Bibliotekbasen som RDF – gamle data på en ny måte!

Av Asbjørn Risan, produkteier BIBSYS

Ingress: **UiO, UiB, UiT, NTNU og BIBSYS har fått støtte av Nasjonalbiblioteket til å kartlegge aktuelle ønsker og behov for bibliotekdata og utvikle en pilot for å konvertere Bibliotekbasen til RDF. Dette tilrettelegger for enklere og raskere utvikling av nye tjenester og økt bruk av bibliotekdata ved å koble Bibliotekbasen til den semantiske veven.**

Nye løsninger på gamle problem

Tenk deg at du sitter i informasjonsskranken i biblioteket og får følgende referansespørsmål:

- Den rumenske ambassaden ønsker en oversikt over hvilke norske bøker som er oversatt til rumensk, hvem som har oversatt og hvilke forlag bøkene er utgitt på
- En student ber om en oversikt over norske barne-/ungdomsbøker som er oversatt til andre språk de siste fem årene. Hun ønsker liste i et format som gjør at hun selv kan sortere ut fra språk, land og forfatter
- En journalist skriver en artikkel om Per Petterson og lurer på hvor mange språk «Ut å stjæle hester» er oversatt til og hvor mange av disse oversettelsene som er støttet av Nordisk ministerråd

Eksemplene er basert på brukstilfeller fra NORLA[1]

Dette er spørsmål som det er vanskelig og tidkrevende å finne svar på i de tradisjonelle bibliotekkatalogene. Dette fordi informasjonen ikke er søkbar, eller at presentasjonen ikke er tilrettelagt for å gi de oversiktene sluttbrukeren er ute etter.

Du kan søke og få en oversikt over hvilke bøker biblioteket har som er skrevet på rumensk, og det kan hende at du får avgrenset søket til kun norske oversatte bøker, men for å finne ut hvem som har oversatt bøkene og hvilke forlag som har gitt dem ut må du gå igjennom alle bøkene i listen manuelt.

På samme måte er det en utfordring å søke etter oversettelser av «Ut å stjæle hester» som er støttet av Nordisk ministerråd. Informasjonen finnes registeret i Bibliotekbasen, men for å nyttiggjøre seg denne må man gå igjennom alle postene. Istedenfor et tall på hvor mange språk «Ut å stjæle hester» er oversatt til og hvor mange av disse som er støttet av Nordisk ministerråd vil dagens søkesystemer kunne tilby en liste over titler (representert av en MARC-post) som tilfredsstiller søkeuttrykket som ble angitt. Det er en god start, men det er ikke det journalisten egentlig var ute etter.

Utfordringen er at søk og presentasjon av resultatene er basert på fullstendige MARC-poster som representerer «alle» aspekter ved et dokument. Løsningen er å bryte opp den fullstendige representasjonen til mindre enheter; tripler (RDF), som hver for seg uttrykker en bestemt egenskap ved et dokument, f.eks. at et dokument er oversatt til rumensk, eller at oversettelsen har mottatt støtte fra Nordisk ministerråd. Alle triplene fokuserer på faktaopplysninger, lik de fakta som en MARC-post kan inneholde, men der er ingen ytre struktur som bestemmer hva som er med og hva som blir utelatt. All informasjon er likt behandlet, alt er like søkbart og alt kan brukes til alt.

Ved å gjøre beregninger på ulike delmengder (alle som er oversatt til rumensk, alle som har originalspråk norsk og alle som har fått støtte fra Nordisk ministerråd) kan man finne de

[1] http://norla.no/

dokumentene som representerer alle delmengdene og presentere de opplysningene man ønsker. Man kan presentere kun navn på oversetter og ikke hele den bibliografiske posten.

Den semantiske veven

Å gjøre dette vil føre Bibliotekbasen et steg nærmere den semantiske veven som ble beskrevet av Tim Berners-Lee i artikkelen *The Semantic Web* i Scientific American 2001[2].

> *«The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. [….] In the near future, these developments will usher in significant new functionality as machines become much better able to process and "understand" the data that they merely display at present.»*

Den semantiske veven er beskrevet i mange artikler tidligere og det har vært noen prosjekter i Norge for å konvertere eksisterende data til tripler. Artikkelen *Linked data for fremtiden* av Aud Gjersdal i Bibliotekforum nr. 5, 2014 oppsummer disse på en god måte.

To av de viktigste prosjektene er Deichmanske biblioteks prosjekt for å konvertere sin base til RDF-tripler og Rådata Nå! som konverterte BIBSYS Autoritetsregister til RDF-tripler. Sistnevnte prosjekt ble kåret til det viktigste fagbibliotekprosjektet i 2011 av Bok og Bibliotek.

Neste skritt

Det er nå på tide å ta et steg videre og de fire store universitetene (UiO, UiB, UiT og NTNU) har sammen med BIBSYS fått støtte fra Nasjonalbiblioteket til et prosjekt der man skal konvertere Bibliotekbasen og andre relevante registre til RDF. Men til forskjell fra de andre prosjektene skal man ikke kun se på hvordan man kan konvertere et datasett til RDF, men også se på den potensielle nytten og bruken av et slikt datasett.

Bakgrunnen for prosjekt slik det er definert i prosjektsøknaden er:

> *«Norske fag- og forskningsbibliotek forvalter mange informasjonsressurser som skal understøtte utdanning- og forskningsprosessen. Semantisk webteknologi kan bidra til å øke relevansen og verdien av Bibliotekbasen og spesialsamlinger, ved å tilby studenten og forskeren denne kunnskapen på en effektiv og kvalitetssikker måte.*
>
> *[….]*
>
> *Prosjektet har som hensikt å levere konkrete leveranser som i sum vil bygge kompetanse innen semantisk webteknologi. Målet er å utvikle en kraftfull kunnskapsressurstjeneste som hjelper studenten og forskeren til å effektivt finne relevant informasjon.»*

Det er definert totalt 6 leveranser i prosjektet:

1. Interessentanalyse – Hvem vil ha nytte av eller være involvert i prosjektet

[2] http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html

2. Brukstilfeller - Liste med detaljerte brukstilfeller som prosjektet søker å realisere
3. Artikkel - Prosjektet skal levere en artikkel eller liknende som forklarer den semantiske vevens relevans for fag- og forskingsbibliotek
4. Informasjonsressurser - En oversikt over de viktigste informasjonsressurser som finnes i UH-bibliotekene og vise hvordan disse kan knyttes til den semantiske veven
5. Pilot - Det skal etableres en pilot som viser Bibliotekbasen med noen utvalgte ressurser knyttet til den semantiske veven. Piloten skal være basert på brukstilfellene.
6. Sluttrapport - Prosjektet skal lage en rapport for prosjektet med muligheter for videre arbeid

Mange av de tidligere prosjektene har vært fokusert på selve konverteringen av datasettene til RDF, mens det i dette prosjektet kun er en av flere leveranser. Piloten skal være basert på en kartlegging av relevante, faktiske brukstilfeller fra de involverte interessentene, som f.eks. å kunne ta ut lister som beskrevet i eksemplene i begynnelsen av artikkelen.

Prosjektet vil samarbeide tett med Deichmanske bibliotek for å trekke erfaringer fra det prosjektet som de har gjennomført. Prosjektet Rådata Nå! (som var en engangskonvertering) vil også bli tatt opp i prosjektet og det skal etableres en infrastruktur slik at den semantiske representasjonen av Bibliotekbasen og autoritetsregistrene på jevnlig basis vil bli oppdatert med nye og endrede data.

Prosjektet vil også se på muligheter for å etablere et autoritetsregister for Verk og hvordan ulike emneordssystemer kan integreres i BIBSYS Autoritetsregister. RDF forutsetter et sterkere fokus på bruk av autoriserte data, all bruk av fritekst er en stor utfordring når den semantiske betydningen skal avklares. Derfor blir innholdet i autoritetsregisteret til bibliotekbasen viktig, både for person og verk, men også innen emneord og klassifikasjon.

Den semantiske web har ikke rom for ulik tolkning av meningsinnholdet til en term. Det er kun om en term er autorisert at en datamaskin entydig kan klare å resonnere korrekt på grunnlag av dataene. Spesielt viktig er det om der finnes muligheter til å autorisere mot internasjonale autoritetsregister (som f.eks. Dewey), det vil etablere innganger til våre data i et internasjonalt perspektiv.

Det vil også etableres koblinger til andre etablerte datasett, som f.eks. DBpedia (som er RDF-versjonen av Wikipedia), slik at Bibliotekbasen blir en del av den semantiske veven, dvs. Linked Data.

Bibliotekbasen er lisensiert med en åpen lisenes (mer om dette i artikkelen *La de tusen blomster blomstre*, Bok og Bibliotek nr 4., 2013[3]) og bruk av Bibliotekbasen og autoritetsregistrene som tripler vil være fritt tilgjengelig for alle som ønsker å lage tjenester basert på disse dataene.

Et annet sentralt aspekt er at ved å konvertere til RDF bryter man ut av bibliotekspesifikke standarder og formater som i praksis er effektive hindre for bruk utenfor bibliotekssektoren, og tar i bruk etablerte og enklere tilgjengelige teknologier som har potensiale til å nå et større publikum.

Veien videre

Forutsatt at prosjektet er en suksess og piloten blir besluttet videreført og satt i regulær drift vil dette prosjektet og den åpne lisensen danne grobunn for å lettere kunne utnytte det potensialet som ligger i Bibliotekbasen og som bibliotekarer ved norske fag- og forskningsbibliotek gjennom mer enn 40 år har bygget opp.

Prosjektet skal etter planen ferdigstilles til påske i 2016 og vi har forhåpentligvis mye å glede oss til nå som Bibliotekbasen tar steget ut av siloen og inn i den semantiske veven!

[3] http://www.bokogbibliotek.no/la-de-tusen-blomster-blomstre