (daglige) og ”MD5 check sum” kontrol af filerne udføres en gang månedligt.

De enkelte URNs tildeles decentralt af de deltagende universitetsbiblioteker. Selve URN er udformet som et hierarkisk system, hvor det institutionelle tilhørsforhold også er afspejlet:

urn:nbn:de:[designation of library associations]:[official notation of university libraries]-[production number][P]

hvor “production number” er et lokalt tildelt løbenummer og “P” et check-ciffer kalkuleret ud fra hele URN. Check-ciffer kalkuleres og tilføjes URN, når et bibliotek anmelder en ressource i DDBs URN management system. Som det fremgår af ovenstående URN, er der et stærkt ”semantisk” element i udformning af en URN-streng, som kan være problematisk. Vi vender tilbage til dette spørgsmål i afsnittet med anbefalinger.

DDBs system kan anvendes med 1:1 relationer, hvor ét URN peger mod ét URL og 1:n relationer, hvor ét URN peger mod flere URLs.

URN resolver system drives af Bibliotheksservice-Zentrum Baden-Württemberg (BSZ), udviklet under CARMEN-AP4 projekt som afsluttedes i 2002. Systemet videreudvikles i 2002-2004 i EPICUR-projekt. [EPICUR]

Et nordisk URN resolver service projekt har fået støtte fra NORDINFO til etablering af en nordisk URN resolver service. [NORDISKURN] Denne URN resolver service vil anvende NBN namespace, og forventes at blive taget i brug primo 2005. Projektet udføres af de 5 nordiske lande med Uppsala Universitetsbibliotek, Enheden for digital publicering, som projektleder. Der er planer om at anvende open source software som er blevet anvendt af Kungliga Biblioteket i Sverige siden 2002. Projektperioden er 01.01.04 – 31.12.04.

Første del af projektet vil være skabelse af selve URN resolver service. Med tildeling af færre ressourcer end forventet, vil projektet søge andre midler til et senere forløb til f.eks. udvikling og indførelse af metoder til flytning af "arkivpakker" mellem arkiver, anvendelse af METS mm.

Iflg. projektets reviderede tidsplan vil en kravspecifikation være færdig ultimo juli 2004 og implementering af den URN:NBN resolver service påbegyndes primo november 2004.

En arbejdsgruppe under Dansk Standard blev nedsat i april 2003 til at foretage nogle indledende undersøgelser vedr. permanente identifikatorer. I arbejdsgruppens . I arbejdsgruppens rapport [DSPI] fra november 2003 anbefales det (efter en kort gennemgang af de eksisterende modeller), at arbejdet fortsætter med fastlæggelse af den nødvendige infrastruktur for et system baseret på URN navngivningsmodel. Det bemærkes, at arbejdsgruppen ikke kun beskæftigede sig med en infrastruktur under "nbn namespace", men tog et bredere perspektiv. Arbejdsgruppens anbefalinger vedrører også kun URN som navngivningskonvention og ikke systemet som skal bruges til "resolving" af identifikatorer til objekternes lokalisering.

C.4.1.1.2: Handle System – [HANDLE]

Handle-systemet er udviklet af Corporation for National Research Initiatives (CNRI) og bygger på det arbejde, som er blevet udført til at fastlægge den infrastruktur, som er nødvendig for opbygning af digitale biblioteker/samlinger. (Se f.eks. [Kahn]). Handle-systemet er et globalt, distribueret system for navngivning, vedligeholdelse og adressering ("resolving") af identifikatorer for digitale objekter. Disse "identifikatorer" kaldes "handles". [HANDLE]

Systemet er et distribueret system, som består af en Global Handle Registry hvortil er registreret alle Local Handle Services. Identifikatorer tildeles af de registrerede "naming authorities". "Handles" som tildeles ressourcerne, består af et prefix, som er den navngivende myndighed og et suffix, det tildelte unikke navn ("unique local name"). Prefix og suffix separeres af "/". Med Handle-systemet tildeles digitale ressourcer unikke identifikatorer ("handles") som bruges til at lokalisere og få adgang til ressourcen via en "resolverservice".

Udover at være et system for navngivning og lokalisering af ressourcer, er det samtidig et software-system, som frit kan downloades.

Handle-systemet er meget udbredt. Der er tildelt ca.1000 prefixes hvoraf ca.700 anvendes af Digital Object Identifier (DOI) System under International DOI Foundation (IDF) (se et følgende afsnit). CNRI står bag udvikling og vedligeholdelse af Global Handle Registry og registrering er gratis. CNRI forudser, at det i fremtiden sandsynligvis vil blive nødvendigt at indføre betaling for denne registrering af de lokale myndigheder og for vedligeholdelse af systemet.

CNRI tilbyder også en gratis Public Local Handle Service (LHS), hvor alle kan tildele ressource-handles og hvor disse handles kan kobles til et URL. Denne LHS er så ”naming authority” 1030 og brugerne af denne service får tildelt en ”sub-naming authority”, f.eks. 1030.34 (som så er prefix i identifikatoren). Efter man har oprettet sine handles, kan de blive ”resolved” ved at anvende CNRI’s proxyserver eller en plug-in, udviklet af CNRI. (For mere information se http://hdl.handle.net/4263537/4090).

C.4.1.1.3: Digital Object Identifiers – [DOI]

Digital Object Identifiers (DOI) er en specifik implementering af Handle-systemet, og var oprindeligt rettet mod forlagsbranchens bekymringer over den ophavsretslige beskyttelse af digitale ressourcer. DOI blev introduceret i 1997.

Som i Handle-systemet, består et DOI af et prefix, som identificerer den organisation, som er medlem af The International DOI Foundation (IDF) [IDF], efterfulgt af et "/", efterfulgt af et suffix, som er den identifikator der er/har/får ?? tildelt en ressource. Prefixet er den udstedende organisation, og det er denne organisation, som bestemmer, hvordan suffixet (objektets ID) udformes. Det kan f.eks. være et ISBN eller et helt andet identifikationssystem.

Et eksempel: ”10.1045/february2003-featured.collection” er et DOI. ”10.1045” er den registrerede organisation (D-Lib Magazine) og ”february2003-featured.collection” er identifikator for en artikel i februar 2003 nummer med titlen ”Flora and Fauna of the Great Lakes Region”. Ved at anvende IDF ”resolution service” (http://dx.doi.org/10.1045/february2003-featured.collection) får vi adgang til artiklen via URL http://www.dlib.org/dlib/february03/02featured-collection.html.

IDFs ”resolution service” anvender Handle-systemet, udviklet af CNRI.

IDF beskriver DOI systemet som bestående af 4 komponenter [PASKIN]: (1) en nummererings- (identifikations) komponent, som er en implementering af URI/URN, (2) en beskrivelseskomponent af objektet, som er blevet identificeret. Til "beskrivelsen" af objektet bruges metadata baseret på <indecs>-systemet. [INDECS] Indecs (Interoperability of data in e-commerce systems) er metadata til ophavsretslig beskyttelse af digitale objekter. (3) En "resolution" komponent, som giver adgang til det identificerede objekt og, som allerede nævnt, er baseret på Handle-systemet og (4) "policy" komponent, er den organisation som står bag, og som sørger for finansiering og udvikling af DOI systemet.

IDF og National Information Standards Organisation (NISO) har samarbejdet om publicering af en standard (ANSI/NISO Z39.84-2000) for DOI. [DOI/NISO]

Forretningsmodellen bag DOI-systemet er en række ”registration agencies” som betaler et årligt medlemskab af IDF, den non-profit organisation som står bag DOI-systemet. Disse ”registration agencies” kan så udstede objekt identifikatorer, for en pris som de selv fastsætter. For hver DOI som udstedes af en registration agency, betales en afgift til IDF, som er pt. ca. 4 amerikanske cent.

Som beskrevet, udsprang DOI systemet af forlagsbranchen, men modellen i sig selv forudsætter ikke nødvendigvis en kommerciel forbindelse mellem en informationsbruger og en ressource. Medlemmerne af The International DOI Foundation (IDF) repræsenterer mange forskellige typer informationsformidlere bestående af kommercielle virksomheder, biblioteker, organisationer osv.

For en kort gennemgang af en række identifikatorsystemer, henvises til en værdifuld rapport udarbejdet af The Stationers Office (TSO) [TSO]. Rapporten koncentrerer sig om og anbefaler DOI-systemet for anvendelse i den engelske "JISC community " (Joint Information Services Community) og i rapporten er eksempler for anvendelse af DOI i forskellige sektorer såsom forlagsbranchen, fagportaler, universiteter osv. Herunder også eksempler på hvordan DOI's tilknyttede metadatamodel (<indecs> - interoperability of data in ecommerce systems) kan anvendes.

I juni 2004 har TSO meddelt, at de nu vil udstede DOI's gratis, til at fremme anvendelsen, især i det engelske "e-Govenment" regi. (TSOFREE) Finansiering af TSO som "registration agency" vil være i form af en årlig betaling til TSO.

En af de mest velkendte implementeringer af DOI-systemet er CrossRef, en organisation etableret af videnskabelige udgivere, som anvender DOI's til at linke mellem referencer i artikler. Som medlem af CrossRef kan man få sine tidsskriftartikler registreret med en DOI og tilhørende metadata i databasen (som administreres af CNRI). Tilsvarende kan en udgiver forsyne citationer i artikler med *outbound* links, til de referencer som allerede findes i "DOI databasen".

C.4.1.1.4: Persistent Uniform Resource Locator – PURL [PURL]

En Persistent Uniform Resource Locator (PURL) er en "almindelig" URL, som bruger http-protokollens redirection til at slå den rigtige URL op i en PURL registreringsservice. Objektets "rigtige" URL bliver derefter returneret til brugerens browser, som så giver adgang til objektet. En PURL kan kun håndtere 1:1 relationer, én PURL lokaliserer kun ét objekt. PURL systemet blev introduceret af OCLC i 1995, og softwaren kan downloades fra PURL website.

PURL systemet består således kun af en resolver service og ingen konventioner til at skabe standardisering/best practice i navngivning af objekter.

Pr. maj 2004 er der blevet udstedt ca. 600.000 PURLs og resolved ca. 86.000.000 PURLs .

En PURL service er en mellemløsning, dvs. overvejende rettet mod adressering/lokalisering af objekter og uden en stringent objektidentifikation. Som "mellemløsning" er PURL blevet anvendt, mens URN-systemet har været under udvikling. Den danske PURL service (http://purl.dk/) oprettet af DBC, ophørte i starten af 2004.

Der er fornylig blevet fremsat en specifikation for en såkaldt POI – PURL-based Object Identifier - for anvendelse med Open Archives Initiative (OAI) og OAI identifier format i OAI repositorier. [POI] En POI er en URI som anvender OAI identifier som namespace-identifier, f.eks.: *http://purl.org/poi/abcd.dk/doc12345.5* hvor "abcd.dk" er OAI identifier. Der er også et sæt guidelines for hvordan POI's kan blive "resolved". [POIRES] PURL software anbefales til at resolve POIs.

C.4.1.1.5: Archival Ressource Key (ARK) – [ARK]

ARK-systemet er et eksperimentelt system, udviklet ud fra den erkendelse, at adgang til digitalt materiale på længere sigt ikke kun garanteres ved at indføre nogle tekniske

foranstaltninger, som sikrer digitale objekters integritet og ved at sørge for en infrastruktur til navngivning og adressering af objekterne. Der er endnu et led, som er afgørende: det organisatoriske. Kun en institution med den nødvendige og tilstrækkelige organisatoriske infrastruktur, kan i sidste ende sikre adgang over tid.

ARK identifikator er en unik URL til en bestemt ressource. En "ARK URL" linker til 3 ting: (1) metadata tilknyttet et digitalt objekt (2) de(n) fil(er) som udgør objektet og (3) et "commitment statement" fra den institution, som har ansvar for objektet, dvs. den institution som skal sikre adgang til og bevaring af objektet.

ARK syntaksen er:

[http://NMAH/]ark:/NAAN/Name

hvor NMAH er Name Mapping Authority Hostport, NAAN er Name Assigning Authority Number, et unikt nummer for den instans, som har navngivet objektet, med Name. NMAH, en almindelig URL, kan og vil formentlig ændres og hører således ikke til objektets identifikation, kun lokalisering (på et givent tidspunkt). Det som følger efter "ark:" label er objektets permanente identifikator. Når et NMAH for et objekt ændres, skal det aktuelle NMAH findes i NAAN registret (som pt. findes på US National Library of Medicine).

ARK systemet bruger http redirection og derfor kan almindelige browsere bruges. Der er pt. 10 institutioner som er registrerede som Name Assigning Authority.

Som nævnt ovenfor kan en ARK identifikator returnere objektets tilknyttede metadata og en "commitment statement". Det gøres ved at tilføje hhv. et "?" eller to "??" efter Name. Det metadataformat som anvendes nu er et simpelt "label-colon-value"format, Electronic Ressource Citations (ERCs), som tager udgangspunkt i Dublin Core.

ARK systemet er udviklet af US National Library of Medicine og er under afprøvning på University of California Digital Library (http://www.cdlib.org/). "A founding principle of the ARK is that persistence is purely a matter of service, and is neither inherent in an object nor conferred on it by a particular naming syntax. The best an identifier can do is lead users to those services." [KUNZE]

Konklusioner og anbefalinger:

Som et vigtigt led i sikring af adgang til digitale ressourcer og genbrug af ressourcer på tværs af anvendelsesområder, er det afgørende, at der indføres et system, som vil sikre standardiseret navngivning af ressourcer uafhængig af ressourcernes lokalisering, for at opnå vedvarende og global adgang.

Et entydigt navn vil i sig selv ikke give et objekt status som "bevaring sikret for eftertiden" og at separere navnet fra lokalisering vil heller ikke give status som "adgang sikret for eftertiden". Den eneste entitet som kan garantere denne status, er den instans som til enhver tid har erklæret sig som "objektets kustode". Kravene må være, at et objekt kan (1) identificeres entydigt og det ligger i navngiving og metadata, (2) lokaliseres til enhver tid og (3) der findes en bevaringsstrategi for en ressource.

Med baggrund i ovenstående gennemgang af de eksisterende konventioner, systemer og mål, kan der opstilles et sæt anbefalinger, som kan indgå i etablering af et permanent identifikator system.

1. Det er vigtigt, at adskillelse af navn og lokaliseringsmekanisme fastholdes:
Der vil formentlig eksistere flere "konkurrerende" lokaliseringssystemer side om side, så det er vigtigt, at de forskellige objekt-ID'er kan anvendes i de forskellige lokaliseringssystemer, også hvis systemerne skulle skiftes ud, og hvis 2 systemer skulle anvendes parallelt. (Og der vil formentlig på sigt være behov for et globalt resolutionssystem for forskellige navngivningskonventioner og forskellige resolutionssystemer.)

2. URN anbefales som navngivningskonvention:
Det anbefales, at man anvender URN navngivningskonventioner og med brug af nbn namespace og evt. flere sub-namespaces (efter nbn:dk).

Der skal nedfælles regler for citationer af URNs, da et URN i sig selv ikke indeholder lokaliseringsmekanismen og de kan pt. ikke anvendes i browsere uden brug af plug-in.

3. Semantik i objektnavngivning:
Der er en del diskussion (men ingen entydig konsensus) om hvorvidt objektnavne skal være helt uden semantik eller om de til en vis grad kan indeholde et semantisk element. Der synes dog at være konsensus om, at navne bør være forholdsvis neutrale og ikke udelukkende skal bestå af semantiske elementer i form af f.eks. flere institutionsnavne eller en hierarkisk opbygning (f.eks. geografisk eller institutionelle osv.). Det er en fordel, at de er genkendelige og kan "læses" af mennesker. At anvende en ressources automatisk genererede MD5 checksum som ID i et URN (f.eks. urn:dbb-2c3ee69s23ecf567ec7yt5re33ace234), som anvendes f.eks. i Finland, synes at være overdrevet neutralt. (Men det kan være fornuftigt at anvende MD5 checksum i andre sammenhænge). Herudover kan der være en fordel i, at en identifikator indeholder mulighed for branding.

Vi foreslår, at man anvender en biblioteks- eller en universitetsforkortelse som første led i objekt ID efterfulgt af et neutralt løbenummer (som f.eks. kan bestå af et årstal efterfulgt af et løbenummer). Biblioteks- eller universitetsforkortelserne (som selvfølgelig skal tildeles entydigt) er knyttet til institutioner, som kan være flygtige og ikke altid entydige, men her henviser de ikke til en aktuel lokalisering, ansvar eller permanens men til objektets oprindelse. Herudover er de genkendelige, har en mnemonisk karakter og indeholder et element af branding.

En anden mulighed er anvendelsen af International Standard Identifier for Libraries and Related Organizations (ISIL) identifier [ISIL][11]. ISIL identifier er alfanumerisk og består af en landekode (specificeret i ISO-3166-1) som præfiks og en "library identifier" som suffiks. Mellem præfiks og suffiks er et "-". ISIL-koden er på maks. 16 tegn og kan variere i længde.

Et automatisk genereret check-ciffer som en del af objekt-ID (som i det tyske DDB projekt) kan anbefales. Og ligesom det tyske DDB system, kunne MD5 checksum og check-ciffer tildeles i et URN management system, når man lokalt registrerer sin ressource ID.

Inden man anbefaler standardisering af navngivning af objekter, skal man undersøge, om en standardisering er mulig på tværs af lokale systemer.

[11] Anvendelse af ISIL er under diskussion af Dublin Core Collection Description Working Group

I et evt. kommende follow-up projekt vedr. PI'er kan der fremsættes regler for navngivning. Guidelines for navngivning vil også være at foretrække frem for frit slag.

4. Ressourcerne skal navngives lokalt:
ID'er skal tildeles ressourcerne lokalt iht. de krav eller guidelines, som udformes.

5. Relationer til andre instanser/sektorer:
Da det ikke kun er biblioteker, som har interesse i at etablere en PI-service, er det vigtigt, at biblioteker indgår i et samarbejde med andre interessenter om etablering af en URN-resolver service og om udformning af URN navngivningskonventioner, dvs. systematisering af de relevante namespaces og objekt-ID'er. På denne måde vil man sikre en "organisatorisk stabilitet", som er afgørende.

6. Service versus "bare" resolution system:
Permanens (adgang til digitale ressourcer over tid) er ikke kun et spørgsmål om teknik men i afgørende grad et spørgsmål om institutioner. (Gen)anvendelse af en voksende mængde digitale ressourcer på tværs af sektorer, er ikke bare et spørgsmål om at lokalisere men også om at kunne identificere og vælge. Permanens og effektiv anvendelse er begge afhængige af robuste services – ikke kun systemer. Det vil formentlig være afgørende, at system(er) indeholder services, som fremmer andet end "bare" lokalisering. Disse services kunne f.eks. være:

- objektets metadata gøres tilgængelige mhp. genfinding og udvælgelse.
- én ID kan adressere til flere ressourcer (f.eks. flere udgaver af samme værk).
- der knyttes MD5 check-sum til ressourcerne.
- der knyttes et check-ciffer som en del af identifikatorer.
- der foretages logning af benyttelse mhp. statistik.

Hvis man opererer med et system bestående af centrale og decentrale arkiver, f.eks. ifm. pligtaflevering til nationalbiblioteker fra universitetsbiblioteker, vil flg. services være af værdi:

- regelmæssigt link check af lokale ressourcer.
- et system for lokal management af tildelte identifikatorer.
- automatisk høstning af ressourcer ifm. f.eks. pligtaflevering.

7. Valg af resolution service:
Groft sagt er der i dag 3 muligheder for etablering af en resolution service: 1. Man anvender eller udvikler egen resolution software, baseret på urn (som f.eks. i det fælles nordiske NORDINFO projekt). 2. Der etableres en DOI registration service i Danmark og 3. Man anvender PURL software (centralt eller decentralt).

Afgørende er, at det system man vælger, kan tilbyde de nødvendige services, som påpeget i pkt. 6. Det kan kun (1) en egen udviklet resolution service (f.eks. NORDINFO projekt, hvis projektet har den ambition), og (2) DOI, som er udbredt og har den nødvendige organisation bag. DOI anvendes i mange forskellige sektorer: biblioteker, central administration, erhvervsvirksomheder osv. Og da permanente identifikatorer har interesse i forskellige sektorer, kunne det være en fordel at etablere et lokaliseringssystem, som i sig selv ikke var tilknyttet én bestemt sektor, f.eks. det Handle-baserede DOI system.

DOI er samtidig ikke bare en resolution service, og der er f.eks. et velfunderet metadata-system tilknyttet. Systemet er også robust i form af sin internationalt udbredte anvendelse.

Systemet forudsætter oprettelse af en lokal ”registration agency”, som vil have økonomiske forpligtelser til IDF. Dette vil selvfølgelig medføre en betaling for de udstedte DOI’s – et eller andet sted i systemet.

Da overdrivelse ind imellem kan understrege en påstand, kan en service, som udelukkende består af resolution, etableres ved anvendelse af PURL software, som vil være en billig løsning!

8. Bevaring af digitale ressourcer:
Det som giver persistent identifier mening, er ikke kun et spørgsmål om teknik eller overholdelse af internationale navngivningskonventioner og -standarder. Et afgørende aspekt er, at teknikken og standarder varetages og udføres i en organisatorisk infrastruktur, og som dermed bliver en vigtig del af stabiliteten.

I forbindelse med kommende aktiviteter vedr. bevaring, kan det anbefales, at man undersøger hvordan bevaringsstrategier, erklæringer eller garantier kan tilknyttes et objekts metadata, herunder også hvordan f.eks. ”mekaniske” MD5 checksum eller lign. metoder kan gøres eksplicit for brugerne.

C.4.2: Netpublikationer:

I ”pre-ingest” fasen, hvor et digitalt objekt produceres, vil der være et antal formater som anvendes af producenterne. I ”ingest” fasen vil disse formater indlemmes i repositoriet, enten som de er, eller de vil gennemgå en konvertering til format(er), som kan accepteres af repositoriet. F.eks. kan et e-print skrives i MS Word, blive omdannet til PDF-format (enten af forfatteren selv eller arkivet), og blive overført til arkivet.

I formidlingsfasen kan objekterne tilgængeliggøres i forskellige formater. Som regel er det et praktisk og ”politisk” spørgsmål, hvilke(t) format(er) et repositorie vil anvende eller tillade. (”Politisk” i denne sammenhæng er f.eks. hensyntagen til faktorer som knytter sig til formidling, tilgængelighed, usability osv.) . Drejer det sig om e-prints, er mulige formater f.eks. et rent tekstformat (.txt), rich text format (.rtf), MS Word filer og filer i PDF-format eller HTML format.

Om et bestemt fil-format kan anvendes af brugeren er afhængig af det computerudstyr brugeren har, som har indflydelse på repositoriets valg af foretrukne formater. Visse formater kræver ekstra software (f.eks. som ”plug-ins”), og det er som regel hvor ekstra software kan betragtes som standardkomponent, at et format kan anbefales. F.eks. kan PDF-format i dag betragtes som et (næsten) standard format og accepteres af de fleste repositorier, hvorimod anvendelse af f.eks. SVG (Scaleable Vector Graphics) endnu giver problemer for formidling.

C.4.2.1: NetPub:

Netpub er et omfattende webbaseret publiceringsmiljø udviklet af firma Valusoft i samarbejde med IT- & Telestyrelsen til konvertering af Word eller PDF-- filer til HTML-dokumenter. Hele processen, fra upload af Word/PDF-fil det færdige sæt HTML-dokumenter, pligtaflevering og fakturering, foregår i en browser (kun IE). NetPub er beregnet til produktion og publicering af HTML-dokumenter, som overholder standarder for tilgængelighed og metadata opmærkning.

Kort fortalt er fremgangsmåden flg.: Dokumentet uploades til Valuesofts hjemmeside via browseren og konverteres til HTML. Konverteringsprocessen danner grundlag for ”første udgave” af det færdige HTML-dokument, hvor NetPub programmet ud fra Word eller PDF-fil, har ”gættet” sig frem til de elementer, som udgør dokumentet. Herefter bliver brugeren trinvis guided gennem processen med at fastlægge de elementer, som udgør det færdige dokument ud fra ”førsteudgaven”. Der skal rettes/indtastes metadata (f.eks. dokumentets titel, abstract, URL) og brugeren skal fastlægge dokumentets forskellige formateringselementer, som programmet har fundet, f.eks. kapitler, afsnit, overskrifter, billedtekster osv. Disse formateringselementer bruges til at danne en indholdsfortegnelse. Herefter vises hele publikationen i en avanceret editor, så brugeren kan færdigredigere dokumentet.

I redigeringsprocessen er der mulighed for at sætte eksterne link i dokumentet, og billederne kan redigeres og skæres. Af hensyn til tilgængelighed skal hvert billede forsynes med en alternativ tekst, og der kan vælges og redigeres i forskellige designskabeloner til det færdige dokument. I konverteringsprocessen har programmet forsøgt at fastlægge elementer i det færdige dokument, f.eks. billedtekster, overskrifter, fodnoter osv., som brugeren bagefter kan rette/ændre i. Der er også mulighed for at inkludere en søgefunktion til dokumentet som en del af dokumentets layoutskabelon.

Når dokumentet er blevet checket for fejl og efterfølgende godkendt, udfærdiger Netpub-programmet en faktura baseret på antal formaterede sider, og faktura fremsendes. Det færdige dokument zippes som en pakke med egen fil- og mappestruktur, med stylesheet, billedfiler, indholdsfortegnelse og kapitler og med Dublin Core metadata indlejret i HTML-siderne, som downloades og placeres på brugerens server. Brugeren kan også vælge at pligtaflevere dokumentet i samme arbejdsgang.

Der er mulighed for at tegne et års licens med 4 logins for 18.000 DKR. Derudover skal der betales 20 DKR pr. side for konvertering til HTML-format.

C.4.2.2: Tilgængelighed:

På dansk og internationalt plan er der anbefalinger vedr. adgang til offentlige websteder. Disse anbefalinger udarbejdes først og fremmest for at sikre, at handikappede borgere tilbydes lige adgang til den information, som tilbydes ikke-handikappede borgere. Det kan måske hævdes, at implementering af disse anbefalinger også har en normativ effekt på udformning af websider generelt og derfor forbedres tilgængeligheden for alle. I Danmark har den årlige ”Bedst på Nettet”- vurdering [BPN] af offentlige websteder betydet en øget opmærksomhed på brugernes anvendelse af websteder, ikke kun aht. tilgængelighed men anvendelse af websteder som helhed. Bedst på Nettet har systematiseret vurdering af websteder og med fokus på tilgængelighed alene findes der initiativer i Danmark og internationalt.

I Danmark overvåges området af IT og Telestyrelsen gennem Det Koordinerende Informations Udvalg (KIU) [KIU], som anbefaler standarder for hele det offentlige område, og en følgegruppe bestående af eksperter og repræsentanter for handikaporganisationer, har udarbejdet de konkrete anbefalinger vedr. tilgængelighed.

Internationalt foregår standardiseringsarbejde mht. tilgængelighed i World Wide Web Consortiums’ (W3C) [W3C] Web Accessibility Initiativ’s (WAI) [WAI] regi, og som også danner grundlaget for en stor del af den danske dokumentation på området. Hvordan man sikrer, at ens websider er tilgængelige, er et meget veldokumenteret område, og der findes en række konkrete standarder, checklists og værktøjer, som kan anvendes til formålet.

På dansk samles information på webstedet http://www.netsteder.dk/, hvor der er gode råd og links til relevante websteder, både i Danmark og internationalt, som omhandler tilgængelighed. IT og Telestyrelsen har udgivet en vejledning til hvordan man kan sikre, at ens websted er tilgængelig:

"Disse retningslinier stiller det krav, at hjemmesider og netsteder, der tilbydes af statslige informationsudbydere også skal være tilgængelige for brugere med funktionsnedsættelser. Retningslinierne beskriver hvorfor der skal gøres en ekstra indsats for at gøre netsteder tilgængelige for brugere med funktionsnedsættelser. Retningslinierne giver råd om hvordan netsteder kan gøres tilgængelige. Retningslinierne anviser hvorledes netsteder kan testes med henblik på at gøre netsteder tilgængelige." [IT&TNETPUB]

I WAI regi udarbejder Web Content Accessibility Guidelines Working Group (WCAG WG) anbefalinger vedr. tilgængelighed. Version 1.0 af Web Content Accessibility Guidelines, som har status som "W3C Recommendation", udkom i maj 1999[12]. [WCAG1.0] WCAG version 1 består af 14 "guidelines", og inden for hver guideline er et antal "checkpoints", som er konkrete anbefalinger, som vedrører den enkelte guideline, som burde implementeres i websider. F.eks.: "Guideline 1. Provide equivalent alternatives to auditory and visual content", består af 5 forskellige checkpoints, som man skal implementere i websiderne, hvor indholdet er baseret på lyd og grafiske elementer, f.eks. placering af alternative tekstelementer i siderne.

Hver af disse "checkpoints" i de 14 Guidelines er klassificeret i 3 "Priorities", som udgør "certificerings" aspektet af WCAG. "Priority 1" checkpoints **skal** opfyldes, da de er nødvendige for, at bestemte grupper kan anvende en webside. "Priority 2" checkpoints **burde** opfyldes, da bestemte grupper vil finde det svært at anvende en webside og "Priority 3" checkpoints **kan** opfyldes, dvs. adgang vil forbedres, hvis de implementeres. Der findes en checkliste over de "checkpoints", som udgør de 3 "priorities". [WCAGCHECK]

WCAG WG har specificeret 3 "conformance levels", som afhænger af hvilke "priority checkpoints" en webside opfylder. Niveau "A" opfylder alle "priority" 1 checkpoints, niveau "Double-A" (AA) opfylder checkpoints i "priority" 1 og 2 og niveau "Triple-A" (AAA) opfylder alle "priority" 1, 2 og 3 checkpoints.

Udover IT & Telestyrelsen findes Dansk Center for Tilgængelighed (DCFT), (en selvejende institution i tilknytning til Erhvervs- og Boligstyrelsen) [DCFT], som ressourcecenter inden for tilgængelighed og websteder (og andre områder). DCFT driver bl.a. en hotline, hvor man kan henvende sig vedr. enkelte problemstillinger og tilgængelighed, og har udarbejdet et dokument vedr. tilgængelighed, som kan anvendes som en del af en kravspecifikation til udvikling af et websted. [DCFTKRAV]

Der findes online systemer for testning af en websides tilgængelighed ift. de internationale standarder. [BOBBY]

Tilgængelighed og PDF-dokumenter:

PDF-formatets enorme udbredelse som nærmest standard publiceringsformat, kan give problemer for tilgængelighed, da skærmlæsere har problemer med at anvende PDF

[12] Web Content Accessibility Guidelines version 2.0 udkom som W3C draft i marts 2004. Indholdsmæssig forskel i de forskellige guidelines mellem version 1.0 og 2.0 kan ses her: http://www.w3.org/WAI/GL/2004/03/11-mapping.html

dokumenter korrekt. Der er problemer med navigering, tekstsekvensen og manglende alternative tekster til illustrationer. [DCFTPDF] Adobe (firmaet bag Acrobat program til fremstilling af PDF-dokumenter) har fra version 5.0 af programmet gjort det muligt at tilføje tags i de oprindelige dokumenter, hvis de er skrevet i Microsoft Word. Disse tags kan så benyttes af skærmlæsere til at anvende PDF-dokumenter. Ikke alle problemer er løst med disse tags, og der vil selvfølgelig stadig være problemer med ældre PDF-dokumenter og grafiske PDF-dokumenter, som ikke kan læses af skærmlæsere. (Adobe har en løsning til grafiske PDF-dokumenter, som hedder ”Paper Capture”, som er tekstscanning, hvor dokumentets visuelle indtryk bevares). Der findes en kort artikel om problemer med tilgængelighed af PDF-dokumenter på Dansk Center for Tilgængeligheds website. [DCFTPDF]

Referencer:

Standarder og kerne referencer:

[ARK]: *Archival Ressource Key, http://www.cdlib.org/inside/diglib/ark/*, University of California Digital Library

[BOBBY]: *Bobby Worldwide*, http://bobby.watchfire.com

[BPN]: *Bedst på Nettet*, http://www.bedstpaanettet.dk/

[CAPLAN]: *Priscilla Caplan, Metadata fundamentals for all librarians*, American Library Association, 2003

[CEDARS]: *CEDARS - curl exemplars in digital archives*, http://www.leeds.ac.uk/cedars/ (CURL: Consortium of University Research Libraries)

[CEDARSRAP]: *The Cedars Project Report,* http://www.leeds.ac.uk/cedars/admin/CedarsProjectReportToMar01.pdf, 2001

[CEDARSMETA]: *Metadata for digital preservation,* http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html, 2000

[DCFT]: *Dansk Center for Tilgængelighed*, http://www.dcft.dk/

[DCFTKRAV]: *Bilag til kravspecifikation - tilgængelighed til hjemmesider:* http://www.dcft.dk/index.asp?pid=2830

[DCFTPDF]: *PDF og tilgængelighed*, http://www.dcft.dk/index.asp?pid=1510

[DCMI]: *Dublin Core Metadata Initiative Metadata Terms,* http://dublincore.org/documents/dcmi-terms/

[DISSFORMAT]: *Präferenzregelung für die Archivierung elektronischer Publikationen*, http://www.ddb.de/wir/pdf/praeferenzregelung.pdf, 2000

[DISSONLINE]: *Digitale Dissertationen im Internet*, http://www.dissonline.de/

[DISSPI]: *Persistent identifier*, http://www.persistent-identifier.de/, Die Deutsche Bibliothek

[DOI]: *Digital Object Identifiers*, http://www.doi.org/, The International DOI Foundation (IDF).

[DOI/NISO]: Syntax for the Digital Object Identifier, http://www.niso.org/standards/resources/Z39-84-2000.pdf, 2000

[DSPI]: *Identifikatorer til digitale objekter (Bilag til XML-komit møde, s. 3–11),* http://www.oio.dk/files/Materiale._2._sending_til_modet_den_11._august_04.pdf

[EPICUR]: *Enhancement of Persistent Identifier Services - Comprehensive Method for unequivocal Resource Identification (EPICUR*), (urn:nbn:de:1111-2003121811), http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:1111-2003121811

[HANDLE]: *Handle System*, http://www.handle.net/index.html, Corporation for National Research Initiatives

[IANA]: *Internet Assigned Numbers Auhtority*, http://www.iana.org/

[IDF]: *The International DOI Foundation*, http://www.doi.org/

[INDECS]: *Interoperability of data in e-commerce systems*, http://www.indecs.org/

[ISIL]: *International Standard Identifier for Libraries and Related Organizations (ISIL)*, http://www.bs.dk/isil/

[IT&TNETPUB]: *Hjemmesiders tilgængelighed: Statens retninglinier for offentlige hjemmesiders og netsteders tilgængelighed*, http://www.netsteder.dk/publ/tilgaeng/index.html

[IT&TREF]: *Referenceprofilen*, http://www.oio.dk/referenceprofilen, IT & Telestyrelsen

[JISC]: *Feasibility and Requirements Study on Preservation of E-prints: Report Commissioned by the Joint Information Systems Committee (JISC)*, http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf, 2003

[KAHN]: *Kahn Robert, Wilensky, Robert: A Framework for Distributed Digital Object Services*, http://hdl.handle.net/4263537/5001

[KIU]: *Det Koordinerende Informations Udvalg*, http://www.oio.dk/KIU

[KUNZE]: *Towards Electronic Persistence Using ARK Identifiers*, http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf, California Digital Library, University of California

[METS]: *METS*, http://www.loc.gov/standards/mets/

[METSOV]: *METS: An Overview & Tutorial*, http://www.loc.gov/standards/mets/METSOverview.v2.html

[METSREG]: *METS Implementation Registry*, http://sunsite.berkeley.edu/mets/registry/

[NEDLIBMETA]: *Metadata for long term preservation*, http://www.kb.nl/coop/nedlib/results/preservationmetadata.pdf, 2000

[NLA-PI]: *Persistent Identification Systems*, http://www.nla.gov.au/initiatives/persistence/PIcontents.html, 2001

[NLAMETA]: *Preservation Metadata for Digital Collections*, http://www.nla.gov.au/preserve/pmeta.html,1999

[NLNZ]: *Metadata Standards Framework – Metadata Implementation Scheme*, http://www.natlib.govt.nz/files/4initiatives_metaschema_revised.pdf, 2003

[NORDISKURN]: *Garanterad dokumentåtkomst nu och i framtiden*, http://epc.ub.uu.se/niwiki/pmwiki.php

[OAI]: *Open Archives Initiative*, http://www.openarchives.org/
[OAIS]: *Reference model for an Open Archival Informatiuon System (OAIS)*, http://www.ccsds.org/documents/650x0b1.pdf, Consultative Committee for Space Data Systems (CCSDS), 2002

[OCLC/RLG]: *Preserving metadata and the OAIS information model: a metadata framework to support the preservation of digital objects*, http://www.oclc.org/research/pmwg/, OCLC/RLG Working Group on Preservation Metadata, 2002

[OCLC/RLGWP]: *Preservation Metadata for Digital Objects: A Review of the State of the Art*, http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf, OCLC/RLG Working Group on Preservation Metadata, 2001

[OCLCMETA]: Preservation Metadata Element set – definitions and examples, http://www.oclc.org/digitalpreservation/archiving/metadataset.pdf, 2001

[PASKIN]: Norman Paskin, *DOI: A 2003 Progress Report*, doi:http://dx.doi.org/10.1045/june2003-paskin, International DOI Foundation, 2003.

[PDF/AISO]: *Document management – Electronic document file format for long-term preservation*, http://www.aiim.org/documents/standards/ISO_19005-1_(E).doc

[PDF/ARLG]: William G. LeFurgy , *PDF/A: Developing a File Format for Long-Term Preservation*, i: RLG DigiNews, December 15, 2003, Volume 7, Number 6, http://www.rlg.org/preserv/diginews/v7_n6_feature1.html#congress

[PDFPRES]: John Mark Ockerbloom, *Archiving and Preserving PDF Files*, i: RLG DigiNews, February 15, 2001, Volume 5, Number 1, http://www.rlg.org/preserv/diginews/diginews5-1.html#feature2

[POI]: *The PURL-based Object Identifier (POI)*, http://www.ukoln.ac.uk/distributed-systems/poi/, 2004

[POIRES]: *POI resolver guidelines*, http://www.ukoln.ac.uk/distributed-systems/poi/resolver-guidelines/, 2004

[PREMIS]: PREMIS: PREservation Metadata: implementation strategies, http://www.oclc.org/research/projects/pmwg/

[PURL]: *Persistent Uniform Resource Locator*, http://www.purl.org/

[SOAP]: *Simple Object Access Protocol*, http:///www.w3.org/TR/SOAP

[TDR]: *Trusted digital repositories: attributes and responsibilities:* http://www.rlg.org/longterm/repositories.pdf, 2002

[TSO]: *Digital Object Identifiers for publishers and the e-learning Community, A Report for the JISC from TSO*, http://www.jisc.ac.uk/index.cfm?name=project_tso, 2003

[TSOFREE]: *Free digital object identifiers pave way to linking public information*, http://www.tsoid.com/downloads/DOI%20free%20V.4.2%20Final.pdf, 2004

[URI]: *Uniform Resource Identifiers (URI)*, http://www.ietf.org/rfc/rfc2396.txt

[URN]:
Functional Requirements for Uniform Resource Names, http://www.ietf.org/rfc/rfc1737.txt
URN Syntax ,http://www.ietf.org/rfc/rfc2141.txt
URN Namespace Definition Mechanisms, http://www.ietf.org/rfc/rfc2611.txt

[W3C]: *World Wide Web Consortium*, http://www.w3c.org/

[WAI]: *Web Accessibility Initiative*, http://www.w3.org/WAI/

[WCAG1.0]: Web *Content Accessibilty Guidelines 1.0*, http://www.w3.org/TR/WCAG10/

[WCAGCHECK]: *Checklist of Checkpoints for Web Content Accessibility Guidelines 1.0*, http://www.w3.org/TR/WCAG10/full-checklist.html
(Findes på dansk: http://www.sensus.dk/full-checklistdk.html)

[XMP]: *Extensible Metadata Platform (XMP)*,
http://partners.adobe.com/asn/tech/xmp/pdf/xmpspecification.pdf

[XMPADOBE]: *Extensible Metadata Platform (XMP)*:
http://www.adobe.com/products/xmp/main.html