Frbrisering av testdata fra Deichmanske

Formål med eksperimentet

Formålet med den eksperimentelle frbriseringen har vært å undersøke mulighetene for å tolke postene i katalogen i henhold til FRBR modellen. Slik informasjon kan for eksempel brukes i implementasjon av grensesnitt med trefflister og lenking/navigering basert på å FRBR modellens entiteter og relasjoner, biblioteksdata som linked data og andre anvendelser. I dette eksperimentet har vi har sett på forskjellige problemstillinger som hvilke entiteter og relasjoner kan vi systematisk trekke ut av disse dataene, hva er kvaliteten på resultatet og hva er de mest vesentlige årsakene til feiltolkninger, hvilke begrensinger og/eller problemer gir dagens bruk av MARC-formatet og dagens katalogiseringspraksisen etc.

Om testdataene og forsøkene som er gjort

Frbriseringen er utført i to forskjellige runder. I første omgang tok vi for oss 4 forskjellige filer for henholdvis forfatterne Per Petterson (61 poster), Knut Hamsun (842 poster), J.R.R. Tolkien (241 poster), William Shakespeare (588 poster). Resultatet fra denne frbriseringen ble vurdert og vi fant at det var mange problemer med dataene som skapte dårlig resultat. Vi bestemte derfor å gå videre med eksperimentet ved å bedre kvaliteten på postene og dette ble gjort for Per Petterson og Knut Hamsun testsamlingene. Disse ble så frbrisert i en ny runde og vi fant at det er et stort potensiale for å bedre resultatet av frbriseringen også innefor dagens katalogiseringsregler.

Verktøy for frbrisering

Verktøyet som er brukt for frbrisering av postene er utviklet ved Institutt for datateknikk ved NTNU (v/Trond Aalberg) og er en videreføring av den programvaren som ble brukt i et BIBSYS FRBR-prosjekt. Verktøyet består av en database hvor man kan lage regler for å tolke dataene (hvilke felter som identifiserer hvilke entiteter under hvilke betingelser etc). Disse reglene brukes for å generere ei XSLT-fil som er et "program" for å transformere XML dokumenter. XSLT-fila gjør om en MARCXML-fil til en XML-fil som består av ett sett FRBR poster. Hver unik FRBR entitet er en egen post og relasjoner mellom entitetene er kodet som lenker. I praksis tar konverteringen vare på alle MARC-feltene som man ønsker å ha med i

FRBR-postene. Den største utfordringen i bruken av verktøyet er å skrive reglene for hvordan dataene skal tolkes. Selv om mange regler vil være felles for flere kataloger, kan det være vesentlige forskjeller både på grunn av forskjellige MARC-formater og på grunn av forskjellig katalogiseringspraksis.

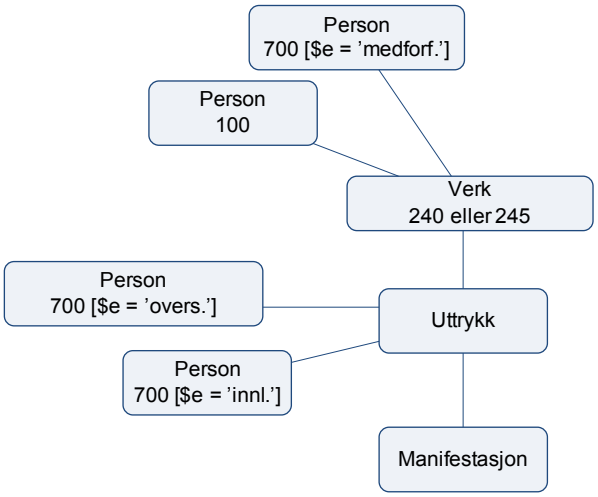
Entiteter og relasjoner vi har sett etter

Entiteter og relasjoner vi har prøvd å identifisere kan best beskrives med et sett av scenarier for typiske FRBR "strukturer" som vi ofte kan finne i bibliotekposter.

Enkel FRBR struktur

Det enkleste av disse strukturene finner vi i poster som beskriver en manifestasjon som inneholder ett uttrykk hvor uttrykket er en realisering av et verk skapt av en enkelt person. Slike poster er det i praksis mange av i de fleste kataloger og den største utfordingen er å finne riktig informasjon om de forskjellige entitetene. Personen som har skapt verket kan vi finne i 100-feltet (110 hvis det er en kroporasjon). Hvis posten har en 240-tittel er denne godt egnet til å identifisere verket, men siden det er mange poster uten orginal- eller standardtittel må vi ofte benytte 245-tittelen i stedet. At manifestasjonen inneholder et uttrykk av dette verket er implisitt og selv om vi ikke har egne felt som forteller oss tittel på uttrykket kan vi vanligvis finne koder som forteller oss språk og hva slags type uttrykk dette er er (tekst, lyd, film etc.). Manifestasjonen kan identifiseres med ISBN-nummer eller annen identifikator, men siden hver katalogpost representerer en unik manifestasjon kan vi også i praksis bruke postnummeret (kontrollfelt 001) som unik id for manifestasjonen. For slike poster fant vi at det største problemet i testsamlingene var mangelfull bruk av 240-felter og en del tilfeller av titler i 240 (og 245) som var skrevet feil. Hvis det mangler orginaltittel i 240 for eksempel for en oversettelse er man nødt til å bruke 245-tittel for å identifisere verket og dermed ender man opp med "falske" verk.

Enkel FRBR struktur med flere personer

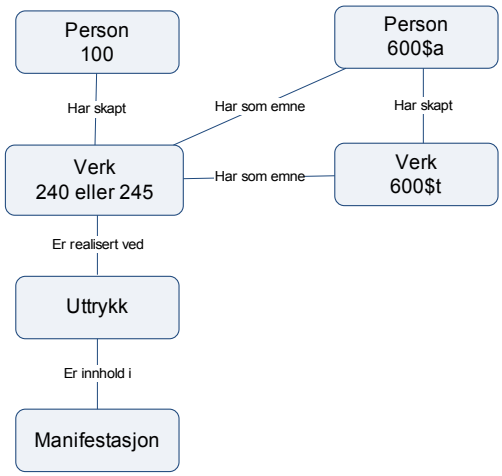


En litt annen variant av den enkle FRBR-strukturen finner vi i poster som beskriver ett verk, men hvor det er katalogisert flere personer (biinnførsler i 700-felt). Denne gjelder kun hvis det ikke er $t i 700-feltet. I tillegg til mulige problemer tilsvarende som for scenarioet over er vi for denne typen

struktur avhengig av å finne ut hvilken entitetestype personene skal relateres når vi frbriserer disse feltene. Medforfatter er skapere av verket og skal derfor relateres til verksentiteten, oversettere har "realisert" uttrykket og skal derfor relateres til uttrykksentiteten. For å koble forskjellige biinnførsler til forskjellige produkt-entiteter er vi avhengig av informasjon som indikerer hva som er riktig kobling. Funksjonskoder i $e som er katalogisert gir i teorien denne informasjonen men i praksis fant vi stor variasjon i skriveform og bruken av funksjonskoder. For biinførsler uten funksjonskode vet vi i praksis ikke om personen skal relateres til verk, uttrykk eller manifestasjon, og ved variasjon i skrivemåte for en gitt funksjonsbetegnelse må vi i praksis kjenne alle mulige varianter som er brukt for å få et riktig resultat.

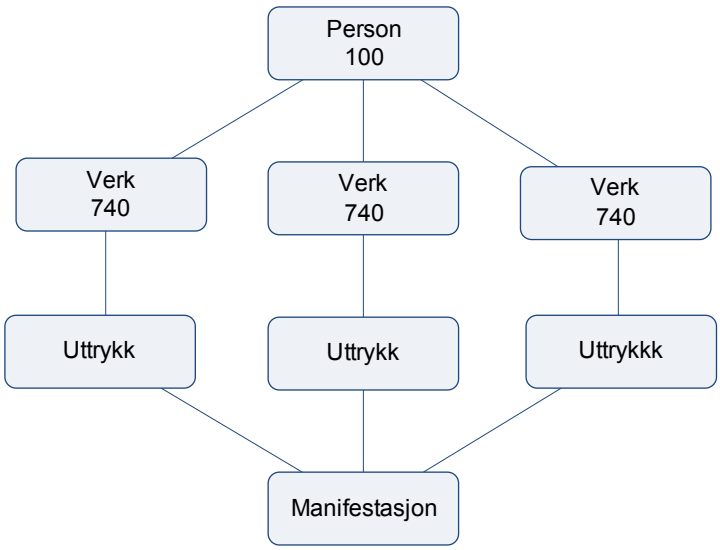
Person og verk som emne

Emneinnførsler på person eller person-tittel kan enkelt identifisere når 600-feltet er brukt til dette. I FRBR kan et verk ha en person eller et annet verk som emne (kan også ha uttrykk eller manifestasjon som emne). For manifestasjoner som bare har ett verk er det implisitt at verket som manifestasjonen inneholder et utrykk av er det som omhandler personene eller tittelen som er katalogisert i 600-feltet. Selv om dette er en tilsynelatende enkel struktur er det også en del problemer ved dette. I norske katalogpostene er vanligvis titler som emne gjengitt ved den norske tittelen. For titler som har ikke-norsk orginaltittel får vi dermed et problem med å identifisere verket som er emne på en konsistent måte.



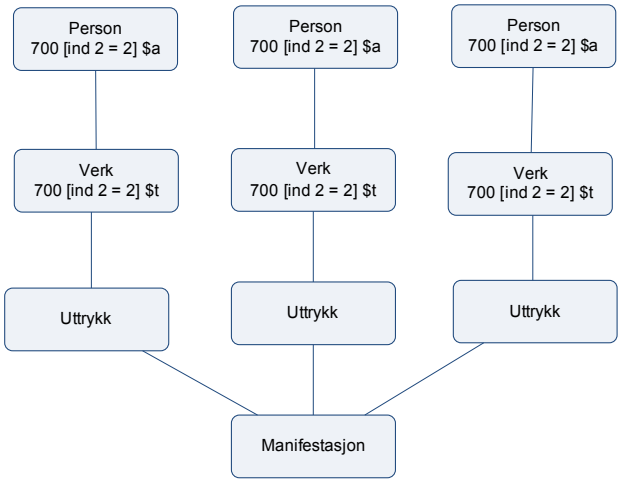
Poster med flere verk av samme person

Problemene med å tolke postene i henhold til FRBR modellen øker etter hvert som informasjonsmengden i en posten øker. I testdataene våre fant vi mange poster for utgivelser som inneholder flere verk. Dette inkluderer bøker med to eller flere romaner, novellesamlinger m.m. Generelt er dette



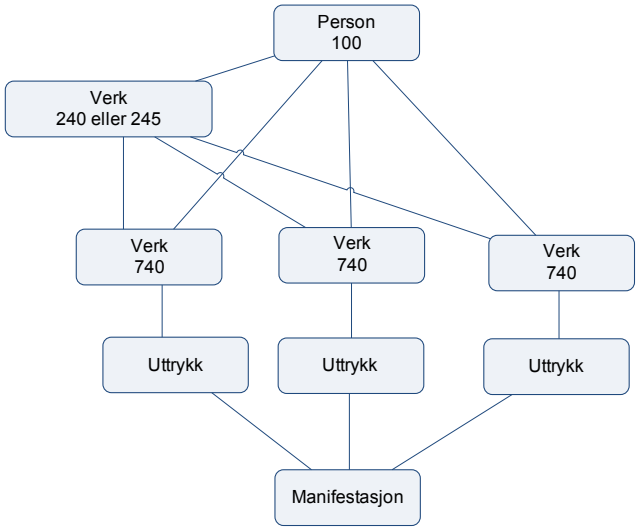
katalogisert ved bruk av 740-innførsler for tittelen og ideelt sett bør vi da kunne identifisere entiteter og relasjoner som vist i figuren. I praksis er det derimot mange problemer knyttet til dette. Biinførsler på titler benyttes til forskjellige formål og det er vanskelig å bruke 740-titler systematisk. I første runde av frbriseringen fant vi noen poster hvor dette kunne identifiseres, men da basert på at det var brukt "Samlede romaner" eller "Samlede verker" i 245-tittelen.Generelt var det vanskelig å identifisere verkene i slike tilfeller også fordi 740-titler vanligvis er den oversatte tittelen (i praksis uttrykkets tittel).

Poster med flere verk av forskjellige personer



Dette er en type struktur som tilsvarer scenarioet over, men med den forskjellen at verkene er skapt av forskjellige personer. Typiske eksempler er essay- eller novellesamlinger som inkluderer verker av forskjellige forfattere. Bruke av 700-felter for slike innførsler gjør at vi enkelt kan finne hvilken person som har skapt hvilket verk (så lenge det ikke finnes medforfattere), men også for dette tilfellet finner vi at vi ofte identifiserer verket feil for oversatte verk på grunn av at biinførsler baseres på uttrykkets tittel.

Flere verk og samlingen som eget verk



Den siste varianten vi har sett på er tilfeller hvor en publikasjon inneholder flere verk, men hvor publikasjonen som helhet også kan oppfattes som et verk. I testsamlingene var det eksempler på novellesamlinger som var utgitt på andre språk og hvor det dermed kan være hensiktsmesig å ha med samlingen som selvstendig verk. Vi finner også novellesamlinger av enkeltforfattere hvor både samlingen og enkeltnovellene kan sees som selvstendige verk. En av de største utfordringene var nettopp å kunne skille disse fra tilfellene hvor vi ikke ønsket å generere egne verk for samlingen. For utgivelser av typen bind 4 i samlede romaner er det av liten interesse for brukerne å få presentert samlede romaner bind 4 som et verk av Hamsun, mens for en novellesamling kan det motsatte være ønskelig.

Erfaringer fra første runde av frbriseringen

Første runde av frbrisering ble utført for alle testsamlingenene og inkluderte regler for:

Poster med "enkel verk" hvor person i 100 felt er skaper av verk med tittel i 240 eller 245. Uttrykk ble identfisert med språkode fra 008 og 041 og formkode i 019. Videre brukte vi

biinførslser for person med funksjonsbetegnelser for innleser, oversetter, medforfatter og redaktør. Største feilkilde i slike poster er relatert til titlene både i 240 og 245. For 240-titler fant vi avvik i skrivemåte samt at det var flere eksempler på at 240-tittel manglet. For 245 var det også eksempler på feil skrivemåte samt at vi endte opp med mange verk som var åpenbare feil/uinteressante verk ("samlede romaner" etc).

Frbrisering av de mer komplekse eksemplene ble bare gjort i begrenset omfang. Det største hinderet for å kunne få meningsfull informasjon ut av 700 og 740 felter var at vi ikke systematisk kunne finne hvilke innførsler som representerte analytiske innførsler eller ikke. I tillegg gav frbrisering av disse feltene mange feil verk på grunn av språket på tittel. Dette var spesielt et problem for Tolkien og Shakespeare hvor verkene ideelt bør identifiseres med engelsk orginaltittel slik den brukes i 240. Her har vi for eksempel tilfeller av 740 innførsler i samme post som både har norsk og engels tittel, eller tilfeller av bare norsk tittel i 740. Det generelle problemet er at vi ikke kan vite om biinførsel er "verk" eller "uttrykk" og vi vet heller ikke om to titler er samme verk eller to forskjellige verk.

Redigering av postene

Basert på resultatene fra første runde så vi at det var mulighet for bedre resultat og vi bestemte at det ville vært interessant å rette feil i postene, legge til informasjon samt se på effekten av en mer konsistent bruk av indikatorene. Endringene som ble gjort er i overensstemmelse med katalogiseringsreglene. For å begrense omfanget av dette arbeidet jobbet vi videre kun med Per Petterson og Knut Hamsun (908 poster). Siden dette er norske forfattere er dette i praksis også en forenkling av noen av de problemene som skyldes språk på tittel i biinførsler.

Endringer som ble gjort i de 908 postene i inkluderte:

- Endring av språkkode i 008 for 5 poster
- Legge til 240$a i 85 poster og rette skrivefeil i 24 eksisterende poster
- Rette skrivefeil i 245$a (eller feil ISBD syntaks) i 6 poster
- Endre første indikator i 245. Før korreksjon hadde 137 poster ind1 = 0 eller blank, mens ind1 = 1var brukt i 774 poster. Etter korreksjon var fordelingen 263 – 651. Den vesentligste endringen her var at i postene etter korreksjon er det 113færre 245 felt som skal tolkes som egne verk.
- Korrigere 700-felter. I opprinnelige poster finner vi 948 700$a subfelter og 545 700$t felter. I de korrigerte postene er antallet redusert til hhv 917 $a og 481 $t.

Endringen skyldes en mer systematisk bruk av 740-feltene der alle innførsler har samme forfatter (som er registrert i 100). Her ble også titler endret til samme skriveform og språk som man ville brukt i 240 for et mindre antall poster, men det er ikke mulig å telle disse endringene på grunn av de øvrige endringer som ble gjort.

- Endre andre indikator i 700-felt for å tydeliggjøre om en innførsel er et eget verk eller ikke. På grunn av endringene som er gjort er det vanskelig å identifisere antallet korreksjoner
- Korrigering av 740 felter. I de opprinnelige postene er det 387 felter, mens det i de korrigerte postene er 873 felter. Årsaken til denne økningen er at nye felter er lagt til for "verks"-skrivemåten og andre indikator er brukt for å kunne skille verkstitlene fra andre titler. Også her ble det gjennomført en mer systematisk bruk av andre indikator, men disse endringene er vanskelig å telle spesifikt.

Resultat etter retting av postene

Etter at postene var rettet ble det utført en ny eksport og en ny runde frbrisering. Resultatet av dette var en tydelig forbedring som i hovedsak kan oppsummeres som:

- Færre "falske" verk både på grunn av bedre kvalitet i 240/245 titler. I tillegg var det i denne runden mulig å ignorere 245 titler som ikke er signifikante verkstitler (basert på indikator).
- Flere "riktige" verk fordi det var mulig å utnytte titler i 700 og 740 i vesentlig større grad siden andre indikator nå kunne brukes for å bestemme om en tittel var et eget verk eller ikke.

I tillegg er det et viktig poeng at man i denne omgangen kunne lage enklere regler. Å bruke indikator-verdier som betingelser er for eksempel vesentlig enklere enn måtte skrive regler basert på innholdet i tittelfelt.

Konverteringen som ble utført etter rettingen av postene var litt forskjellig fra den opprinnelige og resultatet er ikke direkte sammenlignbart mht. antall entiteter og relasjoner. Videre er bruken av Petterson og Hamsun i den siste konverteringen en forenkling av noen problemer siden man i en norsk katalog for norske forfattere får færre problemer mht. språk som er brukt på titler i biinførsler og emneinnførsler.

De rettinger som ble gjort er ikke i konflikt med katalogiseringsreglene og resultatet viser at man vesentlig kan forbedre postene med tanke på "frbr-kvalitet". Konsistent bruk av indikatorer, funksjonsbetegnelser osv. er vesentlig mer viktig i kontekst av FRBR enn i tradisjonelle biblioteksystemer. Tilsvarende krav gjelder for konsistent bruk av standard og orginaltitler.
Dette er dessverre et omfattende arbeid siden det for det meste må gjøres manuelt.