Linked open data

Fra 'cataloguing' til 'catalinking'

af CARSTEN H. ANDERSEN
DIREKTØR
Datadivisionen
/ DBC

Den traditionelle bibliografiske post er i opbrud. De enkelte dataelementer, og i særlig grad autoritetsdata, kommer i fokus, og linkes sammen til virtuelle poster tilpasset brugssituationen. Et skifte til nye internationale standarder, der understøtter linked data, forbedrer bibliografiske datas synlighed og anvendelighed på nettet.

En brønd er intet uden vand. Databrønden er intet uden metadata. To væsentlige udviklinger kommer til at præge de metadata, vi møder i databrønden.

Den ene er den internationalisering, som i disse år præger den bibliografiske udvikling. Et lille eksempel: Da det for år tilbage blev besluttet at opretholde et særligt dansk MARC-format i form af danMARC2, blev det samtidig besluttet, at fremtidige ændringer i det nationale format skal være i overensstemmelse med MARC21. Internationaliseringen forstærkes netop nu af et sæt nye katalogiseringsregler, RDA (Ressource Description and Access), der er under implementering af en række lande, og af udviklingen af en afløser for MARC-formatet under betegnelsen Bibliographic Framework Transition Initiative (BIBFRAME). Uden at foruddiskontere en beslutning om, hvorvidt vi i Danmark skal indføre RDA og/eller BIBFRAME er der ingen tvivl om, at databrønden som en infrastruktur på såvel lokalt som nationalt niveau kommer til at udvikle sig i overensstemmelse med de principper, der ligger bag den internationale bibliografiske udvikling.

Den anden udvikling bunder i den kendsgerning, at biblioteksbrugerne i stigende grad søger information i generelle websøgemaskiner fremfor i bibliotekskatalogen. Derfor skal bibliotekernes metadata integreres i webben og optimeres i forhold til størst mulig synlighed i søgemaskinerne. Databrønden skal stadig være det metadatamæssige omdrejningspunkt, men skal også kunne levere metadata ud på webben.

Et af de væsentligste principper i såvel udviklingen på webben generelt som i den internationale bibliografiske udvikling er linked data. I denne artikel belyses, hvad linked data er, hvordan linked data kommer til udtryk i den bibliografiske verden, og hvilke perspektiver linked data giver for bibliotekerne.

Linked data: fra et net af dokumenter til et net af data

World wide web er i sin nuværende form i udpræget grad rettet mod menneske til maskine-interaktion. Vi arbejder i websider, som i det store hele alene er opmærket med henblik på, at browseren kan layoute siden pænt. Når vi klikker på et link, er det vores tolkning af den sammenhæng, som linket indgår i, der skaber vores forventning til, hvad vi bliver præsenteret for, når linket aktiveres. Der er ingen information knyttet til linket, der muliggør, at en applikation kan foretage en semantisk afkodning af, hvad linket fører til. Man kan tale om 'a web of documents'.

Tim Berners-Lee formulerede visionen om semantisk web, "a web of data that can be processed directly and indirectly by machines". Dvs. et world wide web, hvor websider er opbrudt i entydigt identificerbare og relaterede dataelementer med tilknyttet semantik, og hvor det er applikationens opgave, ud fra semantikken og relationerne sammenholdt med brugerens kontekst, at sammensætte og præsentere relevante data for brugeren. Alt sammen med brug af gængse www-standarder. En linked data-baseret applikation vil kunne afgrænse og samle information forskellige steder på nettet om den i brugerkonteksten rette betydning af begrebet Venus (Er brugeren interesseret i tennisspilleren, gudinden, Venus fra Milo, eller…?).

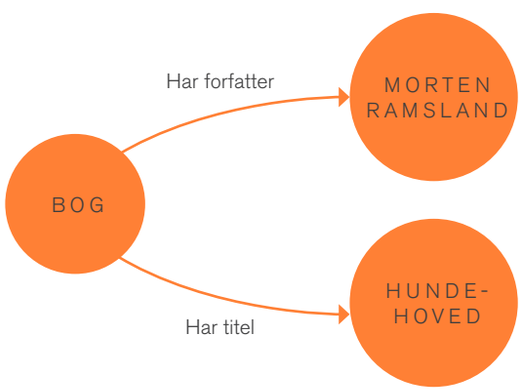
Berners-Lee formulerede fire principper for linked data som en vej til at realisere det semantiske web – her lettere bearbejdet:

1. Brug en URI (Uniform Resource Identifier) som entydig identifikator i stedet for f.eks. forfatterens navn
2. Brug URL'er til de entydige identifikatorer, så de kan linkes til og opløses, f.eks til forfatternavnet
3. Tilbyd yderligere information om f.eks. forfatteren i et standardiseret format, når URI-linket forfølges
4. Suppler informationen med links til relaterede URI'er, f.eks. en beskrivelse af forfatterskabet.

Pointen er, at der anvendes helt gængse www-standarder som HTTP-protokollen og RDF (Resource Description Framework) til at beskrive og koble dataelementer sammen på tværs af webben.

Grundlæggende er der tale om linked data, men oftest taler vi om linked open data fordi det ligger implicit i tænkningen om linked data, at data, der publiceres på denne måde, bør være frit tilgængelige.

I linked data er information udtrykt i RDF og således atomiseret i såkaldte tripler. Navnet skyldes, at en triple består af tre elementer: subjekt – prædikat – objekt, f.eks. Bog – HarTitel – Hundehoved. Prædikater kan også ses som relationer. Subjekter og objekter kan indgå i flere tripler. Tripler beskrives typisk med denne grafiske repræsentation (for forståelsens skyld er ikke anvendt URI'er for subjekt, attribut og objekt):



Når en bibliografisk post publiceres som linked data, vil det typisk være som en mængde tripler. Et eksempel på, hvordan disse tripler kan se ud, ses her:

```
dbcResource:r25741420 a dct:BibliographicResource.
dbcResource:r25741420 a bibo:Book.
dbcResource:r25741420 dct:title "Hundehoved".
dbcResource:r25741420 dct:creator dbcPeople:pA2067402x.
dbcResource:r25741420 bibo:isbn10 "87-621-0454-3".
dbcResource:r25741420 dct:language lexvo:dan.
marcLangCode:dan foaf:focus lexvo:dan.
dbcPeople:pA2067402x foaf:givenName "Morten".
dbcPeople:pA2067402x foaf:familyName "Ramsland".
```

En applikations korrekte forståelse af triplerne opnås ved at anvende veldefinerede ontologier, vokabularer mv. til regulering og beskrivelse af anvendte subjekter, prædikater og objekter. Disse ontologier og vokabularer er udtrykt i RDF-skemaer på nettet, som applikationen kan slå op i. I ovenstående eksempel ses f.eks. anvendt 'dct:title' fra Dublin Core til angivelse af, at 'Hundehoved' skal opfattes som en titel. Da f.eks. forskellige fagdomæner anvender forskellige ontologier, er mapning mellem forskellige ontologier væsentligt for at kunne linke data på tværs af webben.

De tripler, som beskriver et materiale, kan indeholde interne links til andre dataelementer i ens eget datasæt, f.eks. andre dataelementer inden for Nationalbibliografien, eller eksterne links til dataelementer i andre datasæt helt andre steder på webben, f.eks. DBpedia (linked data-versionen af Wikipedia). Der er i princippet ingen forskel på interne og eksterne links, og derved opnås lige præcis et 'web of data', hvor metadata til beskrivelse af et materiale kan sammenstykkes af dataelementer på tværs af webben.

Linked data og bibliografisk udvikling

Biblioteksverdenens tilgang til metadata har været – og er fremdeles – præget af dels ønsket om at beskrive materialer i en samling ud fra kurateringsmæssige behov snarere end slutbrugerbehov, dels af kartotekskortets rammesætning for fortløbende sammenstilling og rækkefølge af dataelementer, jf. ISBD.

Disse karakteristika blev ved overgangen til digital lagring af metadata i vid udstrækning videreført, nu blot i poster. I en linked data-kontekst svarer poster til 'a web of documents'. For at komme over i 'a web of data' fordres, at posterne atomiseres i relaterede dataelementer. Og det er præcis, hvad der er ved at ske i den internationale bibliografiske udvikling.

IFLAs Functional Requirements for Bibliographic Records (FRBR) fra 1998 indvarslede et opgør med ovennævnte karakteristika, om end på et konceptuelt niveau. Med afsæt i definerede (slut)brugerbehov opererer FRBR med entiteter og relationer og organiserer sammenhørende data på forskellige niveauer i form af work, expression, manifestation og item.

Der har efterfølgende været arbejdet på at realisere FRBR eller variationer heraf i bibliotekskataloger, bl.a. i bibliotek.dk's værkvisning og i WorldCat. Men fælles for disse løsninger er, at metadata ikke er skabt og initialt organiseret ud fra FRBR, og FRBR-strukturen er opnået ved en efterfølgende bedst mulig bearbejdning og sammenstilling af data i forbindelse med præsentationen af søgeresultater.

"I en linked data-kontekst svarer poster til 'a web of documents'. For at komme over i 'a web of data' fordres, at posterne atomiseres i relaterede dataelementer. Og det er præcis, hvad der er ved at ske i den internationale bibliografiske udvikling."

"Autoritetsdata bliver et helt centralt omdrejningspunkt for linkning mellem forskellige datasæt på tværs af webben – netop på grund af den entydige identifikation."

// RDA - Ressource Description and Access

De nye katalogiseringsregler, RDA, udarbejdet af Library of Congress, Libraries and Archives Canada, British Library og National Library of Australia, afspejler for alvor FRBR og har hermed en entitets- og relationsorienteret tilgang. RDA er indført eller under indførelse af nationalbiblioteker mv. i bl.a. USA, Canada, England, Australien, Tyskland, Holland, Finland og Schweiz. I Danmark pågår konsekvensanalyse og udarbejdelse af beslutningsgrundlag for overgang til RDA i regi af Bibliografisk Råd.

// BIBFRAME

Det har vist sig i forbindelse med overgangen til RDA, at MARC-formatet ikke i tilstrækkelig grad understøtter den entitets- og relationsorienterede tilgang til strukturering af metadata. Det har affødt BIBFRAME-initiativet, der har til formål at finde en mere hensigtsmæssig måde at strukturere metadata på, end MARC tilbyder. BIBFRAME er tænkt til at favne både katalogisering og udveksling af metadata og omfatter såvel en konceptuel datamodel som konkret udtrykkelse/formatering af datamodellen.

BIBFRAME er helt grundlæggende baseret på linked data-principper. Dermed kan der ved overgang til BIBFRAME for alvor tales om en atomisering af posten eller 'a web of data'. Med BIBFRAME nærmer bibliotekerne sig museerne, som i længere tid har haft en linked data-baseret tilgang til strukturering af metadata, bl.a. udtrykt i Europeanas datamodel (EDM).

BIBFRAME's datamodel adskiller sig fra FRBR's datamodel, bl.a med et andet antal niveauer. Set i et rent biblioteksperspektiv er det ærgerligt og besværliggør endnu engang livet for implementører med hensyn til modelvalg og mapning mellem modeller. BIBFRAME er imidlertid tænkt bredere anvendt end på biblioteksområdet alene.

// Autoritesdata bliver omdrejningspunkt

I fremtiden må vi altså forestille os en bevægelse fra den veldefinerede bibliografiske post i retning af enkeltstående dataelementer – såvel egne dataelementer som dataelementer hentet på nettet – der linkes sammen til at udgøre en virtuel post. Det rejser spørgsmålet om, hvad der tilsammen udgør en sådan virtuel post. Det vil der næppe være ét svar på, derimod vil en virtuel post blive sammenstillet til det specifikke formål, f.eks. visning i lokal kontekst, visning i national kontekst, eksport osv.

Væsentlige elementer i metadata til et materiale eller en ressource er beskrivelsen af ting, steder og mennesker – det være sig som forfatter, emneord, udgivelsessted mv. Der er i biblioteksverdenen tradition for at arbejde med kontrollerede data eller autoritetsdata, dvs. entydigt identificerbare og fast udformede data som f.eks. personnavne og emneord. Disse autoritetsdata har hidtil været et anliggende alene for biblioteksverdenen. Med bibliotekernes metadata struktureret som linked data åbner sig nye perspektiver for deling, genbrug og berigelser på tværs af hele webben. Autoritetsdata bliver et helt centralt omdrejningspunkt for linkning mellem forskellige datasæt på tværs af webben

– netop på grund af den entydige identifikation. I Danmark giver det anledning til en gentænkning af autoritetsdata, og Bibliografisk Råd arbejder pt. på en national strategi for autoritetsdata. I den forbindelse bliver internationalt anerkendte standarder for identifikation som f.eks. International Standard Name Identifier (ISNI) meget væsentlige.

Linked datas perspektiver for bibliotekerne

// Ud hvor brugerne er

Slutbrugerne bredt søger i stigende omfang information via generelle søgemaskiner på nettet og i mindre grad via bibliotekskatalogerne. Bibliotekernes metadata skal derfor integreres i webben, så bibliotekernes ressourcer kan nås med afsæt i søgning i f.eks. Google.

Netop Google står sammen med andre væsentlige søgemaskiner på nettet som Yahoo! og Bing bag schema.org, som definerer, hvordan metadata struktureres for at sikre optimal indeksering i søgemaskinerne.

Schema.org er målrettet alle ressourcer på nettet – ikke blot bibliotekerne – og udtrykker en ret elementær datamodel, der langtfra har den specificitet som kendetegner bibliotekernes metadata. En W3C Community Group, Schema Bib Extend, arbejder på forslag til udvidelse af schema.org's muligheder for at beskrive bibliografiske ressourcer. Schema.org-strukturerede data kan udtrykkes i RDF og harmonerer dermed fint med den bibliografiske udvikling. Der er iværksat et initiativ, der har som mål at opnå optimal indeksering af Databrøndens bibliografiske data – konkret bibliotek.dk's data – i Google ved eksponering af data i schema.org.

Schema.org tilbyder kun en forenklet strukturering af metadata. Biblioteksgrænseflader vil stadig være relevante. De kan tilbyde søgning og formidling på et niveau, som generelle søgemaskiner ikke kan. Databrønden er fortsat central, dels til understøttelse af biblioteksgrænsefladerne, dels som det sted hvor nationalbibliografiske data, bibliotekskatalogiseringsdata og bibliotekernes katalogiseringer skabes/koordineres.

// Bedre formidling

Linked data er født af visionen om det semantiske web. Målet er således at knytte semantik til de enkelte dataelementer, så applikationerne opnår mere viden om dataelementerne: Hvad de udtrykker, differentiering af homonymer osv. Sammenholdt med linked data-principperne: 'Tilbyd yderligere information om f.eks. forfatteren i et standardiseret format når URI-linket forfølges' og 'Suppler informationen med links til relaterede URI'er, f.eks. en beskrivelse af forfatterskabet' giver det mulighed for at lave mere intelligente formidlingsløsninger, der kan tilbyde flere og mere relevante informationer.

// Mere tilgængelige metadata

Bibliotekssektoren har været fantastisk dygtig til at udvikle og anvende standarder inden for sektoren. Til gengæld er det standarder, som er vanskeligt tilgængelige for interesserede uden for sektoren – tænk blot på Z39.50 og ISO2709, som aldrig har vundet indpas uden for bibliotekssektoren. Ved publicering af vores data som linked data bruger vi de HTTP- og XML-standarder, som anvendes helt generelt på webben. Vi åbner dermed for en bredere anvendelse af vores data. Der vil ikke mindst være perspektiver i anvendelse af autoritetsdata på tværs af sektorer, f.eks. biblioteker, museer og rettighedsorganisationer. Hertil kommer, at der med eksponeringen af bibliotekernes metadata i generelle web-standarder åbnes for, at en bredere base af aktører kan lave spændende applikationer til biblioteksbrugerne.

// ABM-samarbejde

Linked data er en velegnet metode til at koble datasæt såvel inden for egen institution som fra forskel-

"Bibliotekerne har et stærkt brand som garant for pålidelige og persistente data og vil kunne spille en central rolle som datahubs, der linker datasæt sammen i et 'web of data'."

lige domæner, f.eks. arkiver, biblioteker og museer. Gode eksempler på dette er Centre Pompidou samt Europeana, hvis nye datamodel EDM bygger på linked data-principper. Museerne arbejder allerede med datamodeller baseret på linked data, og alt peger i retning af, at den pågående konsolidering af dansk museums-it, SARA, tilsvarende vil blive baseret på linked data-principper. Der vil således åbne sig nye perspektiver for dels samarbejde om autoritetsdata, dels formidling af kulturarven på tværs af sektorer.

// Deling af metadata

Linked data giver mulighed for at dele ansvar for metadata i stedet for at skabe og vedligeholde alle data selv. Ultimativt kan en bibliografisk post afløses af en identifikator for det materiale/den ressource, der beskrives, suppleret med links til dataelementer forskellige steder på webben. F.eks. vil autoritative udenlandske navneformer kunne trækkes fra Virtual International Authority File, (VIAF). Databrønden kan i denne vision udvikle sig til en samling af links.

Lige nu er modenhedsgraden af linked data langt fra denne vision, bl.a. vil det ikke være realistisk at basere sig på, at links vil være persistente, så data kan indsamles dynamisk i brugsituationen. En mellemform kan være at høste bibliografiske dataelementer fra webben ud fra links og bygge den samlede beskrivelse af materialet/ressourcen op via denne høstning, på samme måde som supplerende data til de bibliografiske poster i dag høstes til databrønden og relateres til de bibliografiske poster.

// Metadata man kan stole på

Linked data handler fundamentalt om at gøre brug af data, som andre stiller til rådighed. Men hvordan ved vi, om disse data er pålidelige? Hvordan ved vi, at det link, som vi baserer os på, også virker om to år? Bibliotekerne har et stærkt brand som garant for pålidelige og persistente data og vil kunne spille en central rolle som datahubs, der linker datasæt sammen i et 'web of data'.

Udfordringer

Visionen om et 'web of data' har længe været præcis det: en vision, men i de seneste par år er der for alvor sket noget inden for linked data. Ud over Europeana er, inden for biblioteksområdet, Library of Congres, British Library, Deutsche Nationalbibliothek og Bibliothèque Nationale de France stærkt engageret i at publicere datasæt som linked open data. OCLC har udvidet WorldCat med eksponering af metadata i henhold til schema.org. Og DBC har forsøgsvist publiceret en afgrænset del af Dansk Bogfortegnelse som linked open data, primært med vidensopbygning for øje.

Der er altså på den internationale scene rigtig mange initiativer i gang for så vidt angår publicering af egne datasæt som linked open data. Anderledes forholder det sig med at udnytte andres datasæt publiceret som linked open data. Selvom der også på dette område foregår mere og mere, er der stadig en udpræget mangel på applikationer, der udnytter potentialet i linked data. Medvirkende hertil er givetvis, at en af de helt store udfordringer er linkning på tværs af datasæt, specielt muligheder for rationelt at kunne berige egne datasæt med links til andre datasæt. Dette gælder allerede inden for biblioteksdomænet for slet ikke at tale om linkning til andre domæners datasæt. Allermest afhænger mulighederne for generering af links til andre datasæt af entydig fælles identifikation af dataelementer, heraf væsentligheden af autoritetsdata.

En anden udfordring i forhold til at udnytte potentialet i linked data er, at publiceringen af mange datasæt som linked open data stadig har projektkarakter. Det indebærer, at der på nuværende tidspunkt hverken kan påregnes persistens, fuldstændighed eller regelmæssig ajourføring.

Den pågående bibliografiske udvikling vil utvivlsomt medvirke til imødegå disse udfordringer.