NTNU
Det skapende universitet

Påliteligheten til identifiserte verk i bibliografiske data

En undersøkelse om hvorvidt ulike karakteristikker ved bibliografiske data kan angi påliteligheten til identifiserte verk

Tobias Hartvedt Knudsen

Master i informatikk
Innlevert: juni 2017
Hovedveileder: Trond Aalberg, IDI

Norges teknisk-naturvitenskapelige universitet
Institutt for datateknologi og informatikk

Sammendrag

I de siste årene har store mengder bibliografiske data blitt publisert som linked open data. Dataene blir ofte transformert til FRBR eller andre modeller med eksplisitte entiteter for intellektuelle verk og publisert uten fokus på kvaliteten. Denne oppgaven utforsker en metode for å angi påliteligheten til verk som blir generert fra bibliografiske data.

Forskningsspørsmålene i prosjektet: Hvilke kvalitetsproblemer finnes i eksisterende bibliografiske samlinger som er publisert som linked open data? Finnes det data fra verksidentifisering som kan brukes til å angi påliteligheten til genererte verk?

Prosjektet undersøker kvaliteten på eksisterende systemer som har publisert bibliografiske data til linked open data-skyen for å kartlegge utfordringene som eksisterer i dag. I tillegg gjennomføres en analyse av karakteristikkene til bibliografiske poster som kan si noe om påliteligheten til verkene som har blitt generert.

Resultatene fra prosjektet bekrefter at kvaliteten på eksisterende bibliografiske data i linked open data-skyen er dårlig, men at det er mulig å bruke data fra verksidentifisering for å angi påliteligheten til om verket er riktig identifisert.

Resultatene fra dette prosjektet kan brukes til å vurdere påliteligheten til verk som blir identifisert fra bibliografiske poster. Verksidentifisering er første steget for å produsere gode data basert på FRBR-modellen. Bibliografiske data med riktig identifiserte entiteter bidrar til å kunne publisere ryddige data med god kvalitet til linked open data-skyen.

Abstract

During the past years a large number of bibliographic records have been published as linked open data. The data are often transformed to FRBR and similar models with explisit entities for intellectual works and published without attention to data quality. This Master Thesis explores a method to specify the confidence of a work that has been generated form bibliographic data.

The research questions of the project: What are the challenges of data quality in existing bibliographic collections that are published as linked open data? Are there any data available from the identification of works that can be used to evaluate the confidence of generated works.

The project assesses the quality of existing systems that have published bibliographic data to the linked open data cloud, to explore the quality challenges that exist today. Further, the charactersitics of bibliographic records related to confidence of the generated works are analyzed.

The project results confirm that the quality of existing bibliographic data in the linked open data cloud are low. However, it has also been proven that data from the identification of works can be used to assess the confidence of whether the work is correctly identified.

The results of this project can be used to evaluate the confidence of works that have been identified from bibliographic records. The identification of works is the first step to produce high quality data based on the FRBR model. Bibliographic data with correctly identified entities contributes to the publication of well organized data of high quality to the linked open data cloud.

Forord

Jeg vil takke veilederen min, Trond Aalberg, for et godt samarbeid. Jeg er veldig takknemlig for hjelp og raske svar: Trond bruker 34 minutter i gjennomsnitt til å svare på mail.[1]

Takk til mamma og pappa som gjennom fem år har støttet meg, og for å være interesserte i det jeg holder på med.

Til slutt vil jeg takke Eline som har vært der for meg, og for å ha hjulpet meg om preposisjonene mine.

Trondheim, juni 2017
Tobias Hartvedt Knudsen

[1] Utregnet fra mailkorrespondanse i tidsrommet 12. september 2016 til 11. juni 2017

Innhold

Figurer

Tabeller

Kapittel 1

Introduksjon

De siste årene har det blitt stadig mer vanlig å publisere data på nett som linked open data. Flere store biblioteker har publisert bibliografiene sine til linked open data skyen og i forbindelse med dette stilles det nye krav til formatet og modellen som brukes for å lagre de bibliografiske postene. Det vanligste formatet for bibliografiske poster i dag er MARC. MARC-formatet fungerer bra til lagring av katalogposter, men i dag stilles det nye semantiske krav, og krav om interoperabilitet til dataene. Derfor har flere biblioteker rundt om i verden begynt å transformere bibliografiene sine fra MARC til FRBR. Eksempler på bibliotek som har publisert bibliografien sin som linked open data er Library of Congress, The Bibliothèque national de France, The British National Bibliography, Deutsche Nationalbibliothek, Biblioteca Nacional de España og Europeana [2]. FRBR er en kravsspesifikasjonsmodell utviklet av International Federation of Library Associations, IFLA[3]. FRBR er i økende grad tatt i bruk i det siste på grunn av behovet for en formell modell i utviklingen av semantiske web-applikasjoner som linked open data [4].

1.1 Motivasjon

Til tross for at flere og flere biblioteker publiserer data som linked open data, er det ikke alltid lagt inn like mye arbeid i kvalitetssikring og evaluering av dataene som blir publisert. Det er ofte lagt mer arbeid i å få dataene publisert enn å få data med god kvalitet publisert. Det er mange eksempler på bibliotek som har publiserert bibliografiske data til linked open data skyen, men det er få eksempler på bibliotek som har publisert gode og støyfrie bibliografiske data. Bibliografiske kataloger består i flere tilfeller av mange millioner poster. Det kan derfor være vanskelig og tidkrevende å transformere data fra ett format til et nytt, og i særdeleshet til et nytt system som stiller strengere semantiske krav. Derfor er det ønskelig å lage en metode eller et system som kan gjøre denne prosessen automatisk og samtidig sikre og evaluere at kvaliteten på det som blir produsert er god.

Målet med masteroppgaven er å se nærmere på datakvaliteten ved bibliotekene som allerede har publisert bibliografiene sine som linked open data. Dette er for å kartlegge utfordringene som eksisterer i dag og for å kunne bidra til forbedringer. FRBR-modellen baserer seg på å identifisere poster som forskjellige entiteter. Den første av disse entitetene er verk. Et verk representerer den intellektuelle entiteten til en gjenstand. De resterende entitetene i FRBR-modellen kan relateres til verksentiteter. For eksempel er entiteten *uttrykk* en realisering av et verk. Av den grunn er det viktig, i første omgang, å identifisere verkene i bibliografien for så å identifisere de resterende entitetene når

bibliografien skal over til FRBR-modellen og potensielt publiseres som linked open data.

1.2 Forskningsspørsmål

Målsetningen med oppgaven er å se på kvaliteten til eksisterende systemer og de bibliografiske dataene som er blitt publisert som linked open data. Videre er målet å finne hvilke karakteristikker ved en bibliografisk post som kan være med å identifisere om posten representerer et verk eller ikke. Forskningsspørsmålene for dette prosjektet er:

Q_1: Hvordan er tilstanden til kvaliteten på dataene i samlinger som er publisert som linked open data?

- **q_1**: Hvilke kvalitetsproblemer er det i dataene som er publisert som linked open data?

Q_2: Finnes det data fra verksidentifisering som kan brukes for å angi påliteligheten til genererte verk?

- **q_1**: Hvilke data fra verksidentifisering kan brukes for å angi påliteligheten til genererte verk?
- **q_2**: Hvordan kan verk identifiseres uten å skape feilidentifiserte verk?

1.3 Fremgangsmåte

I dette prosjektet var det i første omgang avgjørende å sette seg ordentlig inn i linked open data, FRBR-modellen, MARC-formatet og FRBRiseringsprosessen. For å kunne jobbe med data som var på MARC-formatet og data som var blitt FRBRisert til å passe FRBR-modellen, var det nødvendig å ha en god forståelse for disse konseptene. Videre var det nødvendig å gjøre en analyse av kvaliteten til systemer som hadde publisert data som linked open data. Dette for å vite hvilke datakvalitets-metrikker som er vanlig å bruke i forbindelse med publisering av linked open data. Når datakvalitet skal analyseres er det avgjørende å ha kriterier for hva som gjør kvaliteten til data god, og hva som gjør den dårlig.

Etter å ha skapt klare rammer for hva som kan defineres som god kvalitet innenfor linked open data, ble det gjort en systematisk analyse av de forskjellige utfordringene eksisterende systemer sliter med. Problemene ble kartlagt for å lettere kunne vurdere på hvilken måte utfordringene kunne forbedres. Det ble også undersøkt hvordan andre prosjekter hadde gjennomført transformeringer fra MARC til FRBR, og hva som gjorde noen prosjekter bedre enn andre.

Deretter ble det funnet rådata som kunne brukes i prosjektet. Det var ønskelig å bruke hele kataloger fra kjente bibliotek, og fra forskjellige typer bibliotek. I dette prosjektet ble kataloger på MARC21-format fra fire forskjellige biblioteker brukt. To av bibliotekene var nasjonalbiblioteker, og de to andre var universitetsbiblioteker. Disse fire bibliotekene representerer tre forskjellige land. Dette var viktig for å kunne gi mer representative resultater.

Til slutt ble dataene FRBRisert, og et utvalg poster ble valgt for å analyseres. Dataene ble analysert og forskjellige karakteristikker i de forskjellige verkskandidatene ble undersøkt for å finne ut av hvilke karakteristikker som kunne gi en høyere sannsynlighet for at en verkskandidat kan representere et verk.

1.4 Bidrag

I dette prosjektet er kvaliteten på bibliografiske data publisert som linked open data analysert. Gjennom analysen ble det kartlagt at de prosjektene som har publisert bibliografiske data som linked open data har flere problemer når det gjelder datakvalitet. I prosjektet er målsetningen å finne en metode for å forbedre kvaliteten på bibliografiske data som blir publisert. Dette ble gjort ved å identifisere verk på en bedre og mer nøyaktig måte. Poster som har blitt FRBRisert fra MARC har blitt analysert, og flere karakteristikker har blitt funnet i disse postene som kan brukes til å identifisere verk. Dette er gjort ved å finne de karakteristikkene som er felles for de verkene som er riktig identifisert, og som ikke er tilstede i de verkene som er feilidentifisert. På den måten kan man si at en karakteristikk er statistisk definert for å kunne identifisere verk.

1.5 Struktur for oppgaven

Strukturen for oppgaven er følgende. Kapittel to omhandler MARC-formatet, FRBR-modellen og FRBRisering. I kapittel 3 presenteres teori rundt datakvalitet i forskjellige datakonsepter som databaser, datavarehus og linked open data. I kapittel fire presenteres en analyse av kvaliteten i forskjellige eksisterende systemer som inneholder bibliografiske data som linked open data. Kapittel fem beskriver tilnærmingen til prosjektet og hvordan analysen har blitt gjennomført. I kapittel seks presenteres resultatene fra prosjektet. Konklusjonene og forslag til videre arbeid blir presentert i kapittel syv.

Kapittel 2

FRBRisering

Dette kapittelet gir en innføring i modellene og teknologiene som er relevante for prosjektet. Først blir knowledge extraction beskrevet, og deretter MARC-formatet og FRBR-modellen. Til slutt blir eksempler på FRBRiseringsprosjekter presentert.

2.1 Knowledge extraction

Internett inneholder enorme mengder informasjon, men potensialet til verdensveven er langt fra utnyttet til det fulle. Dataene er åpne og tilgjengelige for hvem som helst, men de er ustrukturerte, har dårlig kvalitet, og er fulle av støy. Informasjonen er lett tilgjengelig, men når den ikke er satt i system har den lav nytteverdi. Knowledge extraction, kunnskapsutvinning, er opprettelse av kunnskap basert på strukturerte eller ustrukturerte data. Kunnskapen må være maskinlesbar som gjør at dataene er mer åpne for andre systemer.

Strukturerte tilgjengelige data vil, i samarbeid med andre teknologier som for eksempel kunstig intelligens, gi mange muligheter som ikke eksisterer i dag. Flere eksempler på mulighetene dette kan skape er beskrevet i Weikum & Theobalds artikkel "From information to knowledge: harvesting entities and relationships from web sources."[5]. De to hovedfordelene ved å strukturere data på en god måte er først å gjøre tidligere utilgjengelig kunnskap tilgjengelig for bruk og læring, og deretter å gjøre denne kunnskapen åpen slik at flere systemer har tilgang til å bruke informasjonen.

Knowledge extraction handler om å skape informasjon som ikke allerede eksisterer, og dette i seg selv kan være en komplisert oppgave. I hovedsak er målet å identifisere forskjellige entiteter i dataene for så å kunne finne relasjonene mellom dem. Det er ikke nok å kun identifisere entiteter og relasjoner, disse må også være riktig identifisert. Dette burde være en selvfølge, men som denne oppgaven skal vise har ikke kvalitet nødvendigvis det fokuset den burde ha.

Det første steget i knowledge extraction er entitetsutvinning. Entiteter kan være alt fra personer, byer og firma, til bøker, sanger, konserter og hendelser. Entitetene må deretter bli klassifisert basert på hva de egentlig representerer. Klassene brukes til å gruppere sammen entiteter som har ting til felles. Eksempler på klasser kan være musikere, gitarspillere, astronauter, lærere, land, noveller og så videre. Klasser kan videre grupperes eller være underklasser av andre klasser. Entiteter kan også eksistere i flere forskjellige klasser. Målet er å kunne klassifisere en entitet i så mange klasser som mulig, som igjen vil øke kvaliteten og mulighetene til videre bruk av entiteten[5].

Entitetsklassifisering byr på problemer blant annet på grunn av antallet forskjellige klasser, og det voksende antallet nye klasser. Det eksisterer flere prosjekter som har som mål å skape en kunnskapsbase som definerer klasser, entiteter og relasjoner. Eksempler på slike prosjekter er DBpedia[6], KnowItAll[7], YAGO[5] og WordNet[8]. Disse baserer seg på forskjellige fremgangsmåter. YAGO [5] for eksempel bruker klasser hentet fra WordNet[8] og Wikipedia. Omfanget av entiteter og klasser er en utfordring, og det er av den grunn ønskelig med automatiske prosesser som identifiserer entiteter og klassifiserer dem. Automatiske prosesser er i stor grad en mer effektiv måte å identifisere entiteter, utfordringen er å ha kvalitet på samme nivå som ved en manuell identifisering.

Identifisering av entiteter er et viktig startpunkt i knowledge extraction. Men for at entitetene skal være nyttige er det også viktig å finne ut hvordan de forskjellige entitetene kan relateres til hverandre. Det finnes utallige mulige relasjoner som kan identifiseres mellom entiteter, og i likhet med entitetene vil flere korrekte relasjoner gjøre dataene mer nyttig. Relasjonutvinning avhenger av at entitetene er riktig identifisert, så det er igjen viktig at entitetsutvinningen er gjennomført på en god måte, og med god kvalitetskontroll.

Relasjoner defineres ofte som tripler, som består av to entiteter og relasjonstypen mellom dem. For eksempel kan det skapes en relasjon mellom forfatteren Agatha Christie og et kjent verk hun har skrevet, eksempelvis *Death on the Nile*. Entitetene vil da være Agatha Christie og *Death on the Nile* og relasjonen vil være at verket ble skrevet av Agatha Christie.

2.2 MAchine-Readable Cataloging

Dagens bibliografiske data er stort sett basert på katalogregler fra forskjellige versjoner av MARC-formatet som for eksempel MARC 21 og UNIMARC. MARC 21 er den vanligste måten for biblioteker å lagre og utveksle biblioteksdata i dag. MARC står for MAchine-Readable Cataloging, der MAchine-Readable betyr at en maskin skal kunne lese og tolke dataene i katalogen. Cataloging betyr å lage katalogposter av informasjonen som skal lagres.

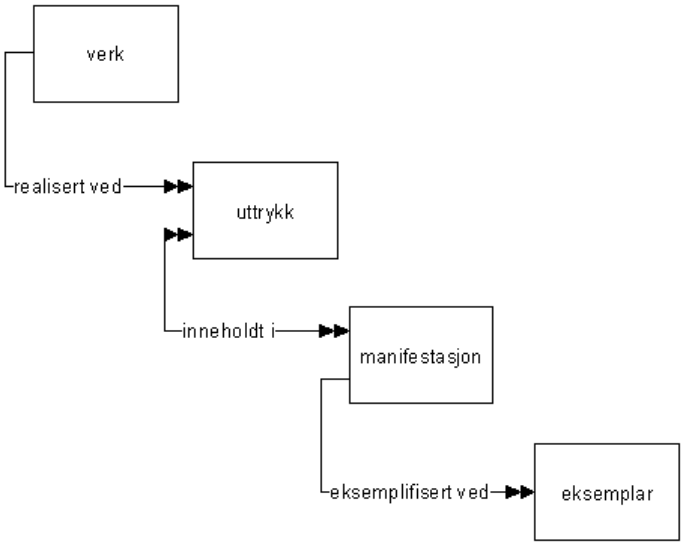
MARC-poster består av felt med informasjon som omhandler gjenstanden det er snakk om, og feltene er representert med 3-sifrede koder. Feltene identifiserer hvilket type felt som følger. Videre deles feltene opp i underfelt som er markert med bokstaver. For eksempel beskriver felt 100 en person, underfelt a er navnet til personen, c er eventuelle titler assosiert med personen og d er datoer assosiert med personen.

2.3 Functional Requirements for Bibliographic Records

Functional Requirements for Bibliographic Records, FRBR, er en entitetsrelasjonsmodell som ble laget for å kunne generalisere bibliografiske data, og som skulle være uavhengig av katalogiseringsmetode og implementasjon. FRBR er en ER-modul som består av entiteter og relasjoner mellom disse entitetene. FRBR ble laget for å gi et nytt perspektiv på struktur og relasjon mellom bibliografiske- og autoritetsposter. Videre var målet å skape et mer presist vokabular for å enklere kunne katalogisere data. Videre ønsker FRBR å bidra til å la bibliotek treffe et bredt spekter av forventninger fra brukere.

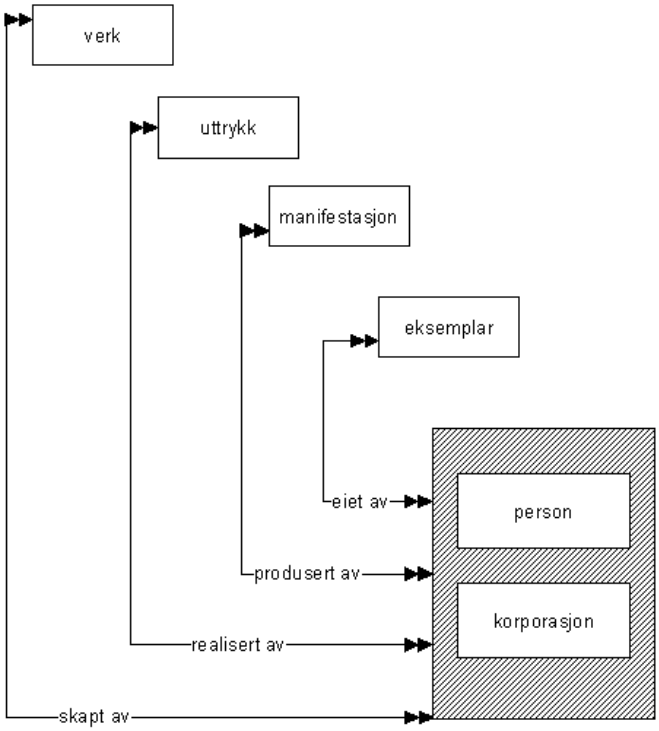
2.3.1 Entitetstyper og grupper i FRBR

FRBR baserer seg på bruken av forskjellige entiteter i bibliografiske data. Disse entitetene er delt opp i tre grupper der gruppe 1 fokuserer på postens representasjon, gruppe 2 inneholder personer og korporasjoner, og den tredje gruppen representerer forskjellige emner for et verk. Gruppe 1 er den mest relevante gruppen for denne oppgaven. Den består av entitetene *verk*, *uttrykk*, *manifestasjon* og *eksemplar*. Verk beskrives som et selvstendig intellektuelt eller kunstnerisk arbeid, og et uttrykk er den intellektuelle eller kunstneriske realiseringen av et verk. Manifestasjon er den fysiske realiseringen av et uttrykk av et verk, og eksemplar er et enkelt eksemplar av en manifestasjon som reflekterer den fysiske formen[1]. Figur 2.1 viser oppdelingen og forholdet mellom entitetene i gruppe 1. De doble pilene beskriver én-til-mange-forhold eller mange-til-mange-forhold.



Figur 2.1: Entitetstyper og hovedrelasjoner i FRBR. Hentet fra [1].

Den andre gruppen, gruppe 2, består av de entitetene som er ansvarlige for det intellektuelle eller kunstneriske innholdet, for den fysiske produksjonen eller for å ta vare på slike arbeider. Innenfor gruppe 2 er person og korporasjon. Person er et individ, og en korporasjon er en organisasjon eller en gruppe av individer og/eller organisasjoner. Figur 2.2 viser typen ansvarsrelasjoner som eksisterer mellom entitetstypene i gruppe 1 og gruppe 2.



Figur 2.2: Entitetstyper og ansvarsrelasjoner i FRBR. Hentet fra [1].

2.3.2 Entitetstyper

Verk er en abstrakt entitet, og den kan beskrives som et selvestendig intellektuelt eller kunstnerisk arbeid. Et verk i seg selv er innholdet eller det intellektuelle arbeidet som ligger til grunn for de forskjellige uttrykkene av verket. Det kan være vanskelig å definere grenser for hva som er et verk og hva som ikke er et verk. Flere faktorer spiller inn når man skal bestemme hva som er et verk, og vurderingen kan være forskjellig fra kultur til kultur, eller mellom nasjonale grupper. For enkelhets skyld vil variasjoner av teksten, i form av for eksempel revisjoner eller oppdateringer av en tidligere tekst være sett på som uttrykk av samme verk. Dette vil også gjelde for forkortelser eller utgivelser av samme tekst. Oversettelser, dubbede versjoner av filmer, musikalske transkripsjoner og arrangementer er alle også forskjellige uttrykk av det samme originale verket[1].

Større modifikasjoner av et verk som omfatter en betydelig grad av uavhengighet, intellektuelt eller kunstnerisk arbeid, vil bli sett på som et nytt verk. Eksempler på dette kan være gjendiktning, bearbeidelser for barn, parodier eller frie transkripsjoner av et musikkverk[1].

Uttrykk er den intellektuelle eller kunstneriske realiseringen av et verk i form av alfa-numerisk, musikalsk eller koreografisk notasjon, lyd, bilde, gjenstand, bevegelse og så videre. Et uttrykk er den spesielle intellektuelle eller kunstneriske formen verket får før det virkeliggjøres. Når et verk blir realisert gjennom spesifikke ord, setninger eller avsnitt er resultatet uttrykket. Forskjellige fysiske former som for eksempel skrifttype og layout vil ikke i seg selv skille to uttrykk av samme verk fra

hverandre, men i den grad formen er et karaktertrekk ved uttrykket vil endringen i form resultere i et nytt uttrykk. Ved å definere uttrykk kan intellektuelle eller kunstneriske innhold, som kan forekomme fra én realisasjon til en annen av samme verk, beskrives. Uttrykk kan brukes til å identifisere for eksempel den spesielle versjonen som er grunnlaget for en oversettelse[1].

Manifestasjon er den konkrete utformingen av et uttrykk av et verk. Dette omfatter et bredt spekter av materialer som for eksempel manuskripter, bøker, kart, plakater, lydopptak, filmer, videoopptak, noter, programvare og så videre. Manifestasjonene representerer alle fysiske gjenstander som har de samme karakteristikker som både intellektuelt innhold og fysisk form. Grensene mellom manifestasjoner baserer seg på både intellektuelt innhold og fysisk form. Fysisk form i dette tilfellet gjelder fremvisningsaspekter, som endring av skrifttype eller skriftstørrelse, endring i fysisk materiale og endring av informasjonsbærende medium[1].

Eksemplar er et enkelt eksemplar av en manifestasjon. Dette kan for eksempel være den fysiske boken som blir lest, eller den fysiske CDen som blir hørt på. Et eksemplar kan ses på som et eksempel på en manifestasjon, som regel lik manifestasjonen selv[1].

Person omfatter både nålevende og avdøde individer. Personer vil i hovedsak være individer som har vært involvert i å skape eller fremføre et verk, for eksempel forfattere, kunstnere, komponister og så videre. Personer som er emne for et verk vil også bli beskrevet, for eksempel i verk som omhandler personen i biografier.[1].

Korporasjon er en organisasjon eller en gruppe individer og/eller organisasjoner som opptrer som en enhet. Dette vil gjelde tilfeldige og etablerte grupper som møter, konferanser, festivaler, utstillinger og så videre. Korporasjoner vil, i likhet med personer, være grupper som har vært involverte i å skape eller realisere et verk, eller som er emne for verket.

FRBR inneholder langt flere funksjoner enn det som er beskrevet her, men for denne oppgaven er det entitetstypene som er beskrevet over som er mest relevante.

2.4 FRBRisering

Forskjellige versjoner av MARC-formatet er de vanligste formatene for katalogisering av bibliografiske data. MARC er et gammelt format som kan være til hinder for bibliotekene i å tilfredstille nye forventninger og behov hos brukere. Videre finnes det langt flere nye formater og datakilder som skal katalogiseres, for eksempel digitale versjoner som ebøker og streams. Det er derfor ønskelig i noen situasjoner å bruke en ny modell for lagring av bibliografiske data. FRBR-modellen er svaret til IFLA Study Group på utfordringene som omhandler MARC-formatet. FRBR tilbyr mer fleksibilitet for å vise data på alle mulige formater, og gir nye muligheter innen søk og visualisering av data[9]. Prosessen med å overføre data fra en katalogiseringsmodell, som MARC, til FRBR 2.3 kalles FRBRisering. Denne prosessen er blitt vanligere og vanligere de siste årene, og flere store biblioteker jobber med å få dataene sine over til FRBR. Eksempler på biblioteker som har gått over til FRBR er den franske nasjonalbibliografien[10], Cervantes[2] og prosjekter som FictionFinder[11]. Disse eksemplene har også publisert dataene som linked open data. Siden bibliografiske data i hovedsak består av store mengder med poster er dette en prosess som må gjøres automatisk for å være gjennomførbart. Det finnes ingen standardisert FRBRiseringsmetode, og forskjellige biblioteker har valgt forskjellige metoder for å ta i bruk FRBR.

2.4.1 FRBRiseringsprosjekter

Det finnes flere forskjellige metoder for FRBRisering, hver med sine fordeler og ulemper. I artikkelen "A Survey of FRBRization Techniques"[9] blir tidligere brukte FRBRiseringsmetoder klassifisert og sammenlignet for å gi en bedre innsikt i FRBRiseringsprosessen. FRBRiseringsteknikkene blir klassifisert innenfor tre forskjellige kriterier: FRBRiseringstype, hvor ekspressiv modell som er brukt, og forbedringene som er gjort for å øke kvaliteten eller effektiviteten til prosessen. Her er tre eksempler på tidligere FRBRiseringsprosjekter:

Bibliothèque nationale de France

Det franske nasjonalbiblioteket, BnF, lanserte store deler av bibliografien sin som linked open data i 2011. Motivasjonen for prosjektet var å gjøre bibliografiske data og ressurser mer tilgjengelig for brukere og maskiner på nett. Det nye systemet lager automatisk sider om forfattere, verk, tema og så videre etter FRBR-modellen basert på hovedkatalogen. Hovedkatalogen består av det digitale biblioteket *Gallica*. Artikkelen *We grew up together: data.bnf.fr from the BnF and Logilab perspectives*[10] beskriver kort hvordan prosessen har foregått.

De bibliografiske dataene til BnF er på MARC-format og de beskriver hvordan BnF har forberedt seg på den semantiske veven i lengre tid gjennom å generere unike og stabile ARK, arkiverbare ressursnøkler. Ressursnøklene er identifikatorer for bibliografiske autoritetsposter og digitale dokumenter. Videre har biblioteket utviklet autoritetsfiler for forfattere, verk, emner og steder som kan brukes som tilgangspunkter for å organisere tilgang til forskjellige typer ressurser. Eksempler på dette er at relasjonen mellom en bok og forfatteren allerede er spesifisert i MARC. Det er ikke bare data fra hovedkatalogen som er brukt, men også data fra mindre spesialiserte databaser. Disse dataene er ikke nødvendigvis katalogisert på samme måte som dataene i hovedkatalogen, men det jobbes med å automatisere denne prosessen. BnF er også koblet til andre datasett gjennom VIAF.

Selve FRBRiseringsprosessen er ikke beskrevet i detalj, men noen prosesser er beskrevet i [10]. Det nye systemet bruker algoritmer basert på maskinlæringsteknikker som tilrettelegger for sammeligning av store databaser. Prosessen med å skape relasjoner mellom allerede eksisterende autoritetsdata og de tilhørende bibliografiske postene er en blanding mellom en automatisk og en manuell prosess. Først blir relasjonene generert med en sammenslåingsprosess som Simon et al. i sin artikkel hevder er svært pålitelig. Deretter blir resultatene som har lavere pålitelighet gått gjennom av katalogiserere og rettere som finner de riktige relasjonene mellom verk og forfattere. Systemet foreslår mulige koblinger som blir sjekket manuelt. Til slutt brukes clustering-algoritmer for å skape autoritetsdata der det ikke eksisterer. For eksempel vil det bli laget verk som blir relatert til eksisterende manifestasjoner. Utover dette er det ikke beskrevet noen tiltak som er gjort for å vurdere og kontrollere kvaliteten på dataene og transformasjonsprosessen.

Biblioteca Virtual Miguel de Cervantes

Biblioteca Virtual Miguel de Cervantes, Cervantes, er en katalog som inneholder 200 000 poster som har blitt overført fra MARC21-formatet til en ny relasjonsdatabase som er modellert basert på FRBR. Dataene er lagret som RDF tripler som bruker RDA-vokabularet til å beskrive entiteter, deres egenskaper og relasjoner. Dataene er blitt publisert som linked open data.

Cervantes begynte med å lage en relasjonsdatabase basert på FRBR, FRAD og FRSAD. FRBR definerer entitetene verk, uttrykk, manifestasjon og gjenstand. FRAD definerer person, korporasjon og familie. FRSAD definerer konsept, objekt, hendelse og sted. Deretter ble det startet en automatisk

prosess for å transformere de gamle postene fra MARC21-format til nye oppføringer i FRBR i den nye databasen. Systemet bruker flere forskjellige teknologier med åpen kildekode i transformeringen. For eksempel ble Hibernate ORM[1] brukt til å definere koblingen mellom egenskaper og kolonner, og klasser og tabeller. Dataene ble deretter lagt inn i objektfelt i RDA-komponentene for RDF grafen ved hjelp av The Apache Jena library[2]. Cervantes tok også i bruk andre vokabularer enn RDA, hvis relasjoner ikke kunne bli beskrevet av RDA-vokabularet.

Biblioteca Nacional de España

Bibliografien til Biblioteca Nacional de España, BNE, ble lansert i 2011 som linked open data etter prosjektet *Linked Data at the BNE*. Prosessen er beskrevet i artikkelen *datos.bne.es: a Library Linked Data Dataset*[12] og *datos.bne.es and MARiMbA: an insight into library linked data*[13]. Dataene som har blitt transformert i dette prosjektet er autoritetsposter som består av metadata som beskriver personer, organisasjoner, verk-titler og emner, og bibliografiske poster. Disse postene er deler av BNE-katalogen og inneholder metadata om både moderne og eldre gjenstander som kart, bilder, innspillinger, manuskripter og så videre.

BNF har brukt MARiMbA[3] til transformeringen av MARC21-poster til RDF. MARiMbA er en prosess som gjør det mulig for personer å manuelt koble MARC21-poster til de riktige RDFS- og OWL-klassene. Systemet går først gjennom alle postene og kartlegger hvilke entiteter de tilhører gjennom en *annotasjonskartlegging*, og deretter blir relasjonene mellom entitetene definert gjennom *relasjonskartlegging* basert på informasjonen i MARC21-posten. MARiMbA bruker eksterne koblinger for å finne like poster i andre datasett, disse datasettene er VIAF, GND, DBpedia, Libris og SUDOC. Til sammen ble det dannet over en halv million koblinger fra poster i BNF til like poster som eksisterer i en av disse eksterne datasettene. Disse koblingene er viktige i valideringen av innholdet.

Når MARC-postene skal overføres til RDF-ressurser brukes MARC-feltene og underfeltene til å definere de ulike typene entiteter i FRBR de skal gjøres om til. Deretter blir relasjonene definert gjennom å bruke feltene i den opprinnelige MARC-posten. Hvis det for eksempel eksisterer et verk og en manifestasjon som omhandler det samme verket, vil systemet generere et uttrykk som blir koblet til verket og manifestasjonen. På den måten kan de nye entitetene og relasjonene bekreftes manuelt ved hjelp av bibliotekarer.

For BNE har både kvaliteten på datakilden og dataene som har blitt publisert vært viktig. I [13] beskriver de hvordan kvaliteten på de originale dataene har en direkte innvirkning på kvaliteten til dataene som har blitt generert. Derfor har de gjennom hele prosessen ikke bare jobbet med selve transformasjonsprosessen, men også gått gjennom kvaliteten på originaldataene. Kvaliteten til RDF dataene har blitt validert ved å bruke retningslinjene for prinsipper i linked data beskrevet i *An empirical survey of Linked Data conformance* [14]. For originaldataene har de gjennomført flere retningslinjer for hvordan de skal håndtere problemene som har oppstått. Disse problemene har for eksempel vært kodefeil i systemet som har produsert gale koblinger mellom MARC-felt og RDF-dataene. Disse feilene ble identifisert av bibliotekarene og rettet opp i. Dette er fordelen med manuell sjekking av dataene som blir produsert. Utover problemer som oppstod under transformeringen er det ikke nevnt noen metode for evaluering av kvaliteten på resultatet. Siden dataene er manuelt sjekket av bibliotekarer vil resultatet kunne ha bedre kvalitet enn et helt automatisk system.

[1] http://hibernate.org/orm
[2] https://jena.apache.org
[3] http://marimba4lib.com

Kapittel 3

Teori

Dette kapittelet går inn på relevante teorier rundt datakvalitet. Først blir kvalitet generelt beskrevet og det blir sett på datakvalitet i databaser og datavarehus. Til slutt presenteres datakvalitet i linked open data, og forskjellige metrikker som brukes for å vurdere kvaliteten til data som er publisert i linked open data-skyen.

3.1 Datakvalitet

Datakvalitet er et viktig aspekt i all type databehandling, om det gjelder knowledge extraction, datatransformering eller informasjonsgjenfinning. ISO9000[15] sin definisjon av kvalitet er *i hvilken grad et sett med egenskaper oppfyller et krav*. Dette betyr at graden av kvalitet avhenger av bruksområdet som dataene skal brukes til. Dette gjelder datakvalitet også. For å kunne måle datakvalitet er det derfor viktig å ha et klart bilde av hva dataene skal brukes til, fordi man ikke kan måle kvaliteten ved å kun se på dataene for seg selv. Et relevant eksempel er hvordan bibliografiske data på MARC-formatet i dag kan ha god kvalitet og være semantisk sterkt i dette formatet. Når man så skal overføre data fra MARC-formatet til et sterkere semantisk system, som FRBR, oppfylles imidlertid ikke de nye kravene og man kan si at datakvaliteten er dårlig hvis ingen tiltak blir gjort.

I databaseverdenen er det vanlig å snakke om for eksempel normalisering for å bedømme kvaliteten til databasen. I dette arbeidet er det mer interessant å se på kvalitet innenfor knowledge extraction fordi oppgavens mål er å se på knowledge extraction i semistrukturerte data.

Populariteten rundt knowledge extraction har økt i det siste og mye på grunn av at flere og flere ønsker å publisere dataene sine som linked open data, beskrevet i 3.2. Utfordringen rundt dagens gjennomføring av publisering av linked open data er at motivasjonen ikke alltid er fokusert på et godt resultat, men i flere tilfeller bare et resultat. Flere og flere publiserer data som linked open data, men det er lite fokus på at kvaliteten til det som blir publisert er god. Flere har beskrevet problemene med uryddig data som blir publisert som linked open data, for eksempel Schlobach og Knoblock i "Dealing with the Messiness of the Web of Data"[16].

Publisering av store data på nett innebærer mange muligheter, men dataene er aldri mer nyttige enn det kvaliteten på dataene er. For å få det beste resultatet, er det nødvendig at dataene har god kvalitet før prosesseringen av dataene[17].

For denne oppgaven er datakvalitet innenfor linked open data mest interessant, men i mange tilfeller er det flere konsepter å hente fra andre felt innen datakvalitet. Derfor vil datakvalitet innenfor

databaser og datavarehus først presenteres, og deretter beskrives kvalitet innenfor linked open data.

3.1.1 Datakvalitet i databaser

Når det er snakk om datakvalitet er datakvalitet i databaser et viktig tema. Flere av prinsippene innenfor kvalitetssikring og kvalitetsvurdering i databaser har mye til felles med andre tema som for eksempel knowledge extraction. Innenfor databaser er det stort fokus på effektivitet, både når det gjelder hastigheter til å lese og å skrive poster, men også med tanke på plass. Derfor vil en database med god kvalitet ha korte skrive- og lesehastigheter, og ikke inneholde unødvendig eller overflødig informasjon. Det finnes flere konsepter som skal sikre at en database har god kvalitet. På modellnivå bruker man normalisering for å forhindre for eksempel overflødig informasjon. På postnivå kan man gjennomføre deduplisering for å fjerne duplikater i databasen. Normalisering må ligge til grunn for å ha god kvalitet i databasen. Duplisering blir brukt som et oppryddningsverktøy i en rotete database som har lav grad av normalisering.

Normalisering

Normalisering er en prosess for å organisere strukturen i en database på en måte som reduserer overflødighet og forbedrer dataintegritet. Normalisering er en måte å optimalisere en database og forenkle designet. Hvis samme data kun er lagret ett sted i databasen, reduserer det sannsynligheten for inkonsistente data.

I hvilken grad en database er normalisert defineres etter hvilken normalform databasen oppfyller. En normalform er et sett med regler som må oppfylles, og jo flere normalformer som er oppfylt jo mer normalisert vil databasen være. For å oppfylle første normalform må alle attributtene i databasen kun inneholde atomære, ikke-oppdelelige verdier. Det eksisterer mange forskjellige normalformer i et hierarki, der én normalform bare kan oppfylles dersom alle normalformene tidligere i hierarkiet er oppfylt.

Normalisering er et konsept som stiller strenge krav til databasemodellen og en normalisert database vil ha langt flere tabeller enn en database som ikke er normalisert. Derfor vil ytelsen til en normalisert database kunne være lavere fordi det må gjøres flere join-operasjoner.

Deduplisering

Deduplisering er prosessen å fjerne duplikater. Duplikater er data som allerede eksisterer eller som det finnes identiske kopier av. I databaser er det snakk om to poster som er identiske. Duplikater er overflødig informasjon som ikke er ønskelig på grunn av at det tar opp unødvendig plass og at det kan skape inkonsistent data. Hvis to poster inneholder samme informasjon som må endres, og kun den ene posten blir endret, risikerer man at den uendrede posten blir brukt senere og da med feil informasjon.

Deduplisering kan enten gjøres fortløpende samtidig som data blir skrevet til databasen, eller det kan gjøres etter at dataene er lagret. Den vanligste måten å deduplisere data på er å gi alle postene en identifikator eller nøkkel basert på informasjonen i posten. Deretter blir identifikatoren hashet, og dermed vil alle identiske poster ha samme hashede identifikator slik at man kan fjerne duplikatene. Denne prosessen kan også fungere for poster som har flere identiske felt, men ikke alle. Hvis identifikatoren blir generert på bare noen av feltene, vil poster som har identiske felt for de feltene som brukes til å generere identifikatoren merkes som duplikat. Dette kan være nyttig dersom

postene ikke inneholder all den samme informasjonen, og man ikke ønsker å miste informasjonen i en sammenslåing.

3.1.2 Datavarehus

Datavarehus inneholder enorme datamengder og er mye brukt i prosessering og håndtering av data i forbindelse med for eksempel *Big Data*. I prosjekter som henter inn store mengder data er datavasking en viktig del av prosessen. I de fleste datavarehus-prosjekter står datavasking for mellom 30 og 80 prosent av utviklingstiden og budsjettet.[18]. I datavarehus har små feil en tendens til å forsterkes. Små feil som menneskelige innsettelsesfeil, overflødige data og korrupte data påvirker verdiene til datavarehuset. På grunn av den store mengden data er det viktig å fjerne eventuelle feil raskt så de ikke sprer seg og skaper store problemer.

Innenfor datakvalitet i datavarehus er det flere steg som kan gjøres for å sikre kvaliteten. For eksempel kan man lage regler for kvaliteten, sjekke for inkonsitens og reparere. Disse stegene må kunne skaleres på grunn av den store mengden data, og det er derfor ofte snakk om å ta et valg mellom *effektivitet versus nøyaktighet*. Big data og datavarehus kan kategoriseres basert på tre V-er: Volume (mengde data), velocity (hastighet på data inn og ut) og variety (forskjellige datatyper og kilder). Disse egenskapene vil alle kunne gjøre jobben med datakvalitetssikring vanskeligere.

På grunn av den store variasjonen av data som kan være i et datavarehus er det svært vanskelig å lage regler for kvaliteten som gjelder hele samlingen, og det vil derfor være nødvendig å lage regler for bare deler av samlingen. Innenfor datavarehus er det noen metrikker som er viktige for å definere god datakvalitet. Disse metrikkene går igjen i mange forskjellige datasystemer og gjelder også databaser og linked open data. Datavarehus må representere entiteter på en konsistent, nøyaktig, komplett, betimelig og unik måte.

3.1.3 Extract, Transform and Load

Extract, Transform and Load, ETL, er prosessen å hente ut data fra eksterne kilder, bearbeide dem og laste dem inn i et system som ofte er et datavarehus. I ETL er *transform*-delen viktig fordi dataene som blir ekstrahert kan ha varierende kvalitet, og det vil derfor være nødvendig med en form for opprydding. Datakvalitet innenfor ETL har flere likheter med datakvalitet i databaser og knowledge extraction.

Artikkelen "The Value of ETL and Data Quality" omhandler utfordringer og prinsipper innenfor datakvalitet i ETL[19]. Det første og viktigste prinsippet er at kvaliteten på originaldataene er god. Henter man inn data av lav kvalitet vil dette forplante seg videre inn i datavarehuset og andre relaterte systemer. Videre er standardisering et viktig prinsipp. Formatet og innholdet i dataene må standardiseres for at dataene skal være enklere å finne og også for enklere å hindre duplikater. Hvis to poster har samme innhold, men er lagret med forskjellige formater på for eksempel navn eller adresse, vil det være vanskeligere å identifisere disse duplikatene.

Til slutt er validering av data en effektiv prosess for å sikre kvaliteten. Validering kan gjøres manuelt, men dette er en tidkrevende prosess og det er ønskelig å gjøre det så automatisert som mulig. Et eksempel på en enkelt automatisert validering kan være å slå opp for eksempel postnummer og adresser mot Posten eller en karttjeneste for å se om de stemmer overens.

3.2 Linked Open Data

Linked Open Data, eller lenkede åpne data på norsk, er data som er åpne i den forstand at de kan brukes, endres og blir re-distribuert av hvem som helst, og de er lenkede fordi de inneholder informasjon om hvordan de er relatert til annen data[2]. Når man snakker om linked open data handler det i hovedsak om å gjøre dataene tilgjengelig på formater som kan leses både av mennesker og maskiner.

Det er mange fordeler ved å publisere bibliografiske data som linked open data. Flere av disse er oppsummert i "W3C Library Linked Data Incubator Final Report"[20]: Det skaper en forbedret måte å navigere kulturell informasjon på, det øker mulighetene for kulturelle data på Internett, det gir en mer varig og robust semantisk modell enn metadata som avhenger av spesifikke datastrukturer, det legger til rette for gjenbruk av data på tvers av kulturarver, og det tillater muligheten for å unngå bibliotekssspesifikke formater, som for eksempel MARC.

3.2.1 Datakvalitet i linked open data

Linked open data og FRBR er to termer som ofte blir nevnt sammen i bibliotekssammenheng, og det er flere gode grunner til å kombinere de to. Data som bruker FRBR-modellen kan dra nytte av annen informasjon om relasjoner som eksisterer som linked open data. På denne måten vil linked open data bidra til å øke kvaliteten til dataene som er på FRBR, og disse dataene vil igjen kunne øke kvaliteten til nye data. Linked open data kan også brukes til å verifisere entiteter som brukes i FRBR, dette gjelder blant annet forfattere, verk og manifestasjoner. Ved riktig bruk av linked open data vil relasjonen mellom disse bli identifisert og verifisert[21].

I motsetning til tidlig bruk av semantisk web som inneholdt små mengder data som ble nøye gjennomgått og validert, består de nyere systemene som linked open data av enorme mengder data som er ukomplette, inkonsistente, gale og skiftende[16]. Scholbach og Knoblock beskriver *the Web of Data* som rotete og vanskelig å bruke, og de jobber med hvordan man skal løse dette problemet[16]. Det er flere andre som også prøver å finne en løsning på den dårlige kvaliteten blant data som blir publisert som linked open data. I [16] deles problemet opp i to klasser der den første klassen handler om å lage retningslinjer for å unngå rot i dataene før de i det hele tatt blir publisert. Den andre klassen handler om å gi brukerne infrastruktur og teknikker for å bygge nyttige teknologier som tar i bruk dataene til tross for den dårlige kvaliteten.

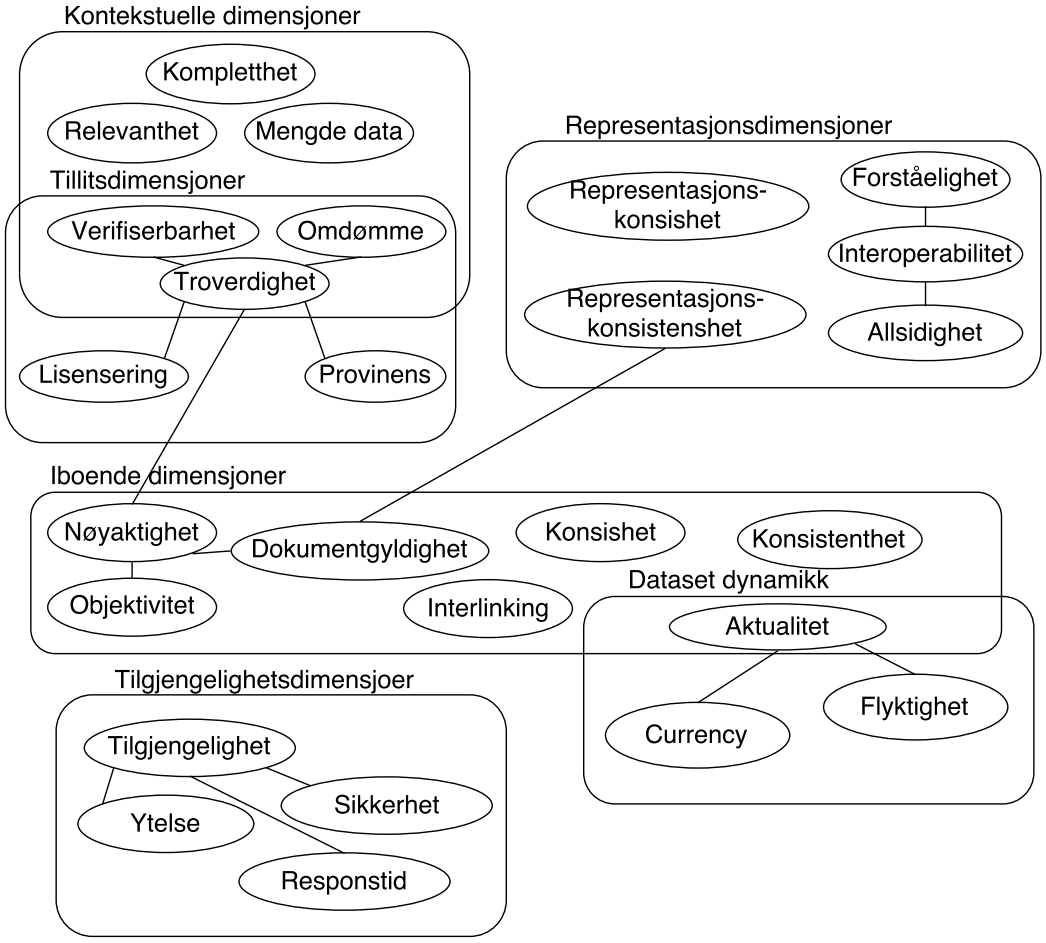
På bakgrunn av dette kan man se hvordan data som blir publisert som linked open data har kvalitetsproblemer som trenger å bli tatt hånd om. Det trengs mer fokus på retningslinjer og teknologier som kan bedre kvaliteten på dataene før de blir publisert, men også systemer som kan rydde opp i data som allerede er blitt publisert. I neste del blir det beskrevet flere metrikker for datakvalitet som vil gjøre det enklere å vurdere kvaliteten til dataene før de blir publisert.

3.3 Metrikker for datakvalitet

God datakvalitet er i mange sammenhenger subjektivt, og det er ikke alltid like lett å definere hva som nettopp er god datakvalitet. Datakvalitet blir ofte oppfattet som tilpasset til dataene, men selv data med lav kvalitet kan være nyttig for visse bruksområder[22]. Det er derfor viktig å definere innenfor hvilke rammer kvaliteten skal vurderes og hvilke egenskaper som gjør at datakvaliteten er god[23]. *Quality Assessment Methodologies for Linked Open Data*[22] er en systematisk gjennomgang av litteratur som omhandler datakvalitet i Linked Open data. De har tatt for seg 21 forskjellige

artikler og laget et konseptuelt rammeverk for metrikker innenfor datakvalitet. Metrikkene er også høyst relevante for denne oppgaven, og for datakvalitet i knowledge extraction.

I "Quality Assessment Methodologies for Linked Open Data"[22] har de identifisert 26 forskjellige metrikker som kan deles inn i 6 forskjellige dimensjoner. Disse dimensjonene er kontekstuelle, iboende, tillits-, tilgjengelighets- og representasjonsdimensjoner, og datasett-dynamikk. Av de 26 metrikkene som er nevnt i [22], er det fem metrikker som oftere er nevnt i litteraturen. Metrikkene er proveniens, konsistens, aktualitet og nøyaktighet. Disse metrikkene har også flere likhetstrekk med noen av de viktigste datakvalitet-kategoriene som er beskrevet i ISO9000:2015[15], som er fullstendighet, validitet, nøyaktighet, konsistens, tilgjengelighet og aktualitet. Dimensjonene som er relevante for denne oppgaven vil nå bli presentert. Figur 3.1 viser alle metrikkene og dimensjonene som er beskrevet i [22], strekene mellom dimensjonene betyr at de er relatert til hverandre.



Figur 3.1: Datakvalitetsmetrikker fra *Quality Assessment Methodologies for Linked Open Data* og relasjonene mellom dem.

3.3.1 Kontekstuelle dimensjoner

Kontekstuelle dimensjoner er dimensjoner som er avhengig av i hvilken sammenheng informasjonen skal brukes.

Fullstendighet

Graden av hvorvidt informasjon ikke mangler. Dette betyr i hvilken grad all nødvendig informasjon er tilgjengelig. Hvilken informasjon som må til for å gjøre en post fullstendig kommer an på type post, og i hvilken sammenheng posten opptrer. Generelt kreves det at informasjonen har en tilstrekkelig dybde, bredde og omfang for oppgaven som er gjeldende. Fullstendighet handler også om fullstendigheten til hele katalogen. Det betyr at all informasjonen som er forventet å være der bør være tilstede. For eksempel vil en database ha lav fullstendighet dersom databasen skal vise til verkene til Agatha Christie, men bare inneholder halvparten av verkene.

Mengde data

Graden av hvorvidt volumet av data er passende for oppgaven som skal gjøres. Mengden data som er tilgjengelig påvirker brukervennligheten, og den må være tilstrekkelig for å kunne gjennomføre oppgavene som er forventet. Hva som er en tilstrekkelig mengde data, baserer seg på systemet det er snakk om og hva det skal brukes til. Bibliografiske data representerer innholdet til et bibliotek eller lignende, så den forventede datamengden bør være tilstrekkelig til å representere innholdet i biblioteket. Dette er et minimum av det som forventes, og mer data som beskriver gjenstandene og entitetene i biblioteket, samt relasjonen mellom dem, øker kvaliteten til systemet.

Relevanse

Graden av hvorvidt informasjonen er aktuell for oppgaven som skal gjøres. Dataene må inneholde informasjonen som er nødvendig for det dataene skal brukes til.

3.3.2 Tillitsdimensjoner

Tillitsdimensjonene er dimensjonene som fokuserer på troverdigheten til dataene.

Proveniens

Den kontekstuelle metadataen som fokuserer på hvordan informasjon skal presenteres, administreres og brukes om opprinnelsen til kilden. Proveniens handler om å vite hvor dataene kommer fra, altså kilden til datene.

Verifiserbarhet

Graden av hvor enkelt det er å verifisere dataene betyr hvor enkelt det er for brukeren å verifisere riktigheten til dataene. Det er viktig at det er enkelt og tydelig for brukerne å kunne vurdere kvaliteten til dataene. Det er flere faktorer som spiller inn i denne vurderingen, blant annet kilden til dataene, hvordan dataene har blitt behandlet og kvaliteten til dataene før en eventuell behandling.

Omdømme

Omdømme er vurderingen fra en bruker for å bestemme integriteten til en kilde. I hovedsak omhandler dette en utgiver, en person, en organisasjon, en gruppe mennesker eller et felleskap, mer enn en karakteristikk av selve dataene. Det er viktig at det er mulig å identifisere utgiveren.

For bibliografiske data kan biblioteket som publiserer dataene bli sett på som utgiver. Et anerkjent bibliotek vil ha et bedre omdømme enn et mindre anerkjent bibliotek, og dataene som blir publisert vil derfor vurderes til å ha høyere kvalitet.

Troverdighet

Graden av hvorvidt dataene er akseptert som å være riktige, sanne, ekte og troverdige. Troverdighet henger sammen med de andre metrikkene under tillitsdimensjonen. For at data skal ha høy troverdighet er det viktig at kilden til dataene har godt omdømme og at det er mulig for brukeren å verifisere denne kilden.

3.3.3 Iboende dimensjoner

Iboende dimensjoner er dimensjonene som er uavhengig av brukerens kontekst. Disse dimensjonene fokuserer på hvorvidt informasjonen representerer verden, og om den er konsistent med seg selv.

Nøyaktighet

Graden av hvorvidt informasjonen representerer verden med riktighet og presisjon, og uten feil. Dataene må inneholde verdier som stemmer med de originale verdiene. Nøyaktighet gjelder både for syntaktisk nøyaktighet, og semantisk nøyaktighet.

For bibliografiske data er det viktig at dataene representerer informasjonen korrekt i form av riktig innhold, at de er riktig identifisert og at relasjonene mellom dataene er riktig. Unøyaktige data vil kunne forplante seg og bli et større problem hvis de relateres til andre data.

Dokumentgyldighet

Gyldigheten til brukbarheten til dokumentet, hvordan det har brukt vokabularer og at dette er korrekt, og riktig bruk av syntaks. For bibliografiske data går gyldigheten også på riktig bruk av felt, entiteter og relasjoner. Dokumentgyldighet gjelder også mer generelle instanser som formatet dataene er lagret på, og at dette må være riktig formatert og ha riktig syntaks.

Duplikater

Graden av hvorvidt det eksiterer identiske entiteter. Dette innebærer om det eksisterer duplikater i dataene eller ikke. Duplikater er støy og skaper problemer for brukere og systemet i seg selv. For bibliografiske data er duplikater en utfordring på flere forskjellige nivå. Duplikater kan være data som inneholder den samme informasjonen, eventuelt med små variasjoner som skrivefeil. Duplikater kan også være poster som ikke nødvendigvis inneholder den samme informasjonen, men som representerer den samme gjenstanden eller entiteten. Eksisterer det to poster som representerer samme forfatter, vil verkene som er relatert til denne forfatteren for eksempel kunne bli splittet i to grupper. Den ene posten av forfatteren kan inneholde informasjon som den andre posten ikke inneholder, og omvendt.

3.3.4 Interlinking

Graden av hvorvidt entiteter som representer den samme informasjonen er sammenslått. Dette betyr enten at entiteter som representerer den samme informasjonen er markert med for eksempel en *sameAs-relasjon* eller entitetene kan identifiseres med lik id. I linked open data er ikke duplikater nødvendigvis et stort problem, så lenge duplikatene er markert og man vet hvilke entiteter som representerer den samme informasjonen.

Konsistens

Hvorvidt informasjon ikke motsier annen informasjon. Dette gjelder både innad i ett datasett, men kan også bety at informasjonen ikke motsier informasjon fra andre datasett.

Konsishet

Hvorvidt data ikke inneholder overflødig informasjon. To forskjellige attributter som inneholder samme informasjon regnes som overflødig informasjon.

3.3.5 Tilgjengelighets-dimensjoner

Tilgjengelighets-dimensjonene er tilgjengelighet, ytelse, sikkerhet og responstid. Dette er viktige metrikker for brukbarheten til et system, og har mye å si for brukerne. Disse metrikkene handler mer om systemet som blir brukt, og ikke nødvendigvis om dataene som er i systemet i seg selv. De er derfor ikke beskrevet i mer detalj her.

3.3.6 Representasjonsdimensjoner

Representasjonsdimensjonene omhandler representasjonskonsistens, forståelighet, interoperabilitet og allsidighet. Disse metrikkene omhandler hvordan dataene er presentert og håndtert. Fordi målsetningen for oppgaven er å se på dataene i seg selv, er ikke disse metrikkene relevante for denne oppgaven.

3.3.7 Datasett-dynamikk

Metrikkene som inngår i datasett-dynamikk er samtidighet, flyktighet og aktualitet.De handler om innholdet i dataene og hvordan det skal være oppdatert og aktuelt for å være nyttig for brukeren. Disse metrikkene er mindre relevante for denne oppgaven, og er derfor ikke beskrevet i mer detalj.

3.4 Datakvalitet i denne oppgaven

I denne oppgaven er også datakvalitet viktig. Målet er å kunne lage en metode som kan brukes for å bedre kvaliteten til bibliografiske data som bruker FRBR-modellen og som blir publisert som linked open data. I databaseverdenen er normalisering viktig for å unngå overflødighet og for å sikre at postene er unike. I datavarehus er mengden data en utfordring, og man må gjøre et valg mellom effektivitet og nøyaktighet. Disse konseptene er også viktige i oppgaven. Utfordringen til mange linked open data-systemer er at de består av store mengder data med mye duplikater og overflødig informasjon. Flere skritt kan tas for å bedre kvaliteten til biblioteksdata ved å hindre duplikater i å bli publisert.

Dette prosjektet ser på muligheten for å kunne identifisere verk med høy sannsynlighet for at de er riktig identifisert. Gjennom en slik metode vil det være mulig å unngå duplikater ved å slå sammen de postene som representerer den samme entiteten, eller ved å koble de sammen med en relasjon. Videre er det flere metrikker innenfor datakvalitet i linked open data som er relevant for denne oppgaven. Fullstendighet er viktig for å kunne ha den riktige informasjonen om en post som er nødvendig for å identifisere verket. Videre vil flere poster gjøre sannsynligheten større for at et verk kan identifiseres. Metrikker som proveniens og verifiserbarhet er viktige for å kunne stole på dataene som blir jobbet med, og vil kunne sikre representative resultat. Det viktigste for kvaliteten i dette prosjektet er at dataene er korrekte, og ikke inneholder duplikater. Postene må være riktig identifisert som verk, og det må ikke eksistere flere poster som representerer det samme verket.

3.5 Validering

Det er nødvendig med god kvalitet på dataene som skal bli analysert, men det er også viktig at dataene har god kvalitet etter at de er blitt prosessert. Dette kan man sikre ved å validere dataene. En slik prosess vil kunne være knowledge extraction, informasjonsgjenfinning eller i dette tilfellet FRBRisering og verkidentifikasjon. For å kunne validere resultatene må man først vite hva som er korrekt, og deretter undersøke resultatene.

Det er vanlig å dele inn resultatposter i for eksempel informasjonsgjenfinning inn i fire grupper. Den første gruppen er de relevante postene som ble gjenfunnet, den andre er de relevante postene som ikke ble gjenfunnet, den tredje er irrelevante poster som ble gjenfunnet, og den siste gruppen er de postene som er irrelevante og som ikke ble gjenfunnet. Det er vanlig å visualisere dette i en forvirringsmatrise som vist i Tabell 3.1. Tabellen viser også hvordan de forskjellige postene blir markert basert på om de er relevante, gjenfunnet og så videre.

	Gjenfunnet	Ikke-gjenfunnet
Relevant	True Positive(TP)	False Negative(FN)
Irrelevant	False Positive(FP)	True Negative(TN)

Tabell 3.1: Forvirringsmatrise.

3.5.1 Precision

Precision er en av de vanligste metodene for å validere et informasjonsgjenfinningssystem, og sier noe om hvor mange av de gjenfunnede postene som er relevante. En annen måte å formulere dette på er at precision er sannsynligheten for at posten som er hentet ut er relevant. Formelen for utregning av precision er:

$$\frac{|TruePositive|}{|TruePositive + FalsePositive|}$$

Precision sier noe om hvor mange av de relevante postene som blir gjenfunnet, og irrelevante poster vil regnes som støy i resultatene. Dette er ikke ønskelig og derfor kan precision brukes som et mål på hvor mye av informasjonen som blir gjenfunnet som er interessant.

3.5.2 Recall

Precision fungerer bra for å få finne ut av riktigheten til postene som er blitt hentet ut. Problemet er at antall poster som ikke er hentet ut ikke er kjent. Recall brukes for å synliggjøre hvor mange av de relevante postene som har blitt hentet ut, med andre ord sannsynligheten for at en gitt relevant post vil bli hentet ut. Formelen for recall er:

$$\frac{|TruePositive|}{|TruePositive + FalseNegative|}$$

3.5.3 Accuracy

Accuracy sier noe om hvor nøyaktig gjenfinningen av poster er både med tanke på at de riktige postene blir identifisert som riktige, og at de gale postene blir identifisert som gale. Formelen for accuracy er:

$$\frac{|TruePositive + TrueNegative|}{|TruePositive + TrueNegative + FalsePositive + FalseNegative|}$$

Kapittel 4

Analyse av dagens kvalitetsstatus

Dette kapittelet inneholder en analyse av eksisterende systemer som har publisert bibliografiske data til linked open data-skyen. Det er et kjent problem at data som er publisert som linked open data har kvalitetsproblemer. Denne analysen skal undersøke hvordan landskapet ser ut.

4.1 Metode

Målet med denne analysen er å få et overordnet blikk på noen av resultatene som har blitt produsert etter FRBRisering. Analysen er en sjekk av kvaliteten til dataene som er publisert som linked open data og som blir presentert for brukeren. I hovedsak er det metrikkenes riktighet og støy som er blitt analysert. Støy omhandler duplikater og feilidentifiserte poster.

Analysemetoden går ut på å gjøre oppslag på kjente forfattere i de forskjellige systemene, for så å gjøre en manuell sjekk av resultatet. Systemene som var med i analysen var FictionFinder[11], Bibliothèque nationale de France[10], Cervantes Virtual[2] og Deutsche Nationalbibliothek[24]. Analysen gikk ut på å finne verkene som er identifisert og relatert til de forskjellige forfatterne som er blitt valgt ut. I tillegg har det blitt gjort søk på ett kjent verk av hver forfatter for å gi et bedre inntrykk av mulige duplikater og støy per verk.

Videre ble resultatet analysert basert på antall treff, riktigheten av disse treffene, og hvordan de er blitt identifisert. Videre ble det sett på om noen poster for eksempel var identifisert som verk, og hvordan de stod oppført, for eksempel om poster var gruppert. Resulterte søket i mange treff var det viktig å gå gjennom flere av treffene for å se hvordan statusen var blant alle treffene. Videre ble det gjort en analyse på verk. Det ble søkt på et kjent verk av de samme forfatterene. Dette ble gjort for å se om søk etter spesifikke verk ga noen andre resultater enn forfattersøket. Når søk etter verk blir gjort, er det enklere å finne duplikater og andre feilidentifiserte poster.

Denne prosessen ble gjennomført for fem forskjellige forfattere med fem forskjellige nasjonaliteter. De utvalgte forfatterne ble valgt basert på ulike faktorer. Den første faktoren var at forfatterne skulle ha forskjellig nasjonalitet og verk på originalspråk. Dette var for å se hvordan de forskjellige systemene håndterer verk på forskjellige språk, og om nasjonalitet har noe å si på riktigheten til resultatene. Forfatterne ble også valgt på grunn av hvor populære de er både nasjonalt og internasjonalt. Det var ønskelig å se på forfattere som har skrevet verk som har blitt oversatt til flere språk. Videre var det viktig at forfatterne hadde produsert flere verk slik at sannsynligheten for flere poster var høyere. Forfatterne som ble valgt var den britiske krimforfatteren Agatha Christie, den franske

forfatteren Jules Verne, den spanske forfatteren Miguel de Cervantes, den norske forfatteren Knut Hamsun og den tyske forfatteren Günter Grass.

For å ha et utgangspunkt i verkene som hver forfatter har skrevet ble Wikipedia brukt. Akkurat til denne analysen er Wikipedia en god kilde, men hvis Wikipedia hadde en riktig oversikt over alle verkene til alle forfatterne i verden, ville identifiseringen av verk vært en smal sak. Utfordringen er at Wikipedia har veldig mye og god informasjon om de mest kjente forfatterne. For mindre kjente forfattere vil det være mangelfull informasjon eller ingen informasjon i det hele tatt. For forfatterne som er brukt i denne analysen er Wikipedia et bra oppslagsverk for å vite hvilke verk hver forfatter har skrevet. En av grunnene til dette er den hyppige deltagelsen i endring og retting av artiklene som omhandler kjente forfattere. For eksempel artikkelen "Agatha Christie bibliography"[1] er endret og revidert 285 ganger av 128 forskjellige brukere siden 2009. Den hyppige revideringen vil sikre at informasjonen er riktig, og jo flere mennesker som er med på å endre artikkelen, øker sannsynligheten for at artikkelen ikke inneholder feil. Artikkelen om Knut Hamsun[4] er revidert 2254 ganger siden 2003.

Navn	Nasjonalitet	Antall verk
Agatha Christie	Britisk	73[1]
Jules Verne	Fransk	54[2]
Miguel de Cervantes Saavedra	Spansk	5[3]
Knut Hamsun	Norsk	38[4]
Günter Grass	Tysk	33[5]

Tabell 4.1: Forfatterne som er brukt i analysen.

Verkene som ble valgt for denne analysen var *Murder on the Orient Express* av Agatha Christie, *Le Tour du monde en quatre-vingts jours* av Jules Verne, *El ingenioso hidalgo Don Quixote de la Mancha* av Miguel de Cervantes, *Sult* av Knut Hamsun og *Die Blechtrommel* av Günter Grass. Disse verkene ble valgt fordi de er blant de mest kjente verkene av forfatterne som er brukt i analysen. Verkene er oversatt til flere forskjellige språk, og noen er også blitt filmatisert. Dette gjør at resultatet kan gi ulike poster og gjøre oppgaven i å identifisere postene vanskeligere for bibliotekene.

[1] https://en.wikipedia.org/wiki/Agatha_Christie_bibliography
[2] https://en.wikipedia.org/wiki/Jules_Verne_bibliography
[3] https://en.wikipedia.org/wiki/Miguel_de_Cervantes
[4] https://no.wikipedia.org/wiki/Knut_Hamsun
[5] https://en.wikipedia.org/wiki/Günter_Grass

Originaltittel	Norsk tittel	Engelsk tittel	Forfatter
Murder on the Orient Express	Mord på Orienteks-pressen	Murder in the Calais Coach	Agatha Christie
Le Tour du monde en quatre-vingts jours	Jorden rundt på 80 dager	Around the World in Eighty Days	Jules Verne
El ingenioso hidalgo Don Quixote de la Mancha	Den skarpsindige lavadelsmannen Herr Quijote fra La Mancha	The history of the valorous and wittie Knight-Errant Don-Quixote of the Mancha	Miguel de Cervantes
Sult	Sult	Hunger	Knut Hamsun
Die Blechtrommel	Blikktrommen	The Tin Drum	Günter Grass

Tabell 4.2: Verkene som er brukt i analysen.

Disse forfatterne og verkene er kun valgt ut for å eksemplifisere og beskrive kvaliteten til forskjellige systemer. Resultatene vil gi en indikasjon på kvaliteten.

4.2 Resultater

4.2.1 FictionFinder

FictionFinder er en FRBR-basert prototype der det er mulig å søke gjennom bibliografiske poster som fiksjonsbøker, ebøker og lydmaterialer. Systemet fokuserer på bibliografiske data som er over manifestasjons-nivået. Algoritmen samler bibliografiske poster i grupper basert på forfatter og tittelinformasjon. Forfatternavn og titlene brukes til å lage nøkler som igjen brukes til å deduplisere samlingen. Verk blir skapt ved å hente elementer fra forskjellige bibliografiske poster som så blir slått sammen[11].

Agatha Christie

Første søk på Agatha Christie som forfatter gir 2032 forskjellige treff der 1602 er identifisert som bøker. Hver post som er identifisert som et verk viser antall utgaver og også de forskjellige utgavene av verket. De første treffene har riktig identifiserte verk med flere hundre forskjellige utgaver av hvert verk, noe som er positivt. Etterhvert som man går gjennom flere av verkene oppstår det flere duplikater, verk med tittel på andre språk og utgaver med variasjoner i skrivemåte eller tittel. Resultatene som blir presentert senere har også langt færre utgaver. Det sier seg selv at hvis det dukker opp 1602 treff og forfatteren har skrevet 73 noveller, må en stor del av resultatet være støy.

Forfatter/Verk	Totalt antall treff	Antall treff identifisert som verk
Agatha Christie	2032	2032
Murder on the Orient Express	58	58
Murder in the Calais Coach[6]	8	8

Tabell 4.3: Treff på Agatha Christie og *Mord på Orientekspressen* i FictionFinder.

Hvis man søker etter et kjent verk av Agatha Christie som *Murder on the Orient Express*, resulterer det i 48 forskjellige treff på bøker. Noen av resultatene har forskjellige titler som for eksempel: *Murder on the Orient Express : a Hercule Poirot mystery*. De fleste resultatene er enten duplikater av det korrekte verket, eller andre oversettelser. En liten del av resultatet er også samlinger av fortellinger som for eksempel: *Five complete Hercule Poirot novels*, som ikke er et verk av Agatha Christie, men en samling av verk.

Jules Verne

Søk etter Jules Verne som forfatter resulterer i 1049 treff på bøker. I likhet med søket på Agatha Christie er de første treffene tilsynelatende korrekt identifisert. Men de aller fleste resultatene har engelsk tittel og representer oversettelser av boken. Disse oversettelsene inneholder stort sett andre engelske utgaver, med noen unntak. Resultatene med korrekt fransk tittel inneholder utgaver på forskjellige språk, fransk, engelsk, tysk, kinesisk og så videre. Videre består flere og flere av verkene, som blir listet opp senere, av duplikater, titler på andre språk og feilstavede titler. Igjen blir de postene som er identifisert riktig presentert først, og de feilidentifiserte postene blir presentert senere.

Forfatter/Verk	Totalt antall treff	Antall treff identifisert som verk
Jules Verne	1049	1049
Le Tour du monde en quatre-vingts jours	2	2
Around the World in Eighty Days[7]	33	33

Tabell 4.4: Treff på Jules Verne og *Jorden rundt på 80 dager* i FictionFinder.

Ved søk etter verket *Le Tour du monde en quatre-vingts jours*, eller *Jorden rundt på 80 dager* på norsk, dukker det opp to verk med litt forskjellig skrivemåte. Det første verket har 1181 forskjellige utgaver, og det andre har 23 utgaver. Ved søk etter den engelske tittelen *Around the World in Eighty Days*, kommer det 33 forskjellige resultater. Dette søket har flere av de samme karakteristikkene som søket på verket *Murder on the Orient Express*. Resultatene består i hovedsak av duplikater, andre titler på samme verk og oversettelser. Noen resultater har riktig tittel men også annen informasjon i tittelen som for eksempel: *Around the world in eighty days (Vietnamese)*.

[6] Amerikansk tittel
[7] Engelsk tittel

Miguel de Cervantes

Søket etter verk av Miguel de Cervantes har mange likheter med de foregående søkene. Totalt kommer det 917 forskjellige treff på bøker med Miguel de Cervantes som forfatter. De første treffene har flere tusen forskjellige utgaver, men også her har treffene forskjellige skrivemåte for titler. Treffene preges også her i stor grad av duplikater, oversettelser og andre lignende titler.

Forfatter/Verk	**Totalt antall treff**	**Antall treff identifisert som verk**
Miguel de Cervantes	959	959
El ingenioso hidalgo Don Quixote de la Mancha	21	21
The history of the valorous and wittie Knight-Errant Don-Quixote of the Mancha[8]	1	1

Tabell 4.5: Treff på Miguel de Cervantes og *Don Quijote* i FictionFinder.

Når man søker etter det mest kjente verket av Miguel de Cervantes: *El ingenioso hidalgo Don Quixote de la Mancha*, på norsk *Den skarpsindige lavadelsmannen Herr Quijote fra La Mancha*, med full originaltittel resulterer det i 21 forskjellige treff. Søker man med den forkortede tittelen: *Don Quixote* får man 201 treff. Men søker man på den fulle engelske tittelen: *The history of the valorous and wittie Knight-Errant Don-Quixote of the Mancha* dukker det kun opp ett verk.

Knut Hamsun

Søket etter Knut Hamsun returnerer totalt 227 treff. Igjen er de første treffene riktige verk med hundrevis av forskjellige utgaver. Problemet er at verkene ikke har originaltittel. De fleste treffene er identifisert med engelsk tittel, utgavene har flere forskjellige språk og da inkludert utgaver med originaltittel. Igjen er store deler av treffene feilidentifisert og er duplikater med annen skrivemåte på tittel eller tittel på andre språk.

Forfatter/Verk	**Totalt antall treff**	**Antall treff identifisert som verk**
Knut Hamsun	227	227
Sult	2	4
Hunger[9]	5	5

Tabell 4.6: Treff på Knut Hamsun og *Sult* i FictionFinder.

Ved søk etter verket Sult av Knut Hamsun, dukker det opp 4 forskjellige treff der Hamsun er oppført som forfatter. Alle treffene har tittelen på forskjellig språk. Posten med originaltittel inneholder bare 8 utgaver som alle er lydbøker. Posten med engelsk tittel har 700 utgaver på forskjellige språk, inkludert norsk. Søker man på den engelske tittelen til verket, *Hunger*, resulterer det i 5 treff,

[8] Engelsk tittel
[9] Engelsk tittel

der tre av dem er de samme som ved søk på originaltittel. Posten med riktig originaltittel dukker ikke opp på søk etter den engelske tittelen.

Günter Grass

På søk etter Günter Grass returneres 127 treff. Igjen har resultatene i hovedsak tittel på engelsk, og inneholder mange utgaver på forskjellige språk, også med originaltittel. Resultatene inneholder nok en gang duplikater og støy, men treffene som inneholder mange utgaver og som representerer korrekte verk, til tross for engelsk tittel, blir presentert først.

Forfatter Verk	**Totalt antall treff**	**Antall treff identifisert som verk**
Günter Grass	127	127
Die Blechtrommel	9	9
The Tin Drum[10]	3	3

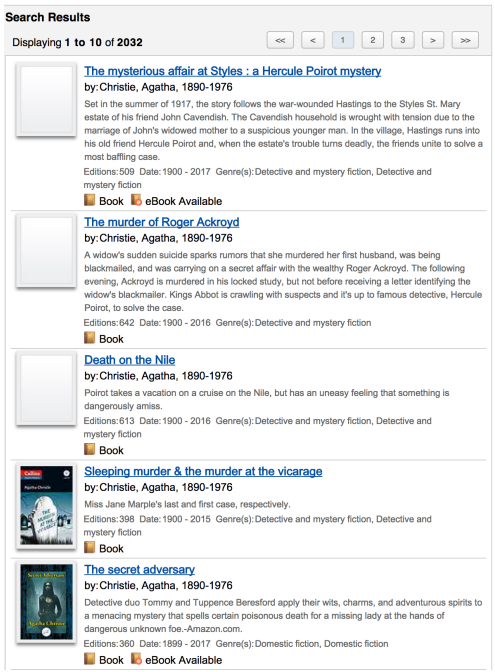
Tabell 4.7: Treff på Günter Grass og *Blikktrommen* i FictionFinder.

Søker man etter *Die Blechtrommel* av Günter Grass resulterer det i 35 treff hvis Günter Grass ikke spesifiseres som forfatter, og 9 hvis det blir spesifisert. Igjen lider resultatene av duplikater på andre språk eller skrivemåter. Treffet som er presentert først har 875 forskjellige utgaver på forskjellige språk, men verket er representert med engelsk tittel. Søker man etter verket med engelsk tittel, resulterer det i 3 treff der to av dem er mer eller mindre like, og én er et samleverk av de tre verkene i trilogien som *Die Blechtrommel* er en del av.

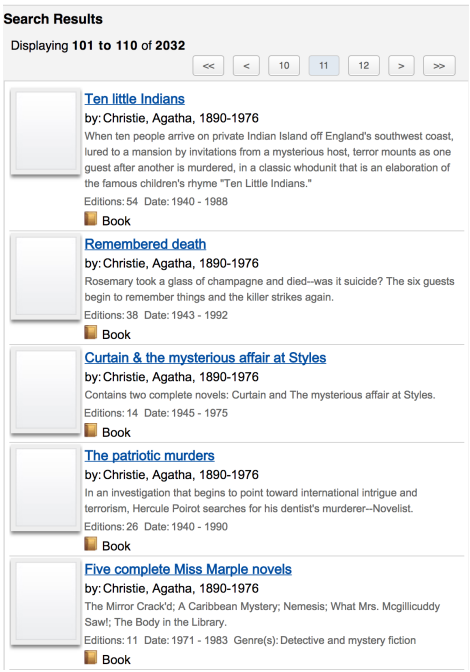
Oppsummering for FictionFinder

Det viser seg at FictionFinder returnerer mange riktige resultater, men den sliter også med store mengder støy. En fordel er at mer riktige resultater blir presentert tidlig, og mindre riktige resultater blir presentert senere. Dette er vist i Figur 4.1a som viser første side med resultater når det blir søkt på Agatha Christie som forfatter. Av de fem verkene som er presentert er 4 riktig identifisert og ett er et samleverk. Videre viser Figur 4.1b resultater fra side 11 der ett av resultatene er en alternativ tittel på et verk, to er verk som ikke har originaltittel, men amerikansk tittel, og de to siste er samleverk. I noen tilfeller er treffene i stor grad riktige med lite støy, men ut ifra denne analysen er det ikke noe tydelig mønster for når dette er tilfellet. For eksempel ga søket på *Jorden rundt på 80 dager* på originalspråket kun to treff, men søket etter *Don Quixote* med full tittel på originalspråket resulterte i mye støy. Verkene står stort sett oppført med engelsk tittel, og ikke originaltittelen. Og siden postene som står oppført med engelsk tittel ikke kommer opp ved søk på verkene med originaltittel, kan det virke som at det ikke er noen kobling mellom samme verket på forskjellige språk.

[10] Engelsk tittel



(a) De første resultatene



(b) Senere resultater

Figur 4.1: Skjermdump av resultater fra FictionFinder etter søk på Agatha Christie som forfatter.

4.2.2 Bibliothèque nationale de France

Den franske nasjonalbibliografien ble publisert som linked open data i 2011. Det nye systemet aggregerer data fra hovedkatalogen til den franske nasjonalbibliografien, det digitale biblioteket Gallica, for så å arkivere dataene til sider om forfattere, verk, tema etter FRBR-modellen [10]. FRBRiseringsprosessen til prosjektet er kort beskrevet i [10], og bruker maskinlærings-teknikker som fasiliterer sammenligningen mellom databaser. Prosessen går ut på først å skape koblinger mellom eksisterende autoritetsdata og de korresponderende bibliografiske postene med hjelp av en sammenkoblingsprosess. For usikre poster foreslår systemet alle potensielle linker mellom postene som blir sjekket manuelt av en operatør. Til slutt brukes klustringalgoritmer for å lage tittel-autoritetsdata der dataene ikke eksisterer.

Agatha Christie

Ved søk på forfatter i den franske nasjonal bibliografien går man inn på selve forfatteren og kan få en oversikt over alle postene som er relatert til forfatteren. For Agatha Christie dukker det opp 1365 poster. Fire av disse er identifisert som verk som inneholder forskjellige utgivelser. Resten er ulike manifestasjoner med stort sett franske titler.

Forfatter/Verk	Totalt antall treff	Antall treff identifisert som verk
Agatha Christie	1365	4
Murder on the Orient Express	1	1
Le crime de l'Orient Express[11]	1	1

Tabell 4.8: Treff på Agatha Christie og *Mord på Orientekspressen* i Bibliothèque nationale de France.

Ved søk etter verket *Murder on the Orient Express* med originaltittel resulterer det i ett verk: *Murder on the Orient-Express : film (1974)*, ingen andre verk, bøker eller andre manifestasjoner. Søker man med den franske tittelen på verket: *Le crime de l'Orient Express* får man samme resultat som med originaltittelen.

Jules Verne

Ved søk etter Jules Verne finnes det 18 forskjellige verk som er relatert til ham. Totalt er det 2072 forskjellige poster. Verkene inneholder alt fra 10 til 100 forskjellige manifestasjoner knyttet til det spesifikke verket. I likhet med det første søket består resten av postene av manifestasjoner.

Forfatter/Verk	Totalt antall treff	Antall treff identifisert som verk
Jules Verne	2072	18
Le Tour du monde en quatre-vingts jours	1	1
Around the World in Eighty Days[12]	0	0

Tabell 4.9: Treff på Jules Verne og *Jorden rundt på 80 dager* i Bibliothèque nationale de France.

Den franske nasjonalbiliografien returnerer ett verk ved søk etter verket *Jorden rundt på 80 dager*, på originalspråket. Dette verket har 93 forskjellige poster knyttet til seg. Ved søk på samme verk med den engelske tittelen, resulterer det i null treff.

Miguel de Cervantes

Søket etter Miguel de Cervantes gir 17 forskjellige verk og totalt 547 poster knyttet til ham. Igjen inneholder verkene alt fra 1 tilhørende manifestasjon til over 500 forskjellige manifestasjoner. Verkene er identifisert med tittel på fransk, og stort sett alle tilhørende manifestasjoner har franske titler.

[11] Fransk tittel
[12] Engelsk tittel

Forfatter/Verk	Totalt antall treff	Antall treff identifisert som verk
Miguel de Cervantes	547	17
El ingenioso hidalgo Don Quixote de la Mancha	1	1
L'Ingénieux Hidalgo Don Quichotte de la Manche[13]	2	2

Tabell 4.10: Treff på Miguel de Cervantes og *Don Quijote* i Bibliothèque nationale de France.

Søker man på *Don Quixote* med full originaltittel, returner det et verk av forfatteren Pierre Duflos. Søker man med den fulle franske tittelen: *L'Ingénieux Hidalgo Don Quichotte de la Manche*, kommer det to verk. Det ene verket representerer den andre delen av historien om Don Quixote med spansk tittel, og det andre verket er riktig identifisert, men med fransk tittel.

Knut Hamsun

Ved søk etter Knut Hamsun resulterer det i to verk som begge er riktig identifisert og med riktig originaltittel. Begge verkene inneholder flere utgaver på stort sett fransk, men også andre språk.

Forfatter/Verk	Totalt antall treff	Antall treff identifisert som verk
Knut Hamsun	2	2
Sult	2	2
La Faim[14]	2	2

Tabell 4.11: Treff på Knut Hamsun og *Sult* i Bibliothèque nationale de France.

Ved søk på verket *Sult* kommer det to resultater som er identifisert som verk, det ene er filmen Sult, og den andre er romanen. Det samme resultatet kommer etter søk på den franske tittelen *La Faim*, men da kommer det også noen urelaterte verk.

Günter Grass

Ved søk etter Günter Grass resulterer det i 13 identifiserte verk der alle har korrekt originaltittel. Totalt er det 135 tekstlige dokumenter knyttet til Grass, der noen grupperes under de identifiserte verkene. Ellers er det totalt 154 dokumenter knyttet til Grass, disse er ikke identifisert som verk, men representerer sanger, illustrasjoner og så videre.

Forfatter Verk	Totalt antall treff	Antall treff identifisert som verk
Günter Grass	135	13
Die Blechtrommel	2	2
Le Tembour[15]	26	26

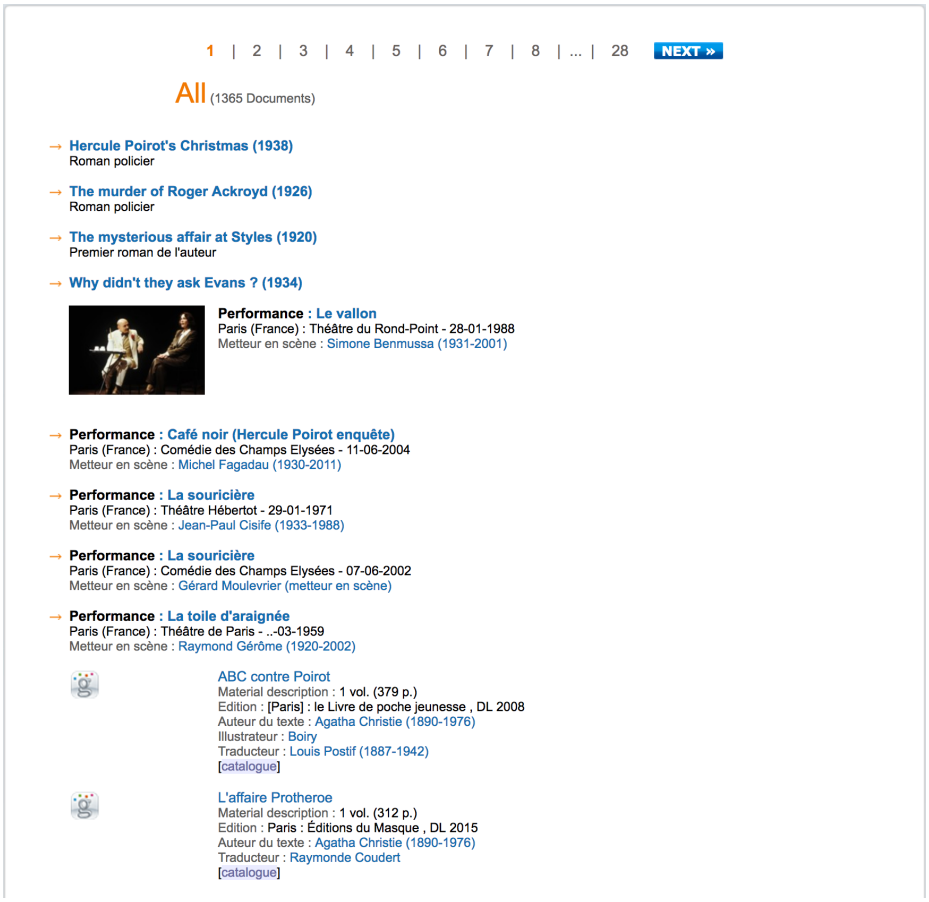
Tabell 4.12: Treff på Günter Grass og *Blikktrommen* i Bibliothèque nationale de France.

[13]Fransk tittel
[14]Fransk tittel

Ved søk etter verket *Die Blechtrommel* kommer det, i likhet med *Sult*, to identifiserte verk der det ene representerer det korrekte verket med originaltittel, og det andre representerer filmatiseringen. Verket som representerer romanen inneholder 15 forskjellige manifestasjoner der de enten har tysk eller fransk tittel.

Oppsummering for Bibliothèque nationale de France

Den franske nasjonalbibliografien har problemer med å representere internasjonale forfattere og titler. Biblioteket består i hovedsak av franske versjoner og liten grad av verkidentifisering. Av de 1365 postene som har en relasjon til Agatha Christie er det bare fire identifiserte verk, vist i Figur 4.2, som i tillegg representerer under 100 poster. En positiv side er at disse verkene har riktig tittel på originalspråket i motsetning til verkene av Miguel de Cervantes. Det samme gjelder verkene som er identifisert av Knut Hamsun og Günter Grass, der alle står oppført med originaltittel.



Figur 4.2: Skjermdump av de første postene som er relatert til Agatha Christie i Bibliothèque nationale de France.

[15] Fransk tittel

4.2.3 Biblioteca Virtual Miguel de Cervantes

Biblioteca Virtual Miguel de Cervantes, Cervantes, er en katalog som består av rundt 200 000 poster som har blitt transformert fra MARC21 til FRBR. Dataene har blitt automatisk omgjort til RDF-tripler, som beskriver entitetene, egenskapene og relasjonene til dataene. Prosessen var en stort sett automatisk prosess med unntak av noen relasjoner som ikke eksisterte i MARC-postene som måtte identifiseres manuelt[2].

Agatha Christie

Ved søk etter Agatha Christie resulterer det i tre verk der to av dem er verk som omhandler Agatha Christie. Det siste verket er *The Mysterious Affair at Styles*, som er riktig identifisert med originaltittel. Det er ingen andre verk av Agatha Christie som dukker opp ved søk etter tittel, så det virker som at innholdet i Cervantes er begrenset og at det begrenser resultatene.

Forfatter/Verk	**Totalt antall treff**	**Antall treff identifisert som verk**
Agatha Christie	3	3
Murder on the Orient Express	0	0
Asesinato en el Orient Express[16]	0	0

Tabell 4.13: Treff på Agatha Christie og *Mord på Orientekspressen* i Cervantes.

Jules Verne

Ved søk på Jules Verne dukker det opp 14 forskjellige verk, der Jules Verne er subjektet i verket. 8 av verkene er identifisert med riktig fransk originaltittel, de 5 resterende verkene er identifisert med den engelske tittelen, og det er ingen duplikater blant verkene. Alle verkene av Jules Verne har riktig relasjon til ham, så det dukker ikke opp andre verk ved søk etter andre titler enn de som allerede er identifisert. Ved søk etter originaltittelen til verkene som er identifisert med engelsk tittel kommer det ingenting opp, så det er ingen relasjon mellom titler av samme verk på forskjellig språk.

Forfatter/Verk	**Totalt antall treff**	**Antall treff identifisert som verk**
Jules Verne	14	14
Le Tour du monde en quatre-vingts jours	0	0
La vuelta al mundo en ochenta días[17]	0	0

Tabell 4.14: Treff på Jules Verne og *Jorden rundt på 80 dager* i Cervantes.

[16]Spansk tittel
[17]Spansk tittel

Miguel de Cervantes

Ved søk etter Miguel de Cervantes kommer det totalt 1247 identifiserte verk der Miguel de Cervantes er forfatter av 410 og subjekt i 837. Flere av verkene er duplikater og forskjellige versjoner av de samme verkene. Det er i tillegg mange andre adapsjoner. Det kan virke som at systemet lister opp forskjellige manifestasjoner, til tross for at verkene manifestasjonene tilhører blir listet opp øverst.

Forfatter/Verk	**Totalt antall treff**	**Antall treff identifisert som verk**
Miguel de Cervantes	410	410
El ingenioso hidalgo Don Quixote de la Mancha	25	25
The history of the valorous and wittie Knight-Errant Don-Quixote of the Mancha[18]	0	0

Tabell 4.15: Treff på Miguel de Cervantes og *Don Quijote* i Cervantes.

Ved søk etter *El ingenioso hidalgo Don Quixote de la Mancha* resulterer det i 25 verk som alle er duplikater av det samme verket. Det første verket har 5 forskjellige manifestasjoner, og resten av verkene har én hver. Søk på den engelske tittelen av verket resulterer i null treff. Det samme problemet er fremtredende ved søk på andre verk av Miguel de Cervantes.

Knut Hamsun

Knut Hamsun eksisterer ikke i Cervantes.

Günter Grass

Günter Grass eksisterer ikke i Cervantes.

Oppsumering for Cervantes

Cervantes-biblioteket inneholder et begrenset antall poster, som gjør at det er en del verk biblioteket ikke inneholder. Det viser seg at hovedutfordringen er duplikater, da det spesielt på verk av Miguel de Cervantes er mange duplikater med tilsynelatende identisk informasjon vist i Figur 4.3. Figuren viser hvordan forskjellige publikasjoner av samme verk er identifisert som ulike verk.

[18] Engelsk tittel

Figur 4.3: Skjermdump av de første verkene som er relatert til Miguel de Cervantes i Cervantes.

4.2.4 Deutsche Nationalbibliothek

Den tyske nasjonalbibliografien begynte å publisere autoritetsdataene sine i 2010 som linked open data og de bibliografiske dataene ble lagt til den allerede eksisterende linked data-tjenesten i 2012[24]. Biblioteket har ikke gjennomført noen FRBRisering på dataene sine, men dataene er gruppert basert på felt, som vil gi lignende resultater som en FRBRisering i dette tilfellet.

Agatha Christie

Ved søk etter forfatteren Agatha Christie i den tyske nasjonalbibliografien resulterer det i 2254 forskjellige treff. Disse treffene er ikke sammenslått basert på verk som de er i de andre bibliotekene, men hvert treff er en manifestasjon av verket. Treffene er i hovedsak tyske oversettelser. Noen av treffene indikerer originaltittel ved å ha originaltittelen i klammeparenteser før den tyske tittelen som for eksempel: *[After the funeral] Der Wachsblumenstrauß*. Det er mulig å filtrere treffene på verk, og i dette tilfellet resulterer det i 8 treff, der alle er riktig identifiserte verk av Agatha Christie. Verkene har tilsynelatende ingen relasjon til resten av postene i biblioteket, til tross for at det eksisterer manifestasjoner blant dataene som baserer seg på verkene som er identifisert.

Forfatter/Verk	**Totalt antall treff**	**Antall treff identifisert som verk**
Agatha Christie	2254	8
Murder on the Orient Express	44	2
Mord im Orient-Express[19]	2	2

Tabell 4.16: Treff på Agatha Christie og *Mord på Orientekspressen* i Deutsche Nationalbibliothek.

Ved søk på verket *Murder on the Orient Express* med originaltittel resulterer det i 44 forskjellige treff. Filtreres resultatene på verk viser resultatet to identifiserte verk der den ene er en film som ikke er relatert til Agatha Christie, og et verk som er riktig identifisert. Det samme resultatet kommer ved søk på den tyske tittelen til verket.

Jules Verne

Ved søk på verk av Jules Verne resulterer det i 18 resultater, der alle har riktig originaltittel. De fleste verkene har ingen publikasjoner relatert til seg, men noen få, som for eksempel *Le Tour du monde en quatre-vingts jours*, har 8 forskjellige publikasjoner knyttet til seg. Totalt har Jules Verne 1254 forskjellige poster knyttet til seg og kun et fåtall av disse er relatert til et verk.

Forfatter/Verk	**Totalt antall treff**	**Antall treff identifisert som verk**
Jules Verne	1254	18
Le Tour du monde en quatre-vingts jours	71	2
Reise um die Erde in 80 Tagen[20]	126	1

Tabell 4.17: Treff på Jules Verne og *Jorden rundt på 80 dager* i Deutsche Nationalbibliothek.

Ved søk på verket *Le Tour du monde en quatre-vingts jours* med originaltittel dukker det opp 71 forskjellige poster og 2 som er identifisert som verk. Det ene identifiserte verket er et skuespill og det andre er det riktig identifiserte verket. Skuespillet er et arbeid basert på boken der Jules Verne har bidratt.

Miguel de Cervantes

Ved søk på Miguel de Cervantes er resultatet stort sett det samme som for de to andre forfatterne. Det dukker opp 9 forskjellige verk, men hvert verk har få eller ingen poster knyttet til seg. Noen av verkene er identifisert med riktig originaltittel, men noen av verkene er feilidentifisert som verk. Videre er det knyttet 364 forskjellige poster til Miguel de Cervantes.

Forfatter/Verk	**Totalt antall treff**	**Antall treff identifisert som verk**
Miguel de Cervantes	364	9
El ingenioso hidalgo Don Quixote de la Mancha	68	1
Der sinnreiche Junker Don Quijote von der Mancha[21]	79	1

Tabell 4.18: Treff på Miguel de Cervantes og *Don Quijote* i Deutsche Nationalbibliothek.

[19] Tysk tittel
[20] Tysk tittel
[21] Tysk tittel

Ved søk på *El ingenioso hidalgo Don Quixote de la Mancha* resulterer det i 79 poster der 1 post er identifisert som verk. Det samme verket kommer ved søk på den tyske tittelen. Verket er riktig identifisert.

Knut Hamsun

Ved søk etter Knut Hamsun resulterer det i 421 treff. Videre gir søk etter verk av Knut Hamsun 15 forskjellige verk der alle er riktig identifisert og med riktig originaltittel. De fleste av verkene har ingen publikasjoner knyttet til seg, mens noen har få publikasjoner knyttet til seg.

Forfatter/Verk	Totalt antall treff	Antall treff identifisert som verk
Knut Hamsun	421	15
Sult	52	2
Hunger[22]	40	1

Tabell 4.19: Treff på Knut Hamsun og *Sult* i Deutsche Nationalbibliothek.

Ved søk etter verket *Sult* returneres to verk, der det ene er riktig identifisert, og det andre verket er urelatert. Ved søk på den tyske tittelen *Hunger*, dukker ikke verket opp. Hvis resultatene filtreres på verk kommer det mange publikasjoner både på søk med original og tysk tittel.

Günter Grass

Ved søk etter Günter Grass der det filtreres på verk, resulterer det i 34 riktig identifiserte verk med originaltittel. Igjen er det et fåtall av verkene som har noen publikasjoner knyttet til seg. Totalt er det 1625 forskjellige treff som er forskjellige publikasjoner av Grass. Stort sett alle verkene av Grass er riktig identifisert som verk. Dette kan ha noe å gjøre med at Grass kommer fra Tyskland.

Forfatter Verk	Totalt antall treff	Antall treff identifisert som verk
Günter Grass	1625	34
Die Blechtrommel	374	2
The Tin Drum[23]	32	0

Tabell 4.20: Treff på Günter Grass og *Blikktrommen* i Deutsche Nationalbibliothek.

Ved søk etter *Die Blechtrommel* resulterer det i 374 publikasjoner som er alt fra bøker, til lydbøker og oversettelser. Filtreres resultatene på verk resulterer det i to treff, der det ene er romanen og det andre er filmatiseringen.

Oppsummering for Deutsche Nationalbibliothek

Den tyske nasjonalbibliografien sliter stort med rotete, inkonsekvente og ufullstendige data. Veldig få poster er knyttet til verk, så flesteparten av postene står som egne manifestasjoner. De riktige dataene er vanskelige å finne og det finnes mange forskjellige duplikater både av personer og andre

[22] Tysk tittel
[23] Engelsk tittel

poster. Dette gjør at det er vanskelig å finne noen fornuftige relasjoner mellom dataene. Postene som er riktig identifisert som verk har stort sett riktig originaltittel som vist i Figur 4.4, noe som er bra da stort sett alle postene kun har informasjon på tysk. Verkene som er identifisert er riktige, men det er få identifiserte verk, noe som kan tyde på at det kun er de verkene som det er helt sikkert at er et verk som blir identifisert som et verk. Verk av internasjonale forfattere som er identifisert har få publikasjoner relatert til seg som vist i Figur 4.5a, men det er flere publikasjoner relatert til verk av nasjonale forfattere vist i Figur 4.5b. Det finnes mange publikasjoner av flere av verkene som ble søkt etter, men disse er ikke relatert til noen av verkene, med noen unntak.

Figur 4.4: Skjermdump av verkene som er relatert til Agatha Christie i Deutsche Nationalbibliothek.

Link zu diesem Datensatz	http://d-nb.info/gnd/4570043-6
Verfasser/Urheber	Christie, Agatha
Titel des Werkes	Murder on the Orient Express
Andere Titel	Mord im Orientexpress (ÖB-Alternative) Murder on the Orientexpress Der rote Kimono Mord im Orient Express Mord im Orient-Express
Quelle	Oxf. companion, DNB
Erläuterungen	Definition: Kriminalroman (1934)
Zeit	erschienen: 1934
Land	Großbritannien (XA-GB)
Sprache(n)	Englisch (eng)
Systematik	12.2p Personen zu Literaturgeschichte (Schriftsteller)
Typ	Werk (wit)
Thema in	1 Publikation 1. *Spannung in verschiedenen Grundtypen der Detektivliteratur* *Engel, Patrick. - Trier : Wiss. Verl. Trier, 2008*
Maschinell verknüpft mit	1 Publikation 1. *Ricardo Piglias "Blanco nocturno". Ein Hybrid zwischen novela de engima und novela negra? [Elektronische Ressource]* *Heckl, Christoph. - München : GRIN Verlag GmbH, 2014*

(a) Skjermdump av posten til verket *Murder on the Orient Express*.

Link zu diesem Datensatz	http://d-nb.info/gnd/4099211-1
Verfasser/Urheber	Grass, Günter
Titel des Werkes	Die Blechtrommel
Andere Titel	Skardinis būgnelis
Quelle	Kindler
Zeit	erschienen: 1959
Land	Deutschland (XA-DE)
Sprache(n)	Deutsch (ger)
Oberbegriffe	Teil von: Grass, Günter: Danziger Trilogie
DDC-Notation	833.914
Systematik	12.2p Personen zu Literaturgeschichte (Schriftsteller)
Typ	Werk (wit)
Thema in	64 Publikationen 1. *Einheit in der Differenz – die innere Struktur des Erzähler-Ichs* *Diller, Franziska. - Hamburg : Kovač, 2015* 2. *Komisches Selbstverkennen* *Socha, Monika. - Hamburg : Kovač, 2014* 3. ...
Maschinell verknüpft mit	3 Publikationen 1. *Realität vs. Fiktion. Günter Grass' "Blechtrommel" als autobiographischer und historischer Roman [Elektronische Ressource]* *München : Science Factory, 2015* 2. *Die Blechtrommel von Günter Grass [Elektronische Ressource]* *Thur, Claudia. - München : GRIN Verlag, 2010* 3. ...
Zugehörige Publikationen	6 Publikationen 1. *[Grass]* *Die Blechtrommel* *Grass, Günter. - Göttingen : Steidl, [2016]* 2. *[Grass]* *El tambor de hojalata* *Grass, Günter. - Barcelona : Debolsillo, mayo, 2016, Primera edición en Debolsillo, (segunda reimpressión)* 3. ...

(b) Skjermdump av posten til verket *Die Blechtrommel*.

Figur 4.5: Skjermdump av verk-poster i Deutsche Nationalbibliothek.

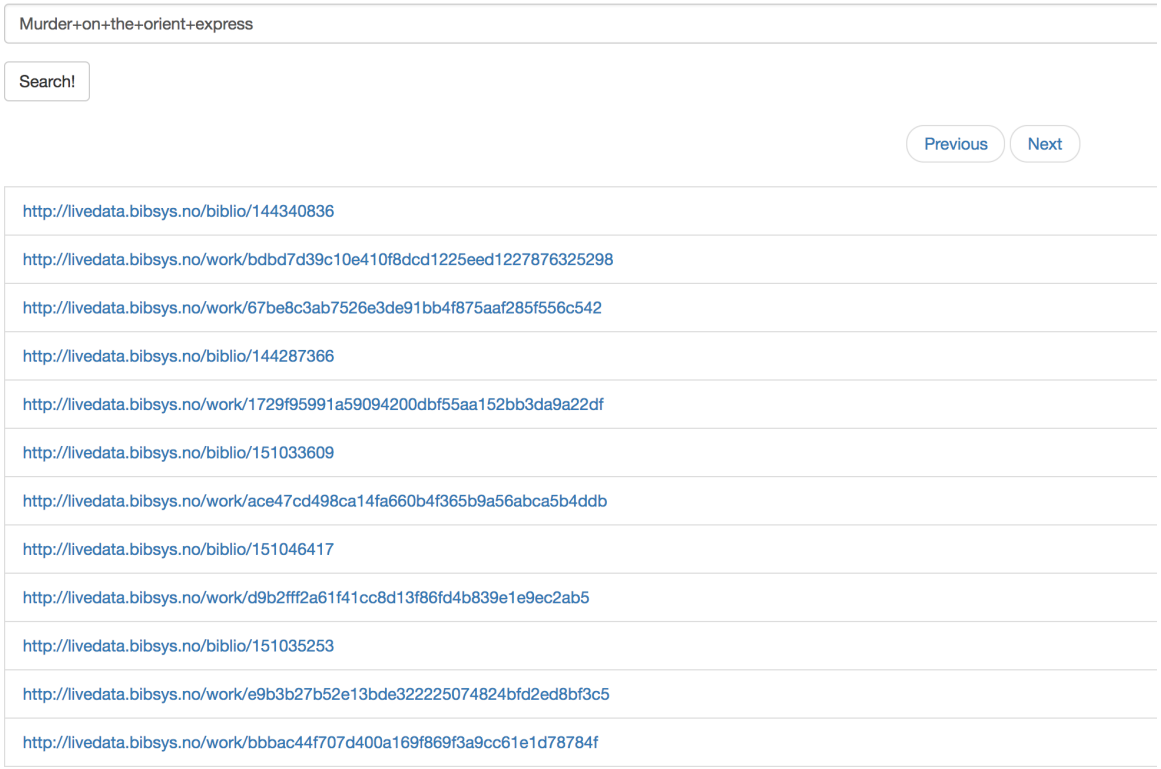
4.2.5 BIBSYS semantisk web

BIBSYS semantisk web er et prosjekt som jobber med automatisk ekstrahering av verk. Prosjektet bruker poster fra BIBSYS Autoritetsregister og Bibliotekbasen for å identifisere verk. Verkene blir identifisert basert på forskjellige regelsett. Regelsettene baserer seg på MARC-felt og underfelt som kan inneholde titler til verk. Ved første kjøring av systemet ble det generert 3 571 053 verkkandidater basert på 4 352 391 bibliografiske poster. Ved andre kjøring ble det generert 4 646 613 verkkandidater basert på 5 530 899 bibliografiske poster. Dette tilsvarer henholdsvis 1,22 og 1,19 bibliografiske poster per verk. Dette er et veldig lavt antall bibliografiske poster per verkkandidat.

For å gjøre en analyse av resultatene til BIBSYS semantisk web har web-grensesnittet som ligger ute blitt sett på. I skrivende stund er SPARQL endepunktet for søk i resultatene nede, og det har derfor blitt brukt den vanlige søkemotoren. Utfordringen er at søket tilsynelatende returnerer alle postene som inneholder ett av ordene fra søke-teksten. For eksempel hvis det søkes etter *Murder on the Orient Express*, returneres alle poster som inneholder et av disse ordene. Søkeresultatet viser kun lenken til hver post med en post-nøkkel, som vist i Figur 4.6, det er derfor vanskelig å få et raskt overblikk over resultatene. Det vises heller ikke et tall på hvor mange poster som er returnert.

Av den grunn er det ikke mulig å gjøre et enkelt oppslag på verk, og antall resultater som blir

returnert vises ikke. Derimot er det mulig å finne fram til nøkkelen til forskjellige forfattere, og videre finne alle postene som er relatert til forfatterne. Derfor ble det valgt å analysere antallet poster som er relatert til de forskjellige forfatterne.



Figur 4.6: Skjermdump av søk etter *Murder on the Orient Express* i BIBSYS semantisk web.

Forfattere

Tabell 4.21 viser antall poster som er relatert til de forskjellige forfatterne. Tallet i parentes viser antallet poster der forfatteren står som *creator*. Tabellen viser også hvor mange av postene som er relatert til forfatteren som er identifisert som verk.

Forfatter	**Totalt antall poster**	**Antall poster identifisert som verk**
Agatha Christie	971(907)	283
Jules Verne	621(550)	202
Miguel de Cervantes	1417(795)	241
Knut Hamsun	7591(2213)	510
Günter Grass	445(267)	86

Tabell 4.21: Poster relatert til forfatterne som er brukt i denne analysen i BIBSYS semantisk web.

Tabell 4.21 viser et stort antall poster relatert til de forskjellige forfatterne. Det er også et stort antall poster som er identifisert som verk. Til tross for at antallet poster som er identifisert som verk er betydelig lavere enn det totale antallet poster, er dette tallet mye høyere en det antallet verk de forskjellige forfatterne har skrevet, som er vist i Tabell 4.2. Det kan tyde på at det er både duplikater og poster som ikke representerer verk som er identifisert som verk.

4.3 Sammendrag

Denne analysen viser noen tendenser blant resultatene etter FRBRiseringsprosesser og kvaliteten til data som er publisert til linked open data-skyen. Støy i resultatene er et gjennomgående problem. Denne støyen er i form av poster som ikke har noen relasjon til verk, duplikater, eller oversettelser som er identifisert som egne verk. Noen av bibliotekene har veldig få identifiserte verk blant mange poster, noe som kan tyde på utfordringer ved identifiseringen av verk basert på postene. Eller det kan tyde på at verk kun blir markert som verk hvis de er helt sikre på at det er et verk. Det er forskjell på hvordan systemene håndterer verk fra andre nasjonaliteter, men stort sett er verkene identifisert med originaltittelen. En annen fellesnevner er at mange poster ikke er relatert til det verket de egentlig er relatert til. Dette er et problem da en av hovedmotivasjonene for å identifisere verk er å kunne gruppere alle uttrykk, manifestasjoner og gjenstander under det korrekte verket de tilhører.

Denne analysen viser at det er et sterkt behov for å forbedre kvaliteten til resultatene av FRBRiseringsprosesser og data som er publisert i linked open data-skyen. I hovedsak gjelder dette bedre verk-identifikasjon, slik at flere verk blir identifisert, som igjen vil redusere antall duplikater og annen støy. Utfordringen vil også være å finne relasjonene mellom poster til det respektive verket de er relatert til.

Kapittel 5

Tilnærming

Dette kapittelet beskriver fremgangsmåten i prosjektet. Først presenteres målet med prosjektet, deretter blir dataene som er brukt og katalogene dataene kommer fra beskrevet. Til slutt beskrives metoden som er brukt.

5.1 Målet

Målet med denne oppgaven er å finne en metodikk som angir påliteligheten til verkene som er generert, som igjen kan bedre FRBRiseringsprosessen. Resultatet av publiseringen av dataene til linked open data-skyen kan bli bedre ved å for eksempel kun publisere de verkene som har høy pålitelighet. Grunnsteinen i en god FRBRisering er identifisering av verk, da verk er det høysete nivået i alle postene. Hvis verket ikke er identifisert har ikke uttrykket som egentlig er relatert til verket et holdepunkt, eller noe som inneholder felleskilden til de andre uttrykkene. På mange måter mister uttrykk og manifestasjonen meningen sin hvis de ikke er relatert til det korrekte verket de realiserer. Tidligere i oppgaven har det blitt sett på verkidentifikasjonen i eksisterende systemer, og resultatet er varierende. En fellesnevner er at resultatene inneholder store mengder støy som består av feilidentifiserte verk. Videre fører dette til at uttrykk og manifestasjoner også blir feilidentifisert fordi de blir relatert til poster som ikke er verk i det hele tatt.

I et vanlig biblioteksystem vil dårlig strukturerte og rotete data ikke nødvendigvis ha andre konsekvenser for brukeren enn en dårlig opplevelse. Mennekser er flinke til å sortere ut det som er nyttig, og filtrere ut det som ikke er interessant. Derfor vil brukerne kunne finne fram til det de ser etter i en samling med duplikater og rotete data, men det vil ta lenger tid. I linked open data er det viktig at informasjonen som blir publisert er av god kvalitet. Linked open data brukes ikke bare av mennesker, men også maskiner, og det stilles derfor høyere krav til riktighet og god kvalitet. Siden dataene som er publisert som linked open data er lenket, er det også viktig at data som representerer det samme er linket sammen.

Når verkene i en samling er identifisert vil prosessen med å relatere uttrykk, manifestasjoner, gjenstander, personer og korporasjoner bli enklere og mer ryddig. Verksidentifisering er derfor et naturlig startpunkt i å forbedre FRBRiseringsprosessen. Målet med denne oppgaven er å finne en metodikk som kan forbedre FRBRiseringsprosessen gjennom å identifisere verk på en enklere måte.

5.2 Data

Dataene som er brukt i denne oppgaven er på MARC21-format, som bruker forskjellige felt for verdiene til postene. MARC-feltene er identifisert med tresifrede nummer, og underfeltene er identifisert med et dollartegn og en bokstav eller et tall. Bibliotekene som er brukt i oppgaven skiller seg fra hverandre på forskjellige måter, både ved bruken av felt, og informasjonen som er lagret i postene. Først skal feltene som er relevante for identifiseringen av verk bli sett på, og deretter hvordan de fire forskjellige bibliotekene har brukt disse feltene. Feltene det er snakk om er:

130

130-feltet, uniform tittel, inneholder den uniforme tittelen som er brukt som hovedoppføring for posten. Hvis et verk er kjent under flere forskjellige titler, brukes 130-feltet for den tittelen som er valgt til å representere verket.

240

240-feltet, uniform tittel, inneholder en uniform tittel som er brukt hvis 130-feltet ikke er brukt. For at 240-feltet skal bli brukt må hovedoppføringen minst inneholde et person-, korporasjon-, eller møtenavn.

245

245-feltet, kjent tittel, inneholder tittel og ansvarsområde for den bibliografiske beskrivelsen av verket. I denne sammenhengen er dette feltet brukt fordi det også kan inneholde annen tittelinformasjon.

246

246-feltet, variasjon av tittel, kan enten inneholde forskjellige former av tittelen som forekommer i forskjellige deler av gjenstanden, deler av tittelen, eller alternativ form på tittelen når forskjellen fra den kjente tittelen, 245-feltet, er så stor at de begge kan være med på å identifisere gjenstanden.

600

600-feltet, emneinnførsel - Personnavn, inneholder navnet på en person. I dette tilfellet er det underfeltet $t som er interessant da det kan inneholde tittelen til et verk.

610

610-feltet, emneinnførsel - Korporasjonsnavn, inneholder navnet på en korporasjon. I likhet med 600-feltet er det underfeltet $t som er interessant da det kan inneholde tittelen til et verk.

611

611-feltet, emneinnførsel - Møtenavn, inneholder navnet på et møte. I likhet med 600- og 610-feltet er det underfeltet $t som er interessant da det kan inneholde tittelen til et verk.

630

630-feltet, emneinnførsel - Uniform tittel, inneholder en emneinnførsel der oppføringselementet er en uniform tittel.

700

700-feltet, tilleggsinnførsel - Personnavn, tilleggsinnførsel der emneelementet er et personnavn. Feltet kan inneholde personnavn der det ikke vil være mer hensiktsmessig å bruke 600-feltet. I likhet med 600-, 610- og 630-feltene inneholder 700-feltet underfeltet $t som kan inneholde "tittel på et verk". 7xx-feltene brukes til tilleggsinformasjon, og brukes forskjellig fra bibliotek til bibliotek.

710

710-feltet, tilleggsinnførsel - Korporasjonsnavn, tilleggsinnførsel der emneelementet er et korporasjonsnavn. Feltet kan inneholde korporasjonsnavn der det ikke vil være mer hensiktsmessig å bruke 610-feltet. I likhet med 700-feltet inneholder 710-feltet underfeltet $t som kan inneholde "tittel på et verk".

711

711-feltet, tilleggsinnførsel - Møtenavn, tilleggsinnførsel der emneelementet er et møtenavn. Feltet kan inneholde møtenavn der det ikke vil være mer hensiktsmessig å bruke 611-feltet. I likhet med 700- og 710-feltet inneholder 711-feltet underfeltet $t som kan inneholde "tittel på et verk".

730

730-feltet, tilleggsinnførsel - Uniform tittel, tilleggsinnførsel der emneelementet er en uniform tittel. Feltet kan inneholde en uniform tittel der det ikke vil være mer hensiktsmessig å bruke 630-feltet.

5.2.1 Katalogene

I denne oppgaven er det benyttet rådata fra fire forskjellige kataloger. Dataene består av hele bibliografien til Harvard Library, University of Michigan, den tyske nasjonalbibliografien og den norske nasjonalbibliografien. Harvard og University of Michigan er begge universitetsbiblioteker, og de to andre katalogene er fra to nasjonalbibliografier. Det vil derfor være forskjeller mellom postene som opptrer i katalogene. Nasjonalbibliografiene vil i hovedsak inneholde flere skjønnlitterære poster, der universitetsbibliotekene vil inneholde faglitteratur. Universitetsbibliotekene vil også kunne inneholde skjønnlitterære poster. Tabell 5.1 viser forekomsten av MARC-feltene i originaldataene i de forskjellige katalogene. Karakteristikkene til katalogene er beskrevet under.

Felt	NNB	Umich	DNB	Harvard
130	0,268 %	1,428 %	0,047 %	1,167 %
240	2,377 %	3,547 %	2,601 %	4,137 %
245	100,000 %	100,000 %	100,000 %	100,000 %
246	16,955 %	-	-	-
600	0,239 %	1,408 %	0,269 %	1,493 %
610	0,019 %	0,192 %	0,202 %	0,211 %
611	-	0,002 %	0,002 %	0,006 %
630	0,361 %	1,217 %	0,343 %	2,060 %
700	0,263 %	2,549 %	4,140 %	3,811 %
710	0,006 %	0,390 %	0,004 %	0,611 %
711	-	0,015 %	0,000 %	0,012 %
730	0,130 %	2,946 %	0,734 %	1,992 %

Tabell 5.1: Forekomsten av MARC-feltene i originaldataene i de forskjellige katalogene.

5.2.2 Den Norske Nasjonalbibliografien

Den norske nasjonalbibliografien, NNB, er den minste katalogen som er brukt i denne oppgaven. Totalt er det rundt 700 000 originalposter som kan være kandidater for verk. NNB skiller seg mest ut med bruken av 246-feltet. NNB er den eneste katalogen som har brukt dette feltet, og nesten 17 prosent av postene inneholder dette feltet. Videre er 700-, 730- og 710-feltet lite brukt sammenlignet med de andre katalogene. 130-feltet er også relativt lite brukt der under 0,3 prosent av postene bruker 130-feltet.

I NNB er 130-feltet som sagt lite brukt, og i hovedsak dukker det opp i poster fra Bibelen, der 130-feltet brukes til å nettopp identifisere at posten kommer fra Bibelen. Videre brukes 240-feltet i stor grad til musikkstykker, arrangementer, sanger og så videre. 245-feltet er i stor grad brukt likt mellom alle katalogene og beskriver tittelen på manifestasjonen. Dette kan både være originaltittelen og for eksempel tittelen på det oversatte verket. NNB skiller seg mest fra de andre katalogene i bruken av 246-feltet da de andre katalogene ikke bruker feltet i det hele tatt. I NNB beskriver 246-feltet originaltittel til verket, der 246 $a er selve originaltittelen og 246 $i forklarer at det er originaltittelen. Videre er feltene 600, 700 og 710 brukt stort sett likt mellom de forskjellige katalogene, men lite brukt i NNB. Disse feltene brukes først og fremst til å beskrive personer som er behandlet som emne i posten, men også i noen tilfeller der verket til en person er behandlet som emne i posten. Feltene 610, 611 og 711 er stort sett ikke-eksisterende hos alle katalogene.

5.2.3 University of Michigan Library

University of Michigan Library, UMICH, er den nest minste katalogen, men har fortsatt dobbelt så mange poster som NNB, med rundt 1,3 millioner poster. UMICH har en mer hyppig bruk av 130-feltet, og poster med 240-felt finnes det også en del av til forskjell fra for eksempel den tyske nasjonalbibliografien. Stort sett skiller UMICH seg lite fra de andre katalogene i hyppigheten av bruken av de forskjellige feltene.

Når det kommer til innholdet i de forskjellige feltene, skiller UMICH seg ut på noen områder. 130-feltet for eksempel, inneholder i stor grad samme tittel som 245-feltet med unntak av at det i tillegg er skrevet et stedsnavn eller en institusjon i parentes etter tittelen. 630-feltet er hyppigere

brukt og brukes i hovedsak for poster som er en del av et større verk, ofte Bibelen eller Koranen, men også mediehus. Igjen er 240-feltet brukt om musikalske verk, og siden 240-feltet er mer brukt i UMICH kan det tyde på at det inneholder flere musikalske verk. Videre er det lite som skiller bruken av de andre feltene fra de andre katalogene.

5.2.4 Deutsche National Bibliothek

Deutsche National Bibliothek, DNB, er den største katalogen med rundt 15 millioner poster. DNB er katalogen med færrest poster med 130-, 240-, 600-, 630- og 710-felt. Poster med 130-, 630- og 710-felt er mer eller mindre ikke-eksisterende.

DNB skiller seg lite fra de andre katalogene i felt-bruken. Først og fremst er 240-feltet brukt med underfeltet $0 som inneholder en eller annen form for kontrollnummer. Dette gjelder også 600-feltet, men 600-feltet er også brukt likt som i de andre katalogene. 730-feltet er interessant da det ofte inneholder en annen tittel enn tittelen fra 245-feltet. For poster som har tysk tittel brukes 730-feltet i noen tilfeller til originaltittel. Poster som inneholder tyske verk der originaltittelen er lagret i 245-feltet inneholder 730-feltet tittelen på et annet språk, stort sett engelsk.

5.2.5 Harvard Library

Harvard Library, Harvard, er den nest største katalogen som er vurdert med over 13 millioner poster. Harvard skiller seg også lite ut i hyppigheten av feltbruken, og er mer eller mindre identisk med UMICH. Det er imidlertid noen forskjeller i selve feltbruken. 130-feltet inneholder en god del originaltitler, men en del poster har den samme tittelen i 130-feltet som i 245-feltet, men med ekstra informasjon i parentes. Denne informasjonen kan for eksempel være et sted, en dato eller kategorien til verket. Videre er 240-feltet hyppig brukt til å beskrive kategorien til posten, og oftest for samleverk. Det kan gjelde både litteratur og musikalske verk. 240-feltet inneholder i stor grad kategorier som novelle, short stories, works, quartets og så videre.

5.3 Dataprosessering

Før dataene ble analysert ble de FRBRisert gjennom systemet som er beskrevet i [25]. Dette systemet gjør en god jobb som FRBRiseringsprosess. Resultatene som kommer ut blir ikke deduplisert på noen måte, og det vil derfor være duplikater blant resultatene. Dataene som er blitt brukt i denne oppgaven er postene som har blitt identifisert som verk av dette systemet. Dette delkapittelet beskriver prosesseringen som er gjort og hvordan verkene er blitt identifisert.

5.3.1 Identifisering av verk

Systemet bruker MARC-feltene beskrevet over til å identifisere mulige verk. Postene blir analysert felt for felt, og ikke post for post. Dette gjør at én post kan bli identifisert som flere forskjellige mulige verkskandidater. I hovedsak blir alle feltene som er beskrevet over brukt til å hente ut mulige verk. Hvor detaljert postene er beskrevet bestemmer hvor detaljerte de foreslåtte verkene blir.

Verdiene som må hentes ut er dateOfWork(dato for verket), formOfWork(formen til verket, mediumOfPerformance(formatet til opptredenen), numberingOfPart(delnummer), otherDistinguishingCharacteristicOfTheWork(andre identifiserende karakteristikker til verket), og key(nøkkel). Disse verdiene kan også brukes videre for enklere å skille verk fra hverandre, og å lettere kunne identifisere

verkene. Systemet vil i første omgang ukritisk gå gjennom katalogene og hente ut alle poster som inneholder de nevnte feltene for så å generere nye poster basert på informasjonen som er gitt. Måten MARC-formatet er laget på gjør at poster som inneholder 130-feltet ikke vil inneholde 240-feltet, og omvendt. Katalogen posten kommer fra, blir også lagret i den nye posten.

Når en post med et ønsket felt er funnet, blir det laget en nøkkel for å identifisere den nye posten. Nøkkelen blir laget basert på forskjellige attributter til posten. Nøkkelen blir så strippet for spesialtegn og mellomrom, og blir lagt sammen til en streng. Strengen vil da være unik for alle verk med samme tittel som er relatert til samme forfatterpost. Deretter blir nøkkelen hashet til en UUID, som lager en unik hash-id for postene som har disse feltene. Id-en blir senere brukt i dedupliseringsprosessen.

5.3.2 Deduplisering

Etter at alle verkskandidatene er generert vil datasettet bestå av mange duplikater av de samme verkene. Dette er støy som er ønskelig å unngå i resultatdatasettet. FRBRiseringen gjør derfor en enkel sammenslåing av postene som tilsynelatende er identiske. For å fastslå postene som er identiske blir den genererte hash-nøkkelen, som er beskrevet over, brukt. Postene som har identisk hash-nøkkel blir slått sammen til én post. I tillegg til dette lagres det informasjon i den nye posten for å beskrive postene som har vært med i sammenslåingen. For det første blir informasjon om antall poster som er slått sammen lagret, samt informasjon om hvilken katalog de kommer fra og feltet de er hentet ut fra. Informasjonen er ikke nødvendigvis viktig for verket i seg selv, men den er viktig i prosessen videre for å identifisere de korrekte verkene. I sammenslåingen kombineres informasjonen i de forskjellige postene slik at ingen informasjon går tapt. Dette kan være verdier som dato, delnummer og så videre, men også relasjoner. Måten nøkkelen er generert på gjør at det vil kunne oppstå flere titler for det samme verket i én og samme post. Titlene vil da i hovedsak avvike i bruken av spesialtegn og mellomrom. Hvis én post har generert flere verkskandidater på grunn av bruk av flere felt som inneholder den samme informasjonen, vil de her bli slått sammen igjen til samme post. Denne prosessen er kun en dedupliseringsprosess som baserer seg på eksakt match, og vil ikke gjøre noe søk på andre mulige duplikater som har små avvik i skrivemåte.

5.4 Testoppsett

I dette prosjektet har verksreferanser og regelbasert prosessering blitt testet. MARC-poster har blitt prosessert til verkskandidater. I disste verkskandidatene har forskjellige karakteristikker blitt vurdert opp mot om de representerer et verk eller ikke.

5.5 Kvalitet

For å kunne se hvilke av verkenes karakteristikker som er med på å påvirke sannsynligheten for at en post kan representere et verk, må postene som faktisk representerer et verk identifiseres. Dette ble gjort manuelt ved å søke på posttittelen, forfatter, dato og så videre på blant annet Google og Wikipedia. Denne prosessen er tidkrevende, men nødvendig for å kunne lage en automatisk prosess som gjør samme jobben. Basert på funnene ble verkskandidatene markert med en av tre metrikker. Den første metrikken er FP, false positive, poster som ikke representerer et verk. Den andre metrikken er TPS, True positive singular, poster som kan identifiseres som verk. Den tredje metrikken er TPD,

True positive duplicate, poster som kan identifiseres som verk, men som er et duplikat av et verk som allerede er identifisert i dette testsettet. Dette betyr at flere verker som blir markert som TPS, potensielt vil være duplikater av andre verker i katalogen.

5.5.1 Kvaliteten til resultatene

I dette prosjektet er det flere faktorer som spiller inn på resultatene. Målet med prosjektet er ikke å teste kvaliteten til FRBRiseringsprosessen, men å finne en kilde til identifikasjon av verk. Resultatene som har blitt produsert avhenger i stor grad av postene i seg selv, men det er også tre viktige ting som påvirker resultatene.

Kvalitet til originaldata

Dataene som har blitt brukt i dette prosjektet har sine styrker og svakheter. Resultatene som blir produsert vil avhenge av flere faktorer som bestemmes av katalogene som er brukt. For det første kommer katalogene fra forskjellige typer biblioteker og fra forskjellige land. Katalogene kan dermed være forskjellige i feltbruk, både i form av hvilke felt som er brukt og hvilken informasjon som er lagret i de forskjellige feltene.

Videre har katalogene også i varierende grad forskjellig innhold fordi to av katalogene kommer fra nasjonalbibliografier og de to andre fra universitetsbibliografier. Dette kan påvirke resultatene, fordi det for eksempel vil kunne være poster som oppstår hyppig i de to nasjonalbibliografiene, og ikke i de to universitetsbibliografiene. Siden katalogene har forskjellige nasjonaliteter vil det både påvirke innholdet, men også hvordan innholdet er lagret. For eksempel vil den norske nasjonalbibliografien inneholde flere norske verk, men den vil også inneholde flere oversettelser enn de andre katalogene. Hvis en katalog inneholder mange oversettelser som ikke er identifisert med originaltittel vil dette påvirke antallet riktig identifiserte verk i prosjektet.

Prosesseringen

Prosesseringen er transformeringen fra MARC-poster til FRBR. Denne prosesseringen er regelbasert og bruker MARC-feltene til å tolke dataene. Det er disse reglene som har sortert ut postene som kan identifisere et verk. Reglene er ikke perfekte, og vil derfor ha en innvirkning på resultatet. Mulige feil kan oppstå, og det kan hende poster som kunne identifisert verk ikke har kommet med i samlingen. Resultatene fra dette prosjektet vil kunne brukes til å lage mer nøyaktige regler som gjøre at resultatet av en FRBRiseringsprosess vil bli bedre. Prosesseringen vil derfor kunne påvirke resultatene i dette prosjektet.

Sammenslåingen

Sammenslåingen av identiske poster er en viktig del av prosesseringen av dataene. Prosessen slår sammen poster som har identisk nøkkel, og denne nøkkelen er basert på innholdet i posten. Det ble ikke gjennomført en mer avansert strengoperasjon enn å fjerne spesialtegn og mellomrom. Dermed ble kun poster som har nesten identisk tittel og annen informasjon slått sammen. Dette gjør at det eksiterer duplikater i dataene som er brukt. Hvis sammenslåingen hadde brukt mer avanserte streng-matchings-teknikker ville det kuttet ned på antall duplikater. Avanserte streng-matchings-teknikker er tidkrevende for store systemer, men det har potensialet til å forbedre resultatene betraktelig.

Duplikater er ikke nødvendigvis et problem i prosjektet i seg selv. Dersom flere nesten identiske poster hadde blitt slått sammen, ville resultatet kunnet blitt mer nøyaktig. Eksempelvis kan man se dette gjennom å analysere karakteristikken "antall ulike kataloger en verkskandidat har opprinnelse i". Da vil to poster med en liten forskjell i skrivemåte kunne produsere to verkskandidater som begge representerer et verk. Den ene posten har opprinnelse i to av katalogene, og den andre har opprinnelse i de to andre. I resultatet vil denne posten bli vurdert som to verkskandidater som begge har opprinnelse i to ulike kataloger. Hvis de hadde blitt slått sammen ville de kunne blitt vurdert som en verkskandidat som har opprinnelse i fire ulike kataloger.

5.5.2 Validering

Dataene i denne testen er hentet ut på bakgrunn av feltene de har brukt, feltene ble beskrevet i 5.2. Feltene som er brukt er de feltene som kan inneholde tittelen på et verk. Dette betyr at alle mulige verkskandidater vil bli hentet ut, med mindre andre felter som ikke inneholder titler er brukt til titler. Av den grunn vil det være irrelevant å beregne recall, beskrevet i 3.5.2, da den vil være hundre prosent. Det er interessant å finne antallet verkskandidater som er hentet ut som er relevante. Siden det i dette tilfellet ikke eksiterer noen negativer, altså poster som ikke er identifisert som verk, vil både precision 3.5.1 og accuracy 3.5.3 bli det samme tallet. Derfor er precision brukt for å validere de forskjellige metrikkene for verkskandidatene. I dette tilfellet gir precision et tall som sier noe om sannsynligheten for at posten som er hentet ut er et verk. Dette tallet vil videre kunne brukes til å anslå sannsynligheten for at verk som har visse karakteristikker faktisk er et verk. Hvis en karakteristikk av verkskandidatene har en precision på 0,9 betyr det at verkskandidater som har den samme karakteristikken har 90 prosent sannsynlighet for å kunne representere et verk.

5.6 Testing

For å enklere kunne identifisere riktige verk har det i denne oppgaven blitt sett etter hvilke karakteristikker hos verkskandidatene som kan være en kilde til identifikasjon av et verk. Dette er ikke et definitivt mål på hva som er, og ikke er, et verk, men det vil være mulig å sette et tall på sannsynligheten for at en post representerer et verk.

Måten dette har blitt gjort på er å foreta et representativt utvalg av verkskandidater. Verkskandidatene som ble valgt ble markert basert på om de kunne representere et verk eller ikke. Duplikater ble også identifisert. Deretter ble forskjellige karakteristikker som kan brukes til å identifisere verk valgt ut. Postene ble analysert basert på karakteristikkene de hadde, og deretter ble disse karakteristikkene målt opp mot antallet verkskandidater som hadde karakteristikken som kunne representere et verk.

5.6.1 Dataene som er blitt analysert

Datasettet som står igjen etter FRBRiseringen består av over 43 millioner verkskandidater. I denne oppgaven er rundt 1600 verkskandidater analysert og karakteristikkene til verkskandidatene er blitt undersøkt. Det kan være vanskelig å finne et utvalg av verkskandidater som kan representere hele katalogen, og derfor er det valgt ut verkskandidater som best mulig kan representere katalogen så nøyaktig som mulig. Dataene er hentet ut for best mulig å kunne finne indikasjoner mellom karakteristikkene og riktigheten. Først ble verkskandidatene som hadde blitt slått sammen av 100, 90, 80, ... , og 10 verkskandidater hentet ut. På grunn av det høye antallet verkskandidater som

bestod av 30, 20 og 10 FRBR-poster, ble det gjort et tilfeldig utvalg av 100 verkskandidater. Siden samlingen i hovedsak består av verkskandidater som kun består av én FRBR-post, ble det valgt ut 200 tilfeldige verkskandidater som ikke hadde blitt slått sammen av flere FRBR-poster. Til slutt ble alle verkskandidatene som var knyttet til én og samme forfatter, Agatha Christie, hentet ut. Av de totalt 1400 som var relatert til Agatha Christie ble 600 tilfeldige verkskandidater brukt.

5.6.2 Hypoteser

Postene som har blitt analysert i oppgaven har flere interessante karakteristikker som kan si noe om sannsynligheten for at en post representerer et verk. Derfor var ønsket å undersøke disse karakteristikkene. Hypotesene for prosjektet er følgende:

H_1: *Har antallet forekomster av et verk noe å si på sannsynligheten for at en verkskandidat kan representere et verk?*

H_2: *Har verkskandidater som står oppført med forfatter høyere sannsynlighet for å kunne identifiseres som et verk?*

H_3: *Kan hvilken katalog verkskandidaten har opprinnelse i si noe om sannsynligheten for at verkskandidaten kan representere et verk?*

H_4: *Kan antallet forskjellige kataloger verkskandidaten har opprinnelse i si noe om sannsynligheten for at verkskandidaten kan representere et verk?*

H_5: *Kan hvilke felt verkskandidaten har opprinnelse i si noe om sannsynligheten for at verkskandidaten kan representere et verk?*

H_6: *Kan antallet forskjellige felt verkskandidaten har opprinnelse i si noe om sannsynligheten for at verkskandidaten kan representere et verk?*

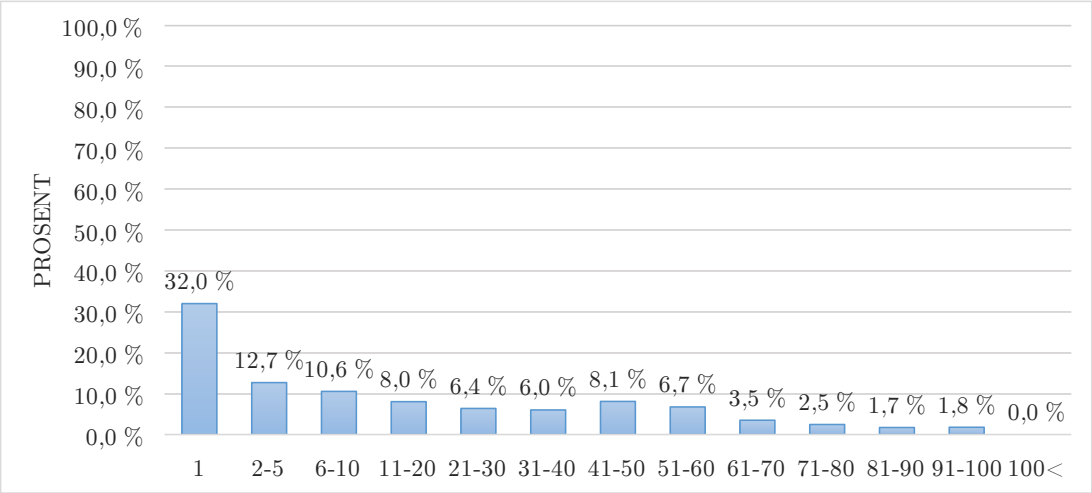
5.6.3 Fordeling

Hvis det eksisterer en eller flere karakteristikker i flere av postene som kan representere et verk, enn i postene som ikke kan det, vil dette kunne brukes til å si noe om verkene som kan representere et verk. Men det er ikke bare karakteristikkene som er viktige i denne sammenhengen.

Antallet poster som har de forskjellige karakteristikkene er også et viktig aspekt, fordi jo flere poster som har karakteristikkene, jo flere poster vil kunne identifiseres som verk. Hvis det for eksempel kun eksisterer én post som har en karakteristikk og denne posten representerer et verk, vil det ikke kunne være et representativt resultat. Man vil heller ikke kunne bruke karakteristikken som et effektivt verktøy for å identifisere verk i andre kataloger. Nedenfor vises karakteristikkene som har blitt analysert og en oversikt over fordelingen av verkene som har disse karakteristikkene.

Antall

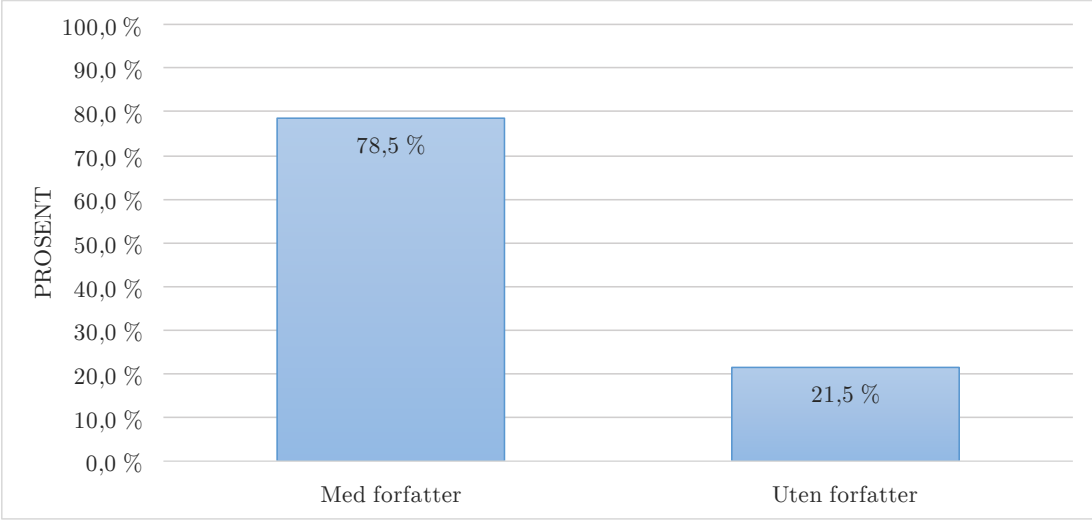
Figur 5.1 viser den prosentvise fordelingen av forekomster av verket. Dette gjelder antallet poster som har blitt slått sammen til én verkskandidat.



Figur 5.1: Den prosentvise fordelingen av verkskandidater og antallet FRBRposter de er slått sammen av.

Forfatter

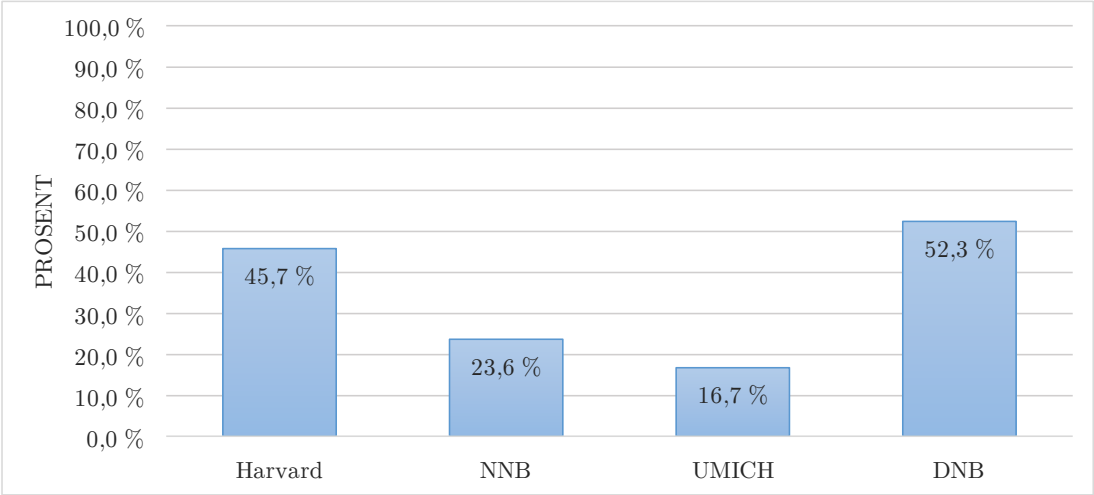
Figur 5.2 viser den prosentvise fordelingen av verkskandidatene som står oppført med en forfatter, eller ikke.



Figur 5.2: Den prosentvise fordelingen av verkskandidater med en relasjon til en forfatter.

Forekomsten i forskjellige kataloger

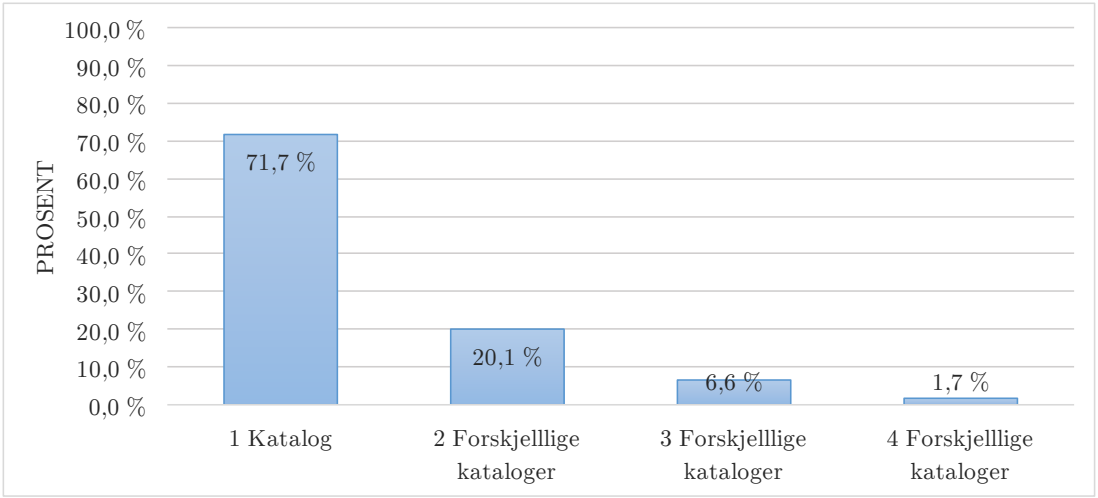
Figur 5.3 viser den prosentvise fordelingen av katalogene verkskandidatene har opprinnelse fra.



Figur 5.3: Den prosentvise fordelingen av katalogene verkskandidatene har opprinnelse i.

Antall forekomster i forskjellige kataloger

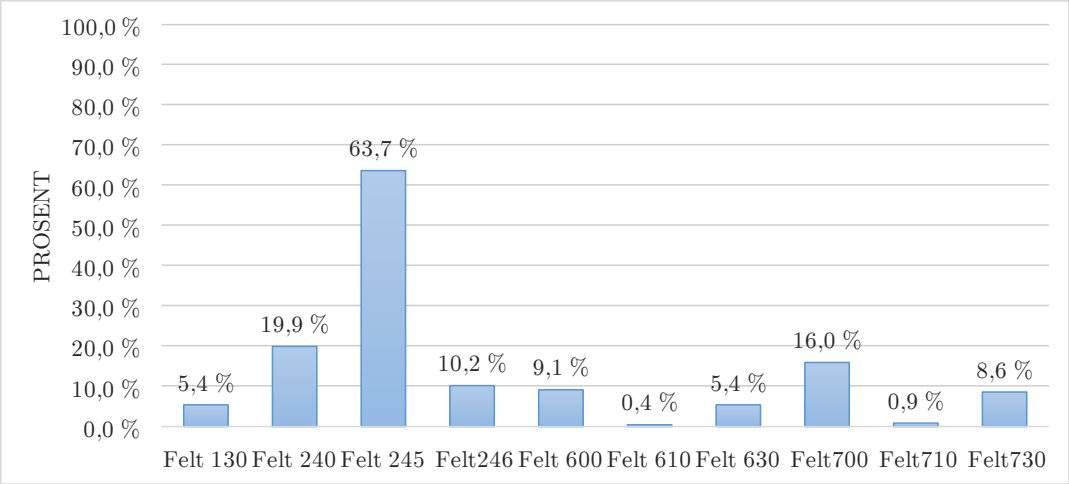
Figur 5.4 viser den prosentvise fordelingen av verkskandidater som har opprinnelse i 1, 2, 3 og 4 forskjellige kataloger.



Figur 5.4: Den prosentvise fordelingen av verkskandidater som har opprinnelse i 1, 2, 3 og 4 forskjellige kataloger.

Feltbruk

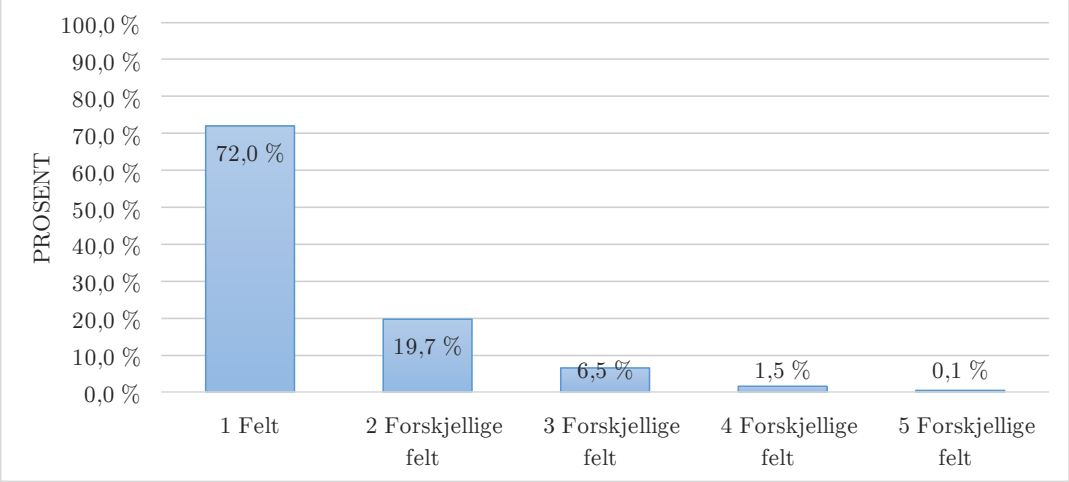
Figur 5.5 viser den prosentvise fordelingen av feltene verkskandidatene kommer fra. Feltene som har blitt sett på er beskrevet i 5.2.



Figur 5.5: Den prosentvise fordelingen av feltene verkskandidatene kommer fra.

Antall ulike felt som er brukt

Som tidligere nevnt kan en post bli hentet ut fra ulike felt, og etter sammenslåingen vil hver post inneholde alle feltene den har blitt hentet ut fra. Figur 5.6 viser den prosentvise fordelingen av verkskandidater som kommer fra 1, 2, 3, 4 eller 5 forskjellige felt.



Figur 5.6: Den prosentvise fordelingen av verkskandidater som kommer fra 1, 2, 3, 4 eller 5 forskjellige felt.

Kapittel 6

Resultater

Dette kapittelet presenterer resultatene som har blitt produsert i prosjektet. Først blir noen begreper forklart, og deretter presenteres resultatene fra analysene.

6.1 Oversikt

Først vil en oversikt over terminologi og karakteristikker bli presentert.

6.1.1 Terminologi

Entitet

Når det snakkes om en entitet i dette kapittelet beskriver det gjenstanden en post representerer. Entiteten kan være en bok, en person, en film, en sang, en manifestasjon, et uttrykk og så videre. Alle postene i en katalog representer en entitet.

Post

Rådataene som har blitt hentet ut i MARC21-format består av poster som representerer entiteter i de gitte bibliotekene. Figur 6.1 viser et eksempel på en post for gjenstanden *Mord på Orientekspressen* av Agatha Christie. Dette eksempelet viser kun informasjonen som er relevant. Posten bruker feltene 240, 245, 246, 700 og det vil derfor bli generert én FRBRposter for denne ene MARC21-posten i FRBRiseringen. FRBRposten vil bli laget på grunn av 240-feltet. Hvis 700-feltet også hadde inneholt underfeltet *t*, ville dette også kunne produsert en FRBRpost.

```xml
<record>
    <datafield ind1="1" ind2=" " tag="100">
        <subfield code="a">Christie, Agatha</subfield>
        <subfield code="d">1890-1976</subfield>
        <subfield code="0">(NO-TrBIB)90057229</subfield>
    </datafield>
    <datafield ind1="1" ind2="0" tag="240">
        <subfield code="a">Murder on the Orient Express</subfield>
        <subfield code="l">Norsk</subfield>
    </datafield>
    <datafield ind1="1" ind2="0" tag="245">
        <subfield code="a">Mord på Orientekspressen</subfield>
        <subfield code="c">Agatha Christie ; oversatt av Axel S. Seeberg
            </subfield>
    </datafield>
    <datafield ind1="1" ind2=" " tag="246">
        <subfield code="a">Murder on the Orient express</subfield>
        <subfield code="i">Originaltittel</subfield>
    </datafield>
    <datafield ind1="1" ind2=" " tag="700">
        <subfield code="a">Seeberg, Axel S.</subfield>
        <subfield code="d">1914-1994</subfield>
        <subfield code="4">trl</subfield>
        <subfield code="0">(NO-TrBIB)90707254</subfield>
    </datafield>
</record>
```

Figur 6.1: Eksempel på en MARC21-post fra NNB.

FRBRpost

Når disse postene blir FRBRisert opprettes det nye midlertidige poster som passer FRBR-modellen. I dette prosjektet har de midlertidige postene blitt kalt for FRBRposter for å hindre forvirring mellom MARC-postene og FRBRpostene. Figur 6.2 er et eksempel på en FRBRpost. Også her vises kun informasjonen som er relevant for dette eksempelet. Denne nye FRBRposten er laget basert på informasjonen fra en MARC21-post som bruker feltet 245. FRBRposten vil i seg selv ikke representere en gjenstand eller et verk, men den kan brukes til å identifisere verk som senere vil kunne brukes til å katalogisere andre poster med samme tittel, forfatter og så videre. Det er FRBRposter som denne, som blir slått sammen til verkskandidater.

```
<record id="christieagatha18901976\cperson#murderontheorientexpress\cwork#"
        type="c:Work" templatename="MARC21-245-Work">
    <frbrizer:confidence rule="MARC21-245-Work" source_confidence="NNB.xml"/>
    <datafield tag="245" ind1="0" ind2="0">
        <subfield code="a"
                type="w:titleOfTheWork">Murder on the Orient Express</subfield>
    </datafield>
    <frbrizer:relationship type="w:author"
    itype="a:authorOf" href="christieagatha18901976\cperson#"/>
</record>
```

Figur 6.2: Eksempel på en FRBRpost på RDF-format.

Verkskandidat

Etter at postene med lik nøkkel blir slått sammen er resultatet ulike verkskandidater. En verkskandidat består av én eller flere poster som er slått sammen, og verkskandidaten inneholder informasjonen som fantes i alle postene som var med i sammenslåingen. Figur 6.3 viser et eksempel på en verkskandidat for verket *Mord på Orientekspressen*. Figuren viser hvordan denne verkskandidaten er slått sammen av poster som er laget fra feltene 240, 245 og 246. Postene som er slått sammen kommer fra katalogene DNB, Harvard og NNB og til sammen er 20 poster blitt slått sammen til denne verkskandidaten. Verkskandidaten er ikke et verk i seg selv, men tittelen verkskandidaten representerer kan representere et verk.

```
1
2  <rdf:Description rdf:about="6130e6e9-6030-3bbb-9117-9aba230c20f5"
3                   rdf:type="c:Work">
4      <w:titleOfWork>Murder on the Orient Express</w:titleOfWork>
5      <w:titleOfWork>Murder on the Orient express.</w:titleOfWork>
6      <w:titleOfWork>Murder on the Orient express</w:titleOfWork>
7      <w:author rdf:resource="99b0c1ba-953c-39a7-a8de-828b3e08ca1c"/>
8      <frbrizer:hasConfidence>
9          <frbrizer:Confidence frbrizer:rule="MARC21-245-Work"
10                              frbrizer:src="dnb" frbrizer:cnt="1"/>
11     </frbrizer:hasConfidence>
12     <frbrizer:hasConfidence>
13         <frbrizer:Confidence frbrizer:rule="MARC21-245-Work"
14                              frbrizer:src="harvard" frbrizer:cnt="5"/>
15     </frbrizer:hasConfidence>
16     <frbrizer:hasConfidence>
17         <frbrizer:Confidence frbrizer:rule="MARC21-240-Work"
18                              frbrizer:src="harvard" frbrizer:cnt="2"/>
19     </frbrizer:hasConfidence>
20     <frbrizer:hasConfidence>
21         <frbrizer:Confidence frbrizer:rule="MARC21-246-Work"
22                              frbrizer:src="nbn" frbrizer:cnt="9"/>
23     </frbrizer:hasConfidence>
24     <frbrizer:hasConfidence>
25         <frbrizer:Confidence frbrizer:rule="MARC21-240-Work"
26                              frbrizer:src="nbn" frbrizer:cnt="3"/>
27     </frbrizer:hasConfidence>
28     <frbrizer:hasConfidence>
29         <frbrizer:Confidence frbrizer:total="20"
30                              frbrizer:type="c:Work"/>
31     </frbrizer:hasConfidence>
32 </rdf:Description>
```

Figur 6.3: Eksempel på en verkskandidat som markeres som TPS på RDF-format.

TPS

True Positive Singular, TPS, er en markering for verkskandidatene som er riktig identifisert som et verk og som ikke allerede har blitt identifisert. For at en verkskandidat skal bli markert som TPS må den inneholde den riktige originaltittelen til et verk. Figur 6.3 som er brukt over er et eksempel på en verkskandidat som vil bli markert som TPS.

TPD

True Positive Duplicate, TPD, er en markering for de verkskandidatene som er riktig identifisert som et verk, men som allerede har blitt identifisert. Det vil si at det finnes en verkskandidat som representerer det samme verket, men som består av et høyere antall poster som har blitt slått sammen. Disse verkskandidatene er riktig identifiserte, men de har ikke blitt fanget opp under

dedupliseringen fordi de avviker på en eller annen måte. Figur 6.4 er et eksempel på en verkskandidat for verket *Mord på Orienteksspressen* som markeres som TPD, fordi det har en *total* på 9, som er tallet på antall poster som er slått sammen. Dette tallet er lavere enn *totalen* til posten i Figur 6.3 som har en total på 20, og i denne oppgaven blir verkskandidatene med høyest total markert først.

```
<rdf:Description rdf:about="2ed992de-1981-3eb6-8099-11cf2834fcc6"
                 rdf:type="c:Work">
    <w:titleOfWork>Murder on the Orient Express</w:titleOfWork>
    <w:titleOfWork>Murder on the Orient Express.</w:titleOfWork>
    <frbrizer:hasConfidence>
        <frbrizer:Confidence frbrizer:rule="MARC21-730-Work"
                             frbrizer:src="dnb" frbrizer:cnt="5"/>
    </frbrizer:hasConfidence>
    <frbrizer:hasConfidence>
        <frbrizer:Confidence frbrizer:rule="MARC21-245-Work"
                             frbrizer:src="dnb" frbrizer:cnt="1"/>
    </frbrizer:hasConfidence>
    <frbrizer:hasConfidence>
        <frbrizer:Confidence frbrizer:rule="MARC21-130-Work"
                             frbrizer:src="harvard" frbrizer:cnt="1"/>
    </frbrizer:hasConfidence>
    <frbrizer:hasConfidence>
        <frbrizer:Confidence frbrizer:rule="MARC21-730-Work"
                             frbrizer:src="harvard" frbrizer:cnt="1"/>
    </frbrizer:hasConfidence>
    <frbrizer:hasConfidence>
        <frbrizer:Confidence frbrizer:rule="MARC21-245-Work"
                             frbrizer:src="harvard" frbrizer:cnt="1"/>
    </frbrizer:hasConfidence>
    <frbrizer:hasConfidence>
        <frbrizer:Confidence frbrizer:total="9" frbrizer:type="c:Work"/>
    </frbrizer:hasConfidence>
</rdf:Description>
```

Figur 6.4: Eksempel på en verkskandidat som markeres som TPD på RDF-format.

FP

False Positive, FP, er en markering for de verkskandidatene som ikke representerer et verk. Dette gjelder verkskandidater som i seg selv ikke er et verk, eller som ikke baserer seg på et verk. Det gjelder også verkskandidater som er basert på et verk, men som ikke har originaltittel. Figur 6.5 er et eksempel på en verkskandidat som representerer en samling av historier skrevet av Agatha Christie. Dette representerer ikke et verk fordi det er en samling av verk, og blir derfor markert som FP.

```
<rdf:Description rdf:about="a178b2eb-2558-3563-9737-f10466ef478f"
                 rdf:type="c:Work">
    <w:titleOfWork >Witness for the prosecution:and other stories</:titleOfWork>
    <w:creator rdf:resource="99b0c1ba-953c-39a7-a8de-828b3e08ca1c"/>
    <frbrizer:hasConfidence
        xmlns:frbrizer="http://idi.ntnu.no/frbrizer/">
        <frbrizer:Confidence frbrizer:rule="MARC21-245-Work"
                             frbrizer:src="harvard"
                             frbrizer:cnt="1"/>
    </frbrizer:hasConfidence>
    <frbrizer:hasConfidence xmlns:frbrizer="http://idi.ntnu.no/frbrizer/">
        <frbrizer:Confidence frbrizer:total="1" frbrizer:type="c:Work"/>
    </frbrizer:hasConfidence>
</rdf:Description>
```

Figur 6.5: Eksempel på en verkskandidat som markeres som FP på RDF-format.

Figur 6.6 er et eksempel på en verkskandidat som representerer verket *Mord på Orientekspressen* med norsk tittel. I dette prosjektet er kun verkskandidater med riktig originaltittel definert til å kunne representere et verk. Derfor vil verkskandidaten vist i figuren bli markert som FP. Verkskandidaten kan representere en manifestasjon av verket, men ikke verket i seg selv.

```
<rdf:Description rdf:about="a178b2eb-2558-3563-9737-f10466ef478f"
                 rdf:type="c:Work">
    <w:titleOfWork >Witness for the prosecution:and other stories</:titleOfWork>
    <w:creator rdf:resource="99b0c1ba-953c-39a7-a8de-828b3e08ca1c"/>
    <frbrizer:hasConfidence
        xmlns:frbrizer="http://idi.ntnu.no/frbrizer/">
        <frbrizer:Confidence frbrizer:rule="MARC21-245-Work"
                             frbrizer:src="harvard"
                             frbrizer:cnt="1"/>
    </frbrizer:hasConfidence>
    <frbrizer:hasConfidence xmlns:frbrizer="http://idi.ntnu.no/frbrizer/">
        <frbrizer:Confidence frbrizer:total="1" frbrizer:type="c:Work"/>
    </frbrizer:hasConfidence>
</rdf:Description>
```

Figur 6.6: Eksempel på en verkskandidat som markeres som FP på RDF-format.

True negative og False negative

Dataene som har blitt sett på i denne oppgaven er alle hentet ut fordi feltene de er hentet fra har en mulighet til å inneholde et verk. Det betyr at postene som har blitt hentet ut er positiver. Det er derfor sannsynlig at postene kan representere et verk. Postene som ikke har blitt hentet ut som kan være verkskandidater er det som blir kalt negativer, enten *True negative*, TN, eller *False negative*, FN. Ingen av postene som har blitt brukt i denne oppgaven vil gå inn i de gruppene. Det er ikke mulig å si noe om poster som er FN fordi det ikke eksisterer en fasit for postene som sier om de er riktige eller ikke.

TN gjelder de postene som ikke har blitt hentet ut og som ikke kan representere et verk. FN gjelder de postene som ikke har blitt hentet ut, men som representerer et verk. Feltene som er brukt

i denne oppgaven for å finne titler på verk er hentet ut fordi det nettopp er disse feltene som blir brukt til titler. Derfor vil postene som er FN i dette datasettet gjelde poster som har brukt felt som ikke skal inneholde en tittel, til å inneholde en tittel.

Representere et verk

Hvis en verkskandidat er riktig identifisert, betyr det at den kan representere et verk og at verkskandidaten kan brukes til å lage en post som holder informasjonen til verket. Videre kan andre entiteter i FRBR som for eksempel person, uttrykk eller manifestasjon relateres til dette verket. En verkskandidat som er en oversettelse av et originalverk vil være et uttrykk av et verk, men det vil ikke kunne representere selve originalverket.

Precision

I denne oppgaven er precision blitt brukt til å gi et tall på riktigheten til verkskandidatene og karakteristikkene de innehar. Precisionen regnes ut basert på antallet verkskandidater som blir merket som TPS, TPD og FP. Tallet som kommer ut ligger mellom 0 og 1, og jo høyere tallet er, jo høyere sannsynlighet er det for at verkskandidaten representerer et verk. I grafene som viser precisionen til de forskjellige karakteristikkene viser den blå søylen precisionen til karakteristikken kun basert på TPS og FP. Den gule delen av grafen viser endringen av precision hvis TPD også blir tatt med i utregningen. Verkskandidatene som er merket med TPD er også riktig identifisert, men de er duplikater av allerede identifiserte verk. Den totale høyden på grafen vil med andre ord være precisionen til alle riktig identifiserte verk, sett bort i fra at noen er duplikater.

6.1.2 Karakteristikker

Totalt ble 1625 forskjellige verkskandidater manuelt identifisert. Metrikken som blir brukt i denne oppgaven for å si noe om kvaliteten baserer seg på antall riktige mot antall gale identifiseringer. Tallet som blir beregnet er precision, og precision beskriver hvor mange av de uthentede postene som er relevante. Videre vil også de korrekte duplikatene, TPD, bli vurdert. Disse vil være med i beregningen ved at verkskandidatene som er markert som TPD vil regnes som riktig identifiserte. Forskjellen disse verkskandidatene gjør i precision-scoren er markert i søylene.

Av de 1625 kandidatene ble 797 identifisert som TPS, 89 som TPD og 739 som FP og gir en precision på 0,54. Dette viser at å hente ut poster basert på kun de utvalgte feltene beskrevet i 5.2, for så å slå dem sammen basert på den konstruerte nøkkelen, lager verkskandidater som har 54 prosent sannsynlighet for å være et verk. Dette er en precision som vil gi dårlige resultater i en FRBRiseringsprosess. Nå skal hypotesene fra 5.6.2 testes for å finne ut om det eksisterer data fra verksidentifiseringen som kan brukes til å angi påliteligheten til de genererte verkene.

6.2 Antall forekomster

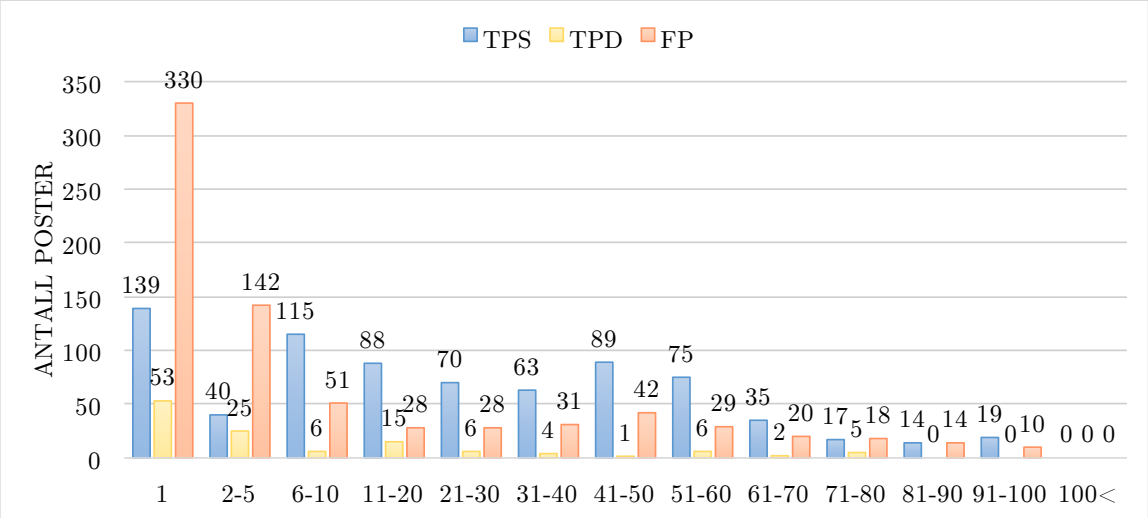
H: *Har antallet forekomster av et verk noe å si på sannsynligheten for at en verkskandidat kan representere et verk?*

Som beskrevet i 5.3.2 blir FRBRposter som har lik nøkkel slått sammen til én post. Antallet FRBRposter som blir slått sammen lagres i den sammenslåtte verkskandidaten. Tallene som har blitt undersøkt sier hvor mange FRBRposter som har blitt slått sammen. Dette er FRBRpostene som har blitt generert i FRBRiserinen, og hver FRBRpost vil ikke representere én enkelt post fra originaldataene på MARC21-formatet. Fordi verkskandidater kan bestå av flere FRBRposter, vil tallene som har blitt undersøkt for antall forekomster derfor ikke nødvendigvis representere antall poster som eksisterer i katalogene, men antall felt som inneholder den entiteten.

Dataene viser at det er stor variasjon i antall forekomster av et verk. For noen verk finnes det mange uttrykk og manifestasjoner, mens andre bare har ett uttrykk. Hvis det eksisterer flere poster av samme entitet, eller det eksisterer flere poster som inneholder informasjon om lignende entiteter, kan det tyde på at disse entitetene bygger på et verk. Det kan derfor være interessant å se om verkskandidater som har et høyt antall forekomster har høy sannsynlighet for å representere et verk.

6.2.1 Fordeling

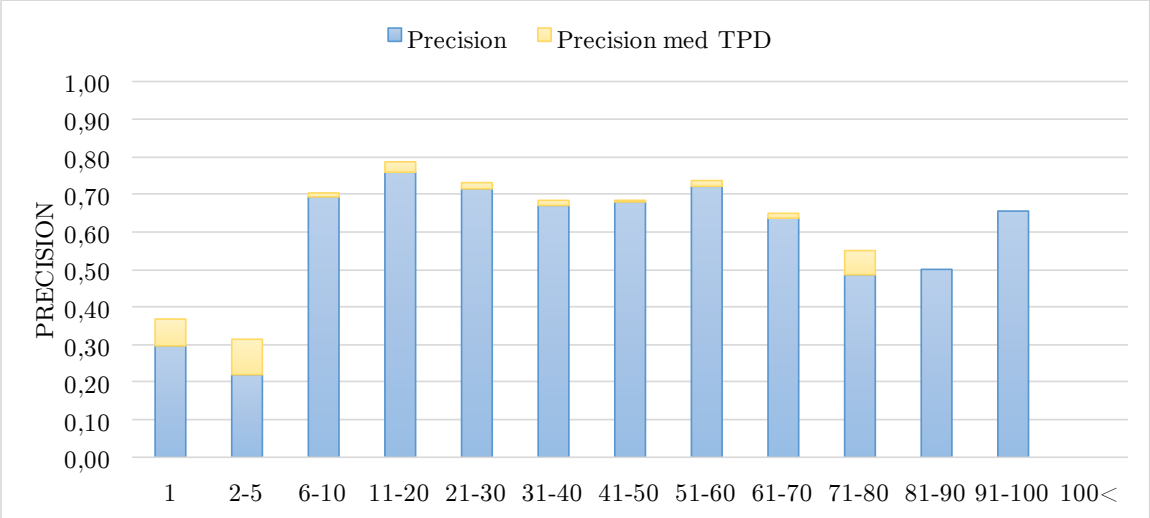
Fordi antallet har stor spredning, er verkskandidatene fordelt i ulike bolker som representerer forskjellige spekter. Den første bolken består kun av verkskandidater som ikke har blitt sammenslått. Dette er en gruppe som i seg selv er interessant å se på og er derfor alene i bolken. Videre deles verkskandidatene som er sammenslått av 2-5 FRBRposter inn i én bolk, 6-10 er en annen bolk og deretter er verkskandidatene delt inn etter 11-20, 21-30 og så videre. På bakgrunn av at noen av verkskandidatene ble hentet ut basert på antallet som beskrevet i 5.6.1, vil bolkene i hovedsak inneholde verkskandidater som nettopp består av akkurat 10, 20, 30 og så videre sammenslåinger. Dette er med unntak av verkskandidatene med relasjon til Agatha Christie, da disse består av et tilfeldig antall. Figur 6.7 viser antallet og fordelingen av verkskandidatene som har blitt identifisert som TPS, TPD og FP.



Figur 6.7: Fordelingen av TPS, TPD og FP for antall FRBRposter som er sammenslått til verkskandidatene.

6.2.2 Precision

Som nevnt tidligere er det metrikken precision som er brukt for å si noe om riktigheten til verkskandidatene og om karakteristikkene som er brukt. Når det kommer til antallet FRBRposter som er slått sammen er det et tydelig skille i dataene. Verkskandidatene som er slått sammen av 1 eller mellom 2 og 5 FRBRposter har en veldig lav precision på under 0,4. Denne precisionen er lavere enn gjennomsnittet til hele samlingen som er 0,54. Resten av verkskandidatene har en jevnt høy precision uten store avvik. Figur 5.1 og Figur 6.7 viser at de to første gruppene, 1 og 2-5, inneholder nærmere 45 prosent av verkskandidatene.



Figur 6.8: Precisionen til verkskandidatene gruppert etter antallet FRBRposter de er sammenslått av.

6.2.3 Analyse

Tallene viser at det er et tydelig skille mellom verkskandidatene som er slått sammen av 1-5 FRBRposter, og de som har blitt slått sammen av flere enn 5 FRBRposter. Dette kan være en god indikasjon for identifisering av verk. Verkskandidater som har flere forekomster enn 5 har en høy precision, noe som viser at hvis en verkskandidat har flere forekomster vil dette øke sannsynligheten for at den kan identifiseres som et verk. Utfordringen er at de fleste verkskandidatene har et lavt antall forekomster. I dette datasettet består nesten 50 prosent av verkskandidatene av mellom 1 og 5 sammenslåinger. På den andre siden kan dette brukes som en indikasjon på hvilke verkskandidater som ikke representerer et verk, siden denne gruppen verkskandidater har en precision på under 40 prosent.

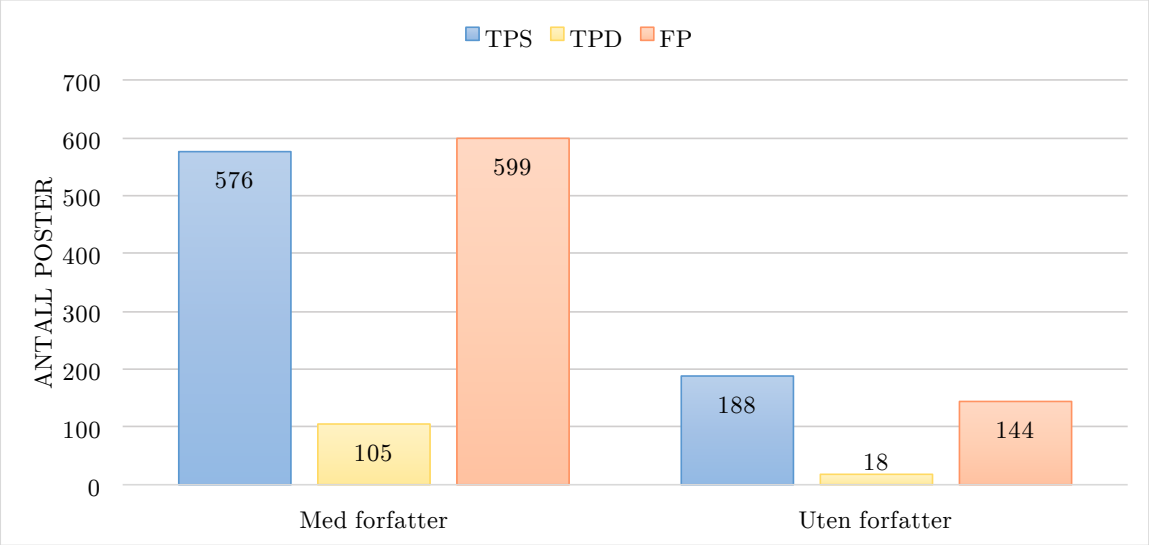
6.3 Med forfatter

H: *Har verkskandidater som står oppført med forfatter høyere sannsynlighet for å kunne identifiseres som et verk?*

Postene som har blitt sett på inneholder forskjellige felt og forskjellig informasjon. Til tross for at postene som har blitt hentet ut representerer bøker og lignende, står ikke alle postene oppført med forfatter. Etter sammenslåingen står mange av verkskandidatene oppført med en forfatter eller en skaper, og det er interessant å se om verkskandidatene har en høyere sannsynlighet for å kunne representere et verk.

6.3.1 Fordeling

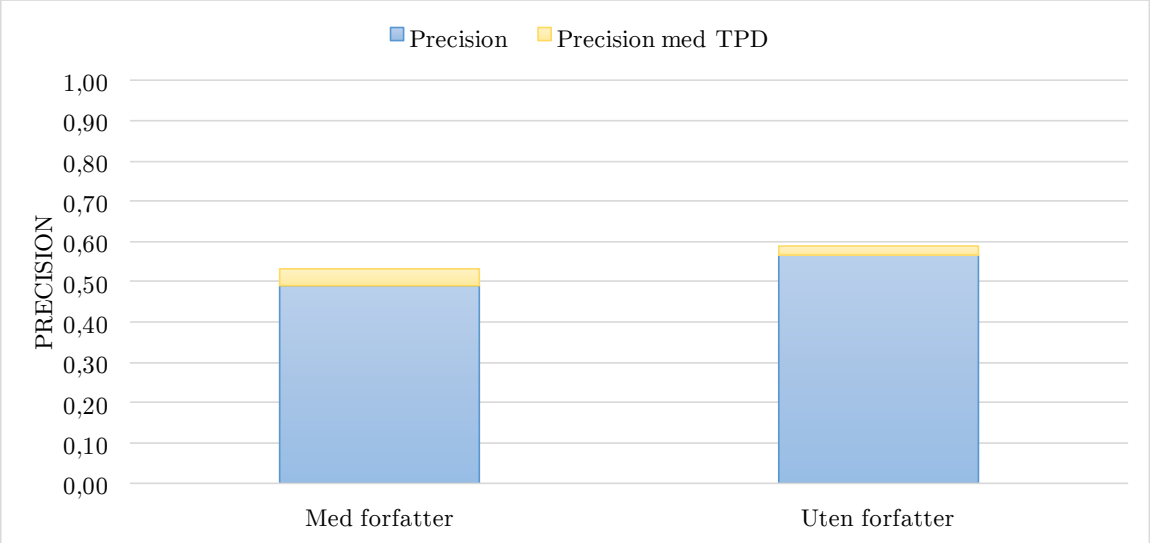
Figur 6.9 viser fordelingen av TPS, TPD og FP for verkskandidatene som står oppført med forfatter, og for verkskandidatene som er oppført uten forfatter. Det er absolutt flest verkskandidater som står oppført med forfatter, med nesten 80 prosent. Fordi en del av postene ble hentet ut på bakgrunn av at de hadde en relasjon til Agatha Christie, vil alle disse postene stå oppført med en forfatter. Tallene vil derfor kunne være påvirket av at noen flere poster står oppført med forfatter enn det som gjelder for hele samlingen.



Figur 6.9: Fordelingen av TPS, TPD og FP for verkskandidater som er oppført med forfatter eller ikke.

6.3.2 Precision

Figur 6.10 viser precisionen til verkskandidatene som står oppført med forfatter og til verkskandidater som ikke står oppført med forfatter. Det viser seg at begge karakteristikkene har en lav precision og postene som ikke er oppført med forfatter har en litt høyere precision enn postene som står oppført med forfatter.



Figur 6.10: Precisionen til verkskandidatene som står oppført med forfatter og uten forfatter.

6.3.3 Analyse

Det viser seg at begge gruppene i denne karakteristikken har en lav precision. Det er ingenting som stikker seg ut i resultatene, og overraskende nok ser det ut til at verkskandidatene som ikke er oppført med forfatter har en høyere precision enn de som står oppført med forfatter. Hverken verkskandidatene som står oppført med forfatter eller de som står oppført uten forfatter har en precision som ligger rundt precisionen til den totale samlingen. Dette viser at denne karakteristikken ikke gir noen indikasjon på om en post representerer et verk eller ikke.

6.4 Kataloger

H_1: *Kan hvilken katalog verkskandidaten har opprinnelse i si noe om sannsynligheten for at verkskandidaten kan representere et verk?*

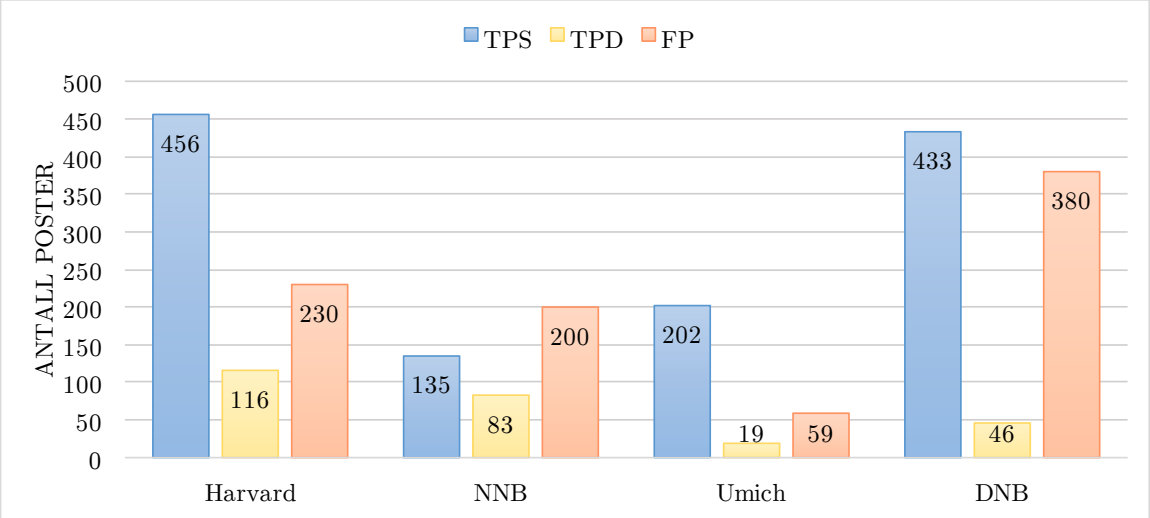
H_2: *Kan antallet forskjellige kataloger verkskandidaten har opprinnelse i si noe om sannsynligheten for at verkskandidaten kan representere et verk?*

Som nevnt i 5.2.1 er det hentet ut data fra fire forskjellige kataloger: Harvard Library(Harvard), University of Michigan(UMICH), den tyske nasjonalbibliografien(DNB) og den norske nasjonalbibliografien(NNB). Etter sammenslåingen av FRBRposter vil noen av verkskandidatene ha blitt slått sammen av FRBRposter fra forskjellige kataloger, og dette blir også lagret i verkskandidaten. Denne analysen er ikke ment for å teste de forskjellige katalogene opp mot hverandre, men for å se på forskjellene mellom forskjellige kataloger. Katalogene som er brukt kommer også fra forskjellige typer biblioteker der Harvard og UMICH er universitetsbiblioteker og DNB og NNB er nasjonalbiblioteker. For denne karakteristikken har det kun blitt sett på forekomsten i forskjellige kataloger, det har ikke blitt sett på hvor mange forekomster det er i de forskjellige katalogene.

En entitet som dukker opp i flere forskjellige kataloger med forskjellig nasjonalitet vil tyde på at entiteten er kjent utover nasjonaliteten den opprinnelig har. Den vil også med høy sannsynlighet ha blitt oversatt til andre språk. Det er interessant å se om entiteter som er kjent i andre land enn opprinnelseslandet har en høyere sannsynlighet for å kunne representere et verk, og å se om utbredelsen av entiteten har noe å si. Utbredelsen vil i dette tilfellet bli beskrevet som antall forekomster i forskjellige kataloger. En post som eksisterer i alle de fire katalogene vil ha stor utbredelse.

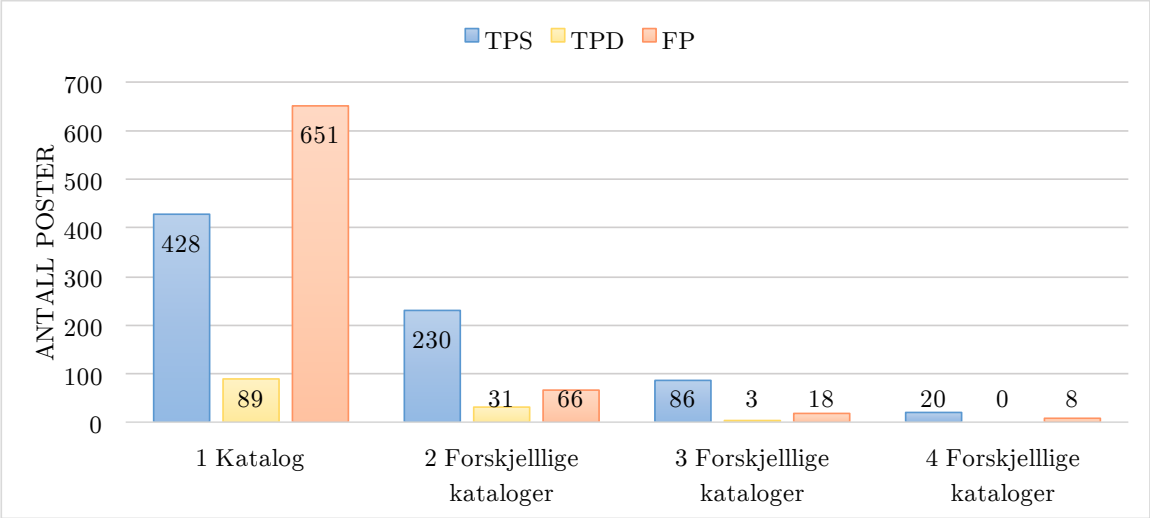
6.4.1 Fordeling

Figur 6.11 viser fordelingen av TPS, TPD og FP hos de fire forskjellige katalogene som er brukt. Grafen viser også at samlingen med data som er blitt analysert inneholder langt flere poster fra Harvard og DNB enn NNB og Umich, som gjenspeiler hele samlingen. Siden verkskandidatene er blitt slått sammen av flere FRBRposter vil en verkskandidat kunne være representert i flere av disse grafene.



Figur 6.11: Fordelingen av TPS, TPD og FP for verkskandidatene og katalogene de har opprinnelse fra.

Figur 6.12 viser fordelingen av TPS, TPD og FP for hvor mange forskjellige kataloger verkskandidatene har opprinnelse fra. Grafen viser hvordan det er en ujevn fordeling i antallet og at hovedmengden av verkskandidatene har opprinnelse i bare én katalog. En av grunnene til at det er få verkskandidater som har opprinnelse i flere forskjellige kataloger kan være på grunn av at to av katalogene kommer fra nasjonalbiblioteker og to kommer fra universitetsbiblioteker. Inneholdet i de forskjellige katalogene vil være forskjellig og det er derfor større sannsynlighet for at en verkskandidat som eksisterer i NNB også vil eksistere i DNB, enn at den vil eksistere i Harvard.

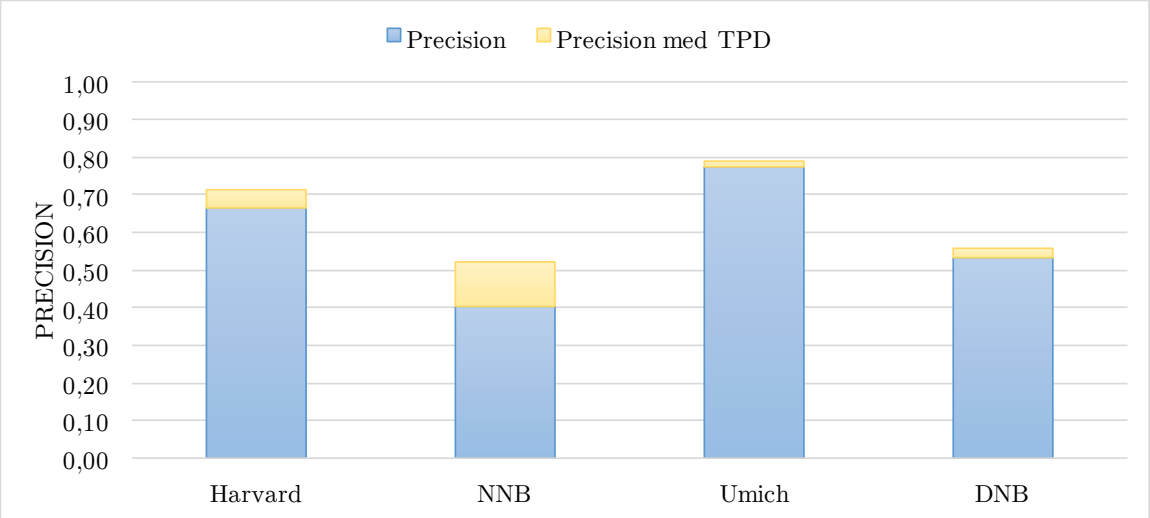


Figur 6.12: Fordelingen av TPS, TPD og FP for verkskandidatene basert på antallet forskjellige kataloger de har opprinnelse fra.

6.4.2 Precision

Først blir katalogene verkskandidatene har opprinnelse fra analysert. Figur 6.13 viser precisionen til verkskandidatene som har opprinnelse i de fire forskjellige katalogene. Figuren viser at University of Michigan har høy precision med nærmere 0,8, og et lite antall duplikater. NNB og DNB skiller seg ut ved at de har relativt lav precision der NNB har 0,52 som så vidt er lavere enn totalen for hele samlingen. DNB har heller ingen høy precision i motsetning til Harvard og Umich.

NNB har et høyt antall duplikater, men som alle vil kunne representere verk. Dette kan tyde på at verkskandidater som er hentet ut har variasjoner i andre felt enn feltene som representerer tittel. I 5.2.1 kan man se hvordan katalogene har ulike praksiser når det gjelder feltbruk og verdiene som er lagret i feltene. Dette kan også ha innvirkning på resultatet. Hvis katalogen inneholder mange poster på det nasjonale språket vil dette for eksempel ha negativ innvirkning på resultatene, fordi det vil være større hyppighet av verkskandidater med ikke-original tittel.



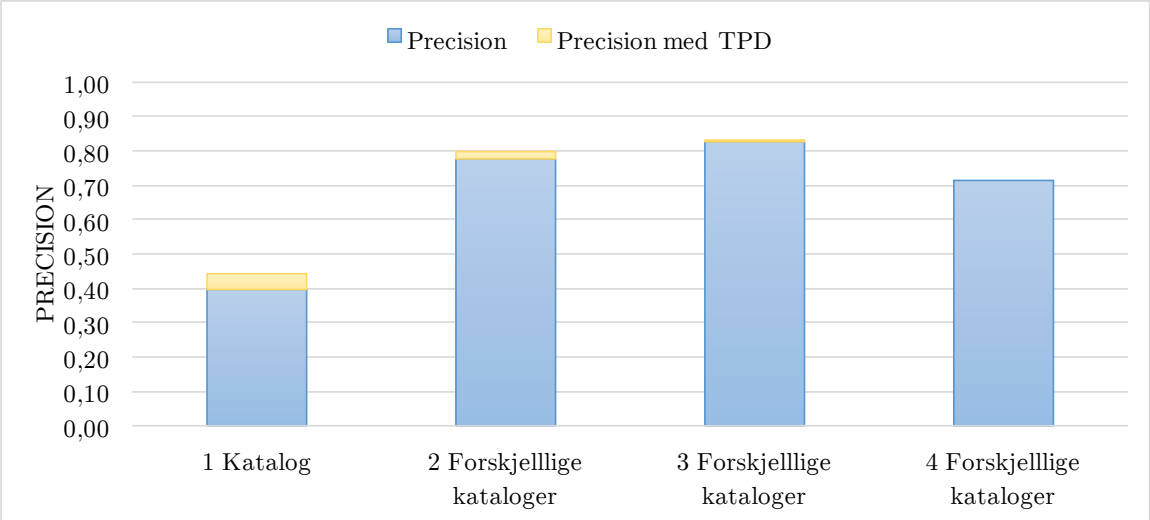
Figur 6.13: Precision for verkskandidater basert på hvilken katalog postene har opprinnelse fra.

Den neste karakteristikken som har blitt sett på er hvor mange forskjellige kataloger en verkskandidat har opprinnelse i. Figur 6.14 viser precisionen til verkskandidatene som har opprinnelse i 1, 2, 3 eller 4 forskjellige kataloger.

Grafen viser et skille mellom verkskandidatene som bare stammer fra én katalog, og de verkskandidatene som stammer fra to eller flere kataloger. Som Figur 6.11 viser er det aller flest verkskandidater som bare har opprinnelse i én katalog. Disse verkskandidatene utgjør over 60 prosent av alle verkskandidatene, og precisionen er på 0,44 som er lavere enn totalen for hele samlingen. Det kan derfor være mer interessant å se på verkskandidatene som stammer fra 2 eller flere forskjellige kataloger.

Figur 6.11 viser at det er få verkskandidater som har opprinnelse i 2 eller flere kataloger. Verkskandidatene som har opprinnelse i 3 og 4 kataloger representerer kun 8 prosent av alle postene, og tallene vil dermed ha en lavere pålitelighet. Verkskandidatene som har opprinnelse i 2 kataloger derimot, representerer 20 prosent av alle postene og har videre en precision på 0,8, som er høyt.

Deles karakteristikkene opp i to bolker der den ene gruppen inneholder verkskandidater som bare kommer fra én katalog, og den andre inneholder verkskandidatene som kommer fra 2 eller flere kataloger vil den ene bolken få en precision på 0,83. Denne gruppen inneholder verkskandidatene som kommer fra 2 eller flere kataloger.



Figur 6.14: Precision for postene basert på hvor mange forskjellige kataloger postene har opprinnelse i.

6.4.3 Analyse

I tallene over vil precisionen til de forskjellige katalogene si mer om de spesifikke samlingene, og gir derfor ikke en bedre forståelse av hvordan verk kan identifiseres. Forskjellene mellom katalogene kan skyldes forskjellig bruk av felt og så videre. Dette kan være interessant for bibliotekene katalogene kommer fra, men fordi tallene sier mer om selve katalogene, er det ikke interessant for denne oppgaven. Hvilken katalog en verkskandidat har opprinnelse i er derfor ingen god indikasjon på om verkskandidaten kan representere et verk.

Antallet forskjellige kataloger en verkskandidat har opprinnelse i viser tydelige forskjeller i precision. Skillet går mellom verkskandidatene som bare har opprinnelse i én katalog og de verkskandidatene som har opprinnelse i flere enn én katalog. Igjen er det en utfordring at de aller fleste verkskandidatene kun har opprinnelse i én katalog med over 70 prosent av verkskandidatene. Dette gjør at karakteristikken ikke nødvendigvis er den beste for å identifisere verk. Deles verkskandidatene inn i to grupper, som beskrevet over, har verkskandidatene som har opprinnelse i flere enn én katalog en høy sannsynlighet for å kunne representere et verk. Verkskandidater som har opprinnelse i én katalog har en precision på 0,44, og precisionen for verkskandidatene som har opprinnelse i flere enn én katalog er 0,8. Derfor vil dette være en god indikasjon på om en verkskandidat kan representere et verk.

6.5 Felt

H_1: *Kan hvilke felt verkskandidaten har opprinnelse i si noe om sannsynligheten for at verkskandidaten kan representere et verk?*

H_2: *Kan antallet forskjellige felt verkskandidaten har opprinnelse i si noe om sannsynligheten for at verkskandidaten kan representere et verk?*

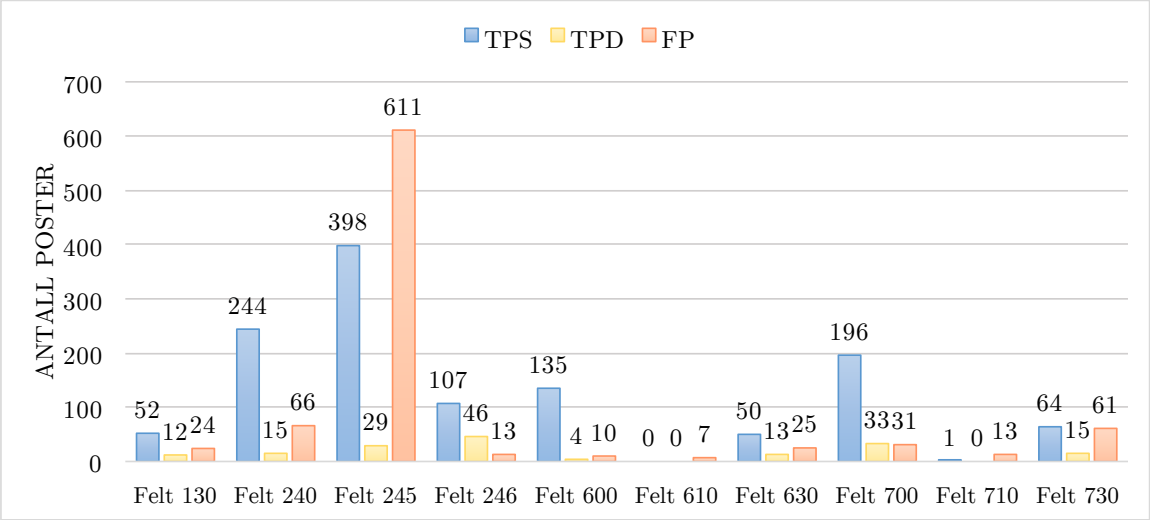
Postene er hentet ut basert på hvilke MARC21-felt de har brukt. Feltene som har blitt sett på er 130, 240, 245, 246, 600, 610, 630, 700, 710 og 730 som er beskrevet i 5.2. Disse feltene er ment for forskjellige bruksområder og inneholder forskjellig informasjon. Likheten mellom alle disse feltene er at de kan inneholde en form for tittel for en entitet. Bruken av feltene er ulik fra post til post, der noen poster inneholder flere av feltene og andre bare inneholder ett. Feltbruken i de forskjellige katalogene har noen forskjeller, men er stort sett lik. Derfor vil noen felt produsere verkskandidater som har større sannsynlighet for å representere et verk enn andre.

Når FRBRpostene blir slått sammen til verkskandidater blir feltene de har opprinnelse i lagret. Noen verkskandidater kan inneholde flere instanser fra det samme feltet, men i denne oppgaven har kun feltene i seg selv blitt sett på. Så om en verkskandidat har opprinnelse i flere forskjellige 245-felt vil dette ikke gjøre noen forskjell i resultatene.

I tillegg til å se på feltene hver for seg har verkskandidatene som har opprinnelse i flere forskjellige felt blitt sett på. Dette vil si verkskandidater som har blitt sammenslått av poster med enten 2, 3, 4 eller 5 ulike felt. Selv om en verkskandidat har opprinnelse i flere felt betyr ikke det nødvendigvis at den har opprinnelse i flere poster. Antallet ulike felt en entitet har opprinnelse i kan imidlertid si noe om utbredelsen til entiteten, noe som er interessant å finne ut av.

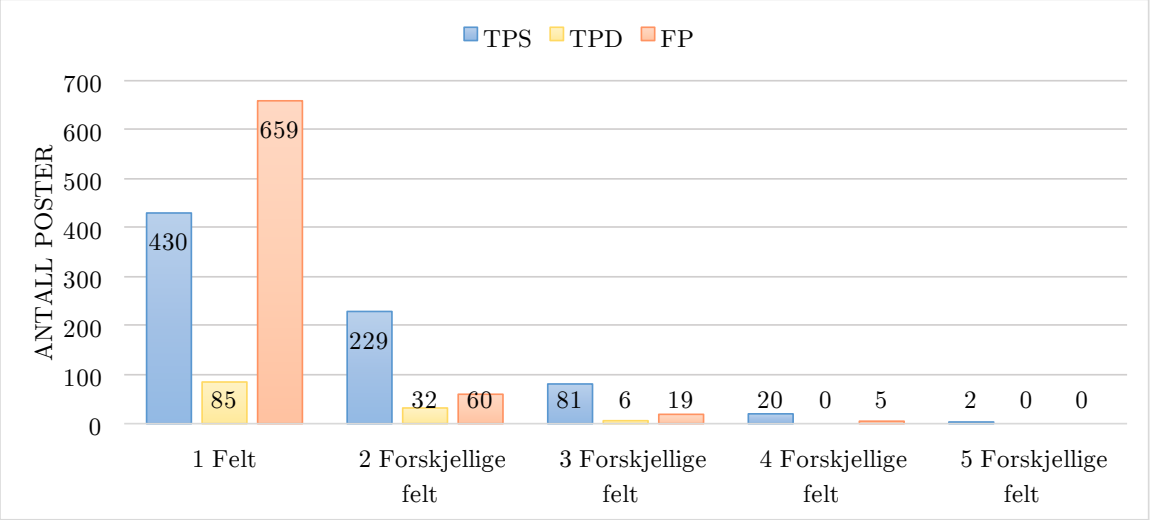
6.5.1 Fordeling

Figur 6.15 viser fordelingen av TPS, TPD og FP for postene som er hentet ut fra de forskjellige feltene som er brukt. Feltbruken er ujevnt brukt i datasamlingen, og hyppigheten av bruken av de forskjellige feltene er også ulik mellom katalogene. Felt 245 dominerer samlingen, mens andre felt som 610 og 710 nesten ikke er i bruk.



Figur 6.15: Fordelingen av TPS, TPD og FP for de forskjellige feltene som er brukt for å hente ut poster.

Figur 6.16 viser fordelingen av TPS, TPD og FP for verkskandidatene som kommer fra enten 1, 2, 3, 4 eller 5 ulike felt. Verkskandidatene som kommer fra bare ett felt representerer over 70 prosent av samlingen. Det at verkskandidatene som kommer fra 4 og 5 forskjellige felt er veldig få kan også ses her. I datasamlingene finnes det også en og annen verkskandidat som har opprinnelse i flere enn 5 forskjellige felt, men disse har blitt utelatt fordi de ikke gir verdi til analysen.



Figur 6.16: Fordelingen av TPS, TPD og FP og antallet forskjellige felt verkskandidaten har opprinnelse fra.

6.5.2 Precision

Figur 6.17 viser precisionen til verkskandidatene som kommer fra de ulike feltene. Her er det flere felt som skiller seg ut både positivt og negativt.

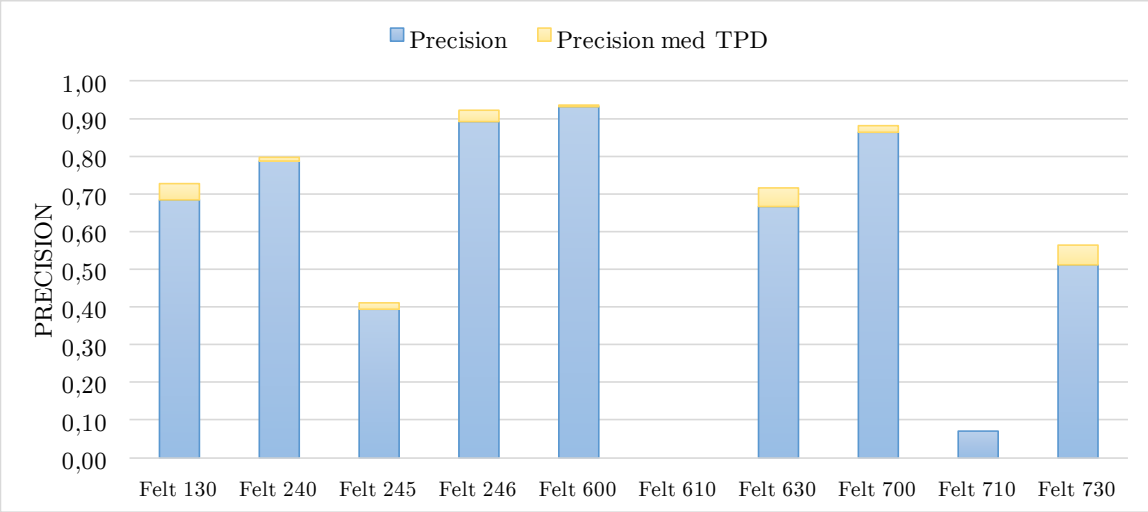
Feltene 610 og 710 skiller seg negativt ut. De er nærmest ikke-eksisterende, da de til sammen kun representerer 1 prosent av verkskandidatene. Videre har de også en veldig lav precision, der felt 610 har en precision på 0 og felt 710 har en precision på 0,07. Feltene 730 og 245 har heller ikke spesielt høy precision, da de ligger rundt den samlede totalen for hele samlingen. Hoveddelen av verkskandidatene kommer fra felt 245, men disse har en lav precision, og dermed lav sannsynlighet for å representere et verk. Felt 245 er stort sett brukt som tittelfelt, men en grunn til at det i denne analysen har så lav precision er at feltet ikke inneholder originaltittel, men også den oversatte tittelen.

Videre har postene som kommer fra feltene 130, 240 og 630 en relativt høy precision som ligger på rundt 0,75. Dette betyr at disse feltene med god sannsynlighet kan representere verk, og feltene er også relativt hyppig brukt. Til slutt kommer feltene 246, 600 og 700.

Felt 700 er det tredje mest brukte feltet i samlingen, og har også en veldig høy precision på 0,88. Feltet er representert i alle samlingene, som betyr at måten bibliotekene har brukt feltet vil være relativt lik mellom de forskjellige katalogene. Felt 700 inneholder mye forskjellig informasjon, og i denne sammenhengen er det underfeltet $t som er blitt brukt fordi dette feltet kan inneholde en tittel.

Feltet 246 eksisterer bare i den norske nasjonalbibliografien som beskrevet i 5.2. Der er felt 246 brukt til originaltittel, noe som naturlig vil gi et høyt antall riktig identifiserte verk, som igjen gir en høy precision. Feltet 246 har en precision på 0,92 som er veldig høyt. Problemet er at dette feltet kun er brukt i den norske nasjonalbibliografien og vil dermed sjelden kunne brukes av andre biblioteker til å identifisere verk.

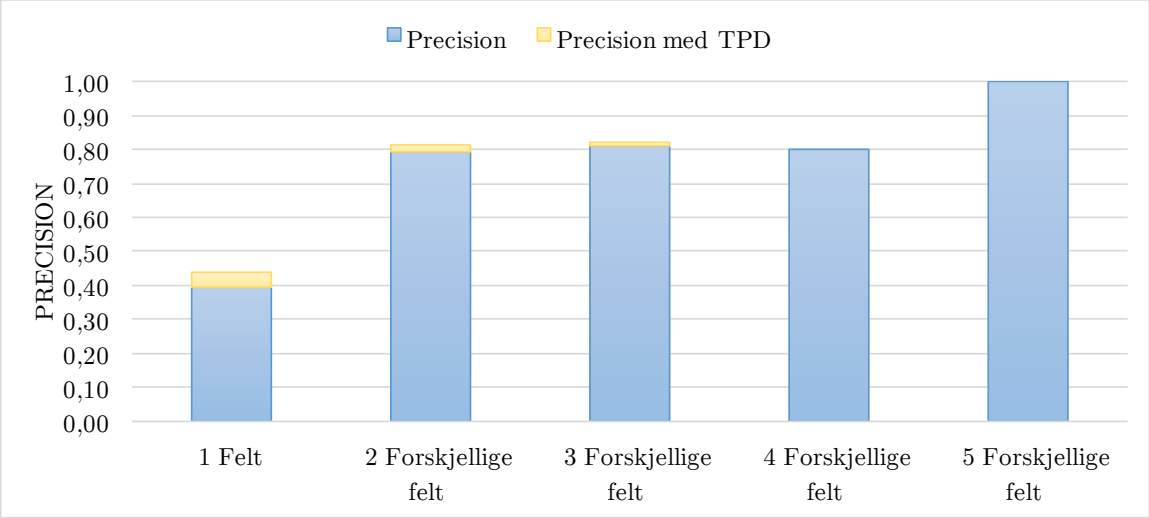
Felt 600 er også relativt godt representert i samlingen. Postene som kommer fra dette feltet har en precision på 0,93 som er den høyeste verdien i samlingen. I likhet med felt 700 er underfeltet $t til feltet 600 brukt til å hente ut verkskandidater. Siden felt 600 er godt representert i alle katalogene er dette en veldig god karakteristikk for identifisering av verk.



Figur 6.17: Precision for verkskandidatene basert på hvilke felt de har opprinnelse i.

Figur 6.18 viser precisionen til verkskandidatene som har opprinnelse i 1, 2, 3, 4 eller 5 forskjellige felt. Disse resultatene har visse likheter som resultatene i 6.4.2. Postene som bare oppstår i 1 felt har en veldig lav precision som er lavere enn totalen for hele samlingen. Denne gruppen består også av de fleste verkskandidatene. Verkskandidatene som kommer fra 2 eller flere forskjellige felt har alle relativt høy precision på rundt 0,8. Som nevnt i 6.5.1 er det svært få poster som har opprinnelse fra 4 og 5 forskjellige felt. Precisionen til verkskandidatene med opprinnelse i 5 forskjellige felt vil derfor være falsk høy på 1,0 siden det kun er snakk om to poster.

Deles verkskandidatene opp i to grupper der den ene gruppen inneholder verkskandidatene som bare har opprinnelse i 1 felt, og den andre gruppen inneholder verkskandidatene som har opprinnelse i 2 eller flere felt, kan dette være en mer representativ karakteristikk som er enklere å benytte seg av i praksis. For denne samlingen vil verkskandidatene som har opprinnelse i bare 1 felt bestå av 72 prosent av postene. Postene som har opprinnelse i 2 eller flere forskjellige felt vil bestå av 28 prosent. Denne siste gruppen vil da ha en samlet precision på 0,81.



Figur 6.18: Precision for postene basert på hvor mange forskjellige felt de har opprinnelse i.

6.5.3 Analyse

Feltbruk er den karakteristikken som har størst variasjon i precision. Noen felt har så lav precision at de enkelt kan utelates i verksidentifisering. Videre er det flere av feltene som har veldig god precision. Disse feltene har også god dekning og hyppighet i alle katalogene som er brukt. Dette er viktig for at karakteristikkene skal kunne brukes til å identifisere verk.

Feltene 240, 246, 600 og 700 har alle en precision som er høyere enn 0,8, som betyr at verkskandidater som har opprinnelse i et av disse feltene vil ha over 80 prosent sannsynlighet for å være et verk. Det er nesten en fordobling hvis det blir sammenlignet med den totale sannsynligheten i samlingen. Feltene 240, 600 og 700 har god dekning og hyppighet i katalogene, men felt 246 er nesten bare representert i den norske nasjonalbibliografien, og vil derfor ikke være nyttig for alle biblioteker. For den norske nasjonalbibliografien vil dette være en utmerket karakteristikk for verksidentifisering.

Antall forskjellige felt en verkskandidat har opprinnelse i gir også interessante funn. Verkskandidater som bare har opprinnelse i 1 felt har en veldig lav precision, og den er lavere enn totalen for hele samlingen. I likhet med karakteristikkene antall, og antall forskjellige kataloger en verkskandidat har opprinnelse i, er ikke dette en god karakteristikk for å identifisere verk. Det er likevel en god måte å luke vekk verkskandidater på, som med høy sannsynlighet ikke vil kunne representere et verk.

På verkskandidater som har opprinnelse i 2, 3, 4 og 5 forskjellige felt er ikke karakteristikkene like nyttige i seg selv. Dette er fordi det er få verkskandidater som går inn i disse kategoriene, som gjør at de ikke vil kunne identifisere verk med god sikkerhet. Det som er interessant å se på er forskjellen mellom verkskandidater som har opprinnelse i 1 felt og de verkskandidatene som har opprinnelse i flere enn 1 felt. Da vil verkskandidatene som har opprinnelse i flere enn 1 felt bestå av nesten 30 prosent av verkskandidatene. I tillegg vil denne gruppen ha en samlet precision på 0,81 som er en god indikasjon på om en verkskandidat kan representere et verk. På den måten er det mulig å konkludere med at alle verkskandidater som har opprinnelse i flere enn 1 felt har stor sannsynlighet for å kunne representere et verk.

6.6 Kombinasjon av karakteristikker

Siden verkskandidatene har flere forskjellige karakteristikker, har det også blitt sett på hvilke resultater som kommer hvis forskjellige karakteristikker kombineres. Kombinasjoner av karakteristikker kan føre til flere funn som vil gjøre det lettere å sortere ut hvilke verkskandidater som kan representere verk. I dette prosjektet har karakteristikken som sier hvilken katalog verkskandidaten har opprinnelse fra ikke blitt tatt med, fordi dette ikke er en analyse av katalogene og de forskjellige karakteristikkene ved katalogene. For eksempel har verkskandidatene som har opprinnelse i Harvard og Felt 130 ikke blitt sett på, fordi denne kombinasjonen bare sier hvor riktig postene fra felt 130 i Harvard-katalogen er. Dette kan være interessant for Harvard, men det er ikke interessant for dette prosjektet.

6.6.1 Metodikk

I dette prosjektet har det blitt valgt å se på kombinasjonen av to og to karakteristikker. Bakgrunnen er at jo flere karakteristikker som blir kombinert, jo færre verk vil inneholde kombinasjonen av alle karakteristikkene. Hvis to og to karakteristikker blir kombinert vil det fortsatt være en del verkskandidater som innehar begge karakteristikkene slik at resultatene kan være representative. Til sammen i denne oppgaven er det brukt 20 forskjellige karakteristikker, dersom man ser bort ifra karakteristikkene som beskriver hvilken katalog posten har opprinnelse i. I dette prosjektet blir kun de 37 mest interessante kombinasjonene presentert. Kombinasjonene som presenteres i denne oppgaven måtte oppfylle tre krav:

1. Den samlede precisionen måtte være over 0,8.
2. Den samlede precisionen måtte være høyere enn precisionen til hver av de brukte karakteristikkene.
3. Det måtte eksistere over 10 verkskandidater som innehadde begge karakteristikkene.

Precisionen til verkskandidatene som innehadde begge karakteristikkene ble regnet ut ved først å finne antall TPS, TPD og FP av verkskandidatene. Deretter ble precisionen utregnet på vanlig måte og verkskandidatene som ble markert som TPD ble medregnet.

Resultatene er presentert med tabeller som viser de ulike karakteristikkene som har blitt kombinert, totalt antall verkskandidater som innehar begge karakteristikkene, precisionen til verkskandidatene som innehar begge karakteristikkene, den høyeste av precisionene som er med i kombinasjonene og forbedringen av precision. Den karakteristikken som hadde den høyeste precisionen av de to som ble kombinert er markert med fet skrift. Forbedringen av precision er regnet ut ved å ta differansen mellom precisionen til verkskandidatene som innehar begge karakteristikkene og precisionen til den av de to karakteristikkene som hadde høyest precision.

I Tabell 6.1 vises kombinasjonen av forskjellige karakteristikker. Første linje viser resultatene fra verkskandidatene som innehadde to karakteristikker. Karakteristikk 1 er "4 Forskjellige kataloger" og Karakteristikk 2 er "Med forfatter". Av disse to karakteristikkene har Karakteristikk 1 høyest precision med 0,71, og dette er vist ved at karakteristikken er skrevet med fet skrift. Totalt var det 16 verkskandidater som både hadde opprinnelse fra 4 forskjellige kataloger, og som stod oppført med forfatter. For disse 16 verkskandidatene var precisionen 0,94. Det betyr at forbedringen på var 0,22 til forskjell fra precisionen til verkskandidatene som bare hadde opprinnelse fra 4 forskjellige kataloger. Tabellene er sortert etter størst forbedring av precision.

6.6.2 Kataloger

Den første karakteristikken som har blitt sett på i kombinasjon med andre er antall kataloger verkskandidaten har opprinnelse i. Totalt var det 15 karakteristikker der den ene karakteristikken var de verkskandidatene som hadde opprinnelse i enten 2, 3 eller 4 forskjellige kataloger, kombinert med andre karakteristikker som oppfylte kravene i 6.6.1. De fleste kombinasjonene hadde liten forbedring i precision, men noen av kombinasjonene hadde veldig stor forbedring. Stort sett består resultatene som er presentert av kombinasjonen mellom to karakteristikker som allerede har en høy precision. Siden verkskandidatene må inneha to karakteristikker vil det naturlig gjelde færre verkskandidater, men for flere av kombinasjonene er det snakk om en god mengde verkskandidater. De karakteristikkene som har hyppigst forekomst er "2 Forskjellige kataloger", og "3 Forskjellige kataloger". Dette er naturlig siden begge karakteristikkene i seg selv har høy precision, og det er en del verkskandidater som faller under disse karakteristikkene.

Karakteristikk 1	**Karakteristikk 2**	**Antall verkskandidater**	**Precision**	**Høyeste precision av karakteristikk 1 og karakteristikk 2**	**Forbedring av precision**
4 Forskjellige kataloger	Med forfatter	16	0,94	0,71	0,22
3 Forskjellige kataloger	**Felt 700**	23	1,00	0,88	0,12
4 Forskjellige kataloger	**Felt 240**	18	0,89	0,80	0,09
2 Forskjellige kataloger	Med forfatter	113	0,88	0,80	0,09
3 Forskjellige kataloger	**Felt 246**	27	1,00	0,92	0,08
3 Forskjellige kataloger	**Felt 600**	23	1,00	0,93	0,07
2 Forskjellige kataloger	2 Forskjellige felt	154	0,86	0,81	0,05
2 Forskjellige kataloger	**Felt 246**	65	0,97	0,92	0,05
2 Forskjellige kataloger	**Felt 700**	109	0,93	0,88	0,05
3 Forskjellige kataloger	Felt 240	39	0,87	0,83	0,04
3 Forskjellige kataloger	Felt 130	15	0,87	0,83	0,03
3 Forskjellige kataloger	3 Forskjellige felt	30	0,87	0,83	0,03
2 Forskjellige kataloger	Felt 240	109	0,83	0,80	0,03
3 Forskjellige kataloger	Med forfatter	51	0,84	0,83	0,01
2 Forskjellige kataloger	**Felt 600**	82	0,94	0,93	0,01

Tabell 6.1: Oversikt over verkskandidatene som har opprinnelse i enten 2, 3 eller 4 forskjellige kataloger, som innehar en annen karakteristikk og som oppfylte kravene i 6.6.1

Analyse

Som tallene i tabellen viser vil det i flere sammenhenger være nyttig å se på verkskandidater som innehar en kombinasjon av karakteristikker. Kombinasjonene som har mest forbedring i precision gjelder et lite antall verkskandidater. For eksempel gjelder de tre første kombinasjonene i Tabell 6.1 kun for 16, 23 og 18 verkskandidater. Verkskandidatene som har opprinnelse i "2 Forskjellige

katalogerkarakteristikken og også står oppført med forfatter har både en stor forbedring i precision, og totalt var det 113 verkskandidater som innehadde begge karakteristikkene. Hvis denne kombinasjonen i tillegg sammenlignes med verkskandidatene som står oppført med forfatter, er det en økning fra 0,49 til 0,88.

Karakteristikkene som er kombinert med antall forskjellige kataloger verkskandidaten har opprinnelse i, som er vist i Karakteristikk 2-kolonnen, er helt uavhengige av hverandre. Flere av disse kombinasjonene kan være effektive indikatorer på om verkskandidater kan representere et verk. Flere av kombinasjonene har en precision på opp mot, og over 0,9, som vil si at med 90 prosent sannsynlighet vil en verkskandidat som innehar en kombinasjon av en av to av karakteristikkene i radene i Tabell 6.1 kunne representere et verk.

6.6.3 Feltbruk

De neste karakteristikken som har blitt sett på i kombinasjon med andre karakteristikker er antall ulike felt en verkskandidat har opprinnelse i. Kombinasjonen antall forskjellige kataloger verkskandidaten har opprinnelse i, vist i Karakteristikk 1-kolonnen, og andre karakteristikker, vist i karakteristikk 2-kolonnen, er vist i 6.2. Karakteristikkene som er kombinert med antallet ulike felt verkskandidatene har opprinnelse i er vist i Karakteristikk 2-kolonnen. Igjen er det stor variasjon i antall verkskandidater som innehar begge karakteristikkene. Likevel er det flere av kombinasjonene der mange verkskandidater innehar begge karakteristikkene som gjør at tallene kan være representative.

Flere av kombinasjonene er mellom karakteristikker som ikke er helt uavhengige. For eksempel er det noen felt som i stor grad opptrer i kombinasjon med andre felt, som felt 600 og 700. Derfor er det ikke overraskende at noen av disse kombinasjonene har et stort antall verkskandidater som tilfredsstiller begge karakteristikkene.

Karakteristikk 1	Karakteristikk 2	Antall verkskandidater	Precision	Høyeste precision av karakteristikk 1 og karakteristikk 2	Forbedring av precision
3 Forskjellige felt	Felt 240	71	0,96	0,82	0,14
3 Forskjellige felt	**Felt 700**	15	1,00	0,88	0,12
3 Forskjellige felt	Med forfatter	46	0,91	0,82	0,09
3 Forskjellige felt	**Felt 600**	15	1,00	0,93	0,07
4 Forskjellige felt	Felt 130	15	0,87	0,80	0,07
2 Forskjellige felt	**Felt 700**	125	0,94	0,88	0,06
2 Forskjellige felt	**Felt 246**	54	0,98	0,92	0,06
2 Forskjellige felt	Med forfatter	110	0,85	0,81	0,04
2 Forskjellige felt	Felt 240	73	0,84	0,81	0,02
3 Forskjellige felt	**Felt 246**	57	0,93	0,92	0,01
2 Forskjellige felt	**Felt 600**	115	0,94	0,93	0,01

Tabell 6.2: Oversikt over verkskandidatene som har opprinnelse i enten 2, 3 eller 4 forskjellige felt, som innehar en annen karakteristikk og som oppfylte kravene i 6.6.1

Analyse

Flere av kombinasjonene mellom karakteristikkene som er presentert i Tabell 6.2 viser gode resultater i form av høy precision. For noen av kombinasjonene er det et lite antall verkskandidater som innehar begge karakteristikkene, men i hovedsak er det flere kombinasjoner der en stor andel verkskandidater innehar begge karakteristikkene. I hovedsak har flere av karakteristikkene som har blitt kombinert allerede en høy precision, men flere av kombinasjonene er med på å øke precisionen. Blant annet har verkskandidatene som står oppført med forfatter i utgangspunktet en lav precision, mens verkskandidatene som står oppført med forfatter og i tillegg har opprinnelse i enten to eller tre forskjellige felt, en veldig høy precision. Det er 71 verkskandidater som har opprinnelse i tre forskjellige felt der ett av disse feltene er felt 240. Disse verkskandidatene har i tillegg en precision på 0,96. Videre er også verkskandidatene som har opprinnelse i 2 forskjellige felt, der et av dem er felt 700 eller felt 600, gode indikasjoner på om de kan representere et verk. Det samme gjelder verkskandidatene som har opprinnelse i 2 forskjellige felt og som står oppført med forfatter. For disse kombinasjonene er det over 100 verkskandidater som tilfredsstiller begge karakteristikkene og de har en høy precision på 0,94, 0,85 og 0,94.

6.6.4 felt

Den siste karakteristikken som har blitt sett på i kombinasjon med andre karakteristikker er hvilke felt verkskandidatene kommer fra. Der denne karakteristikken er kombinert med enten antall forskjellige kataloger verkskandidaten har opprinnelse i eller antall forskjellige felt verkskandidaten har opprinnelse i, er beskrevet i 6.6.2 og 6.6.3. I hovedsak er det snakk om kombinering av verkskandidater som kommer fra minst to forskjellige felt. To av kombinasjonene er verk som kommer fra felt

700 og felt 246, som i tillegg står oppført med forfatter. Igjen gjelder flere av kombinasjonene et lite antall verkskandidater, men noen av kombinasjonene gjelder et større antall verkskandidater som er mer representativt. For disse kombinasjonene er forbedringen av precision litt lavere enn i 6.6.2 og 6.6.3, men kombinasjonene gir fortsatt nyttige resultater. På grunn av reglene som er brukt for å hente ut FRBRposter fra postene, vil for eksempel felt 245 bli brukt hvis felt 240 eksisterer i samme posten. I denne sammenslåingen vil de verkskandidatene som både har opprinnelse fra felt 240 og felt 245, gjelde verkskandidatene som har opprinnelse fra felt 240 fra en katalog, og felt 245 fra en annen katalog. Verkskandidaten kan også være slått sammen av to ulike poster fra samme katalog.

Karakteristikk 1	**Karakteristikk 2**	**Antall verkskandidater**	**Precision**	**Høyeste precision av karakteristikk 1 og karakteristikk 2**	**Forbedring av precision**
Felt 240	**Felt 700**	25	0,96	0,88	0,08
Felt 246	Felt 700	12	1,00	0,92	0,08
Felt 240	Felt 245	132	0,87	0,80	0,07
Felt 240	**Felt 246**	67	0,97	0,92	0,05
Felt 700	Med forfatter	89	0,92	0,88	0,04
Felt 245	**Felt 700**	11	0,91	0,88	0,03
Felt 240	**Felt 600**	25	0,96	0,93	0,03
Felt 246	Med forfatter	77	0,95	0,92	0,03
Felt 245	**Felt 246**	95	0,94	0,92	0,02
Felt 600	Felt 700	138	0,94	0,93	0,01

Tabell 6.3: Oversikt over verkskandidatene som har opprinnelse i enten felt 240, 245, 246, 600 eller 700, som innehar en annen karakteristikk og som oppfylte kravene i 6.6.1

Analyse

Kombinasjonene som er presentert i Tabell 6.3 gir interessante funn, der alle kombinasjonene, med unntak av én, har en precision på over 0,9. karakteristikkene som er brukt som for eksempel felt 246, 600 og 700 har allerede høy precision i seg selv, og disse feltene i kombinasjon med andre karakteristikker vil derfor ikke gi veldig store utslag i forbedringen av precision. Til tross for dette vil for eksempel verkskandidater som både kommer fra felt 240 og 245 ha en ytteligere stor sannsynlighet for å kunne representere et verk. Andre kombinasjoner som verkskandidatene som kommer fra felt 600 og 700, verkskandidatene som kommer fra felt 245 og 246 og verkskandidatene som kommer fra felt 700, samt står oppført med forfatter, er alle gode indikatorer på om verkskandidater kan representere et verk. Dette er på grunn av den høye precisionen og det høye antallet verkskandidater som innehar begge karakteristikkene. Flere av kombinasjonene er av felt som til vanlig opptrer sammen, og det er derfor ikke overraskende at det er hyppige forekomster av verkskandidater som kommer fra begge disse feltene. Alle karakteristikkene som er med i disse kombinasjonene, med unntak av verkene som er oppført med forfatter, er gode indikasjoner på om verkskandidaten kan representere et verk. Tallene viser at karakteristikkene i kombinasjon med hverandre øker sannsynligheten ytterligere for

å kunne representere et verk. For eksempel vil verkskandidatene som kommer fra både felt 240 og 245, i denne samlingen, ha 97 prosent sannsynlighet for å representere et verk.

6.6.5 Oppsummering

Analysen viser at det i flere sammenhenger vil være nyttig å se på kombinasjoner mellom to og to karakteristikker for å kunne identifisere verk. Kravene som ble stilt til verkskandidatene som ble presentert i 6.6.1, gjør at det kun er noen av resultatene som ble presentert her. En av utfordringene med karakteristikkene som er kombinert er at de allerede har høy precision. På mange måter vil det være mer nyttig å finne kombinasjoner mellom karakteristikker som allerede har lav precision, og på den måten finne flere verkskandidater som kan representere verk som ikke allerede har blitt funnet. Man kan derfor argumentere for at kravene som ble stilt i 6.6.1 er for strenge og utelater viktige funn. Men Tabell 6.4 viser de 25 verkskombinasjonene som har høyest økning i precision. Kombinasjonene i denne tabellen er enten presentert i analysen, eller så er det så få verkskanididater som innehar begge karakteristikkene at det ikke gir mening i å ta dem med.

Til tross for at de fleste av karakteristikkene som er brukt har god precision, vil de i kombinasjon med hverandre enda tydeligere identifisere verkskandidater som kan representere verk. Flere av kombinasjonene har et høyt antall verkskandidater som inneholder begge karakteristikkene. De vil derfor være et representativt funn som kan brukes til verksidentifisering. Det vil være mulig å analysere kombinasjoner av ytterligere karakteristikker, men da vil resultatene i større grad lide av få verkskandidater som tilfredsstiller alle karakteristikkene.

Karakteristikk 1	Karakteristikk 2	Antall	Precision	Høyeste precision av karakteristikk 1 og karakteristikk 2	Forbedring i precision
Felt 630	4 Forskjellige kataloger	2	1,00	0,72	0,28
Med forfatter	**4 Forskjellige kataloger**	16	0,94	0,71	0,22
4 Forskjellige felt	**4 Forskjellige kataloger**	2	1,00	0,80	0,20
Med forfatter	**4 Forskjellige felt**	3	1,00	0,80	0,20
4 Forskjellige felt	**3 Forskjellige kataloger**	7	1,00	0,83	0,17
Felt 240	**3 Forskjellige felt**	71	0,95	0,82	0,13
Felt 700	3 Forskjellige felt	15	1,00	0,88	0,12
Felt 700	4 Forskjellige kataloger	2	1,00	0,88	0,12
Felt 240	4 Forskjellige kataloger	18	0,89	0,80	0,09
Med forfatter	**3 Forskjellige felt**	46	0,90	0,82	0,08
Felt 246	Felt 630	1	1,00	0,92	0,08
Felt 246	3 Forskjellige kataloger	27	1,00	0,92	0,08
Felt 246	4 Forskjellige felt	2	1,00	0,92	0,08
Felt 246	Felt 700	12	1,00	0,92	0,08
Felt 240	**Felt 700**	25	0,96	0,88	0,08
Med forfatter	**2 Forskjellige kataloger**	113	0,87	0,80	0,07
Felt 130	4 Forskjellige kataloger	5	0,80	0,73	0,07
Felt 600	3 Forskjellige felt	15	1,00	0,93	0,07
Felt 600	4 Forskjellige kataloger	2	1,00	0,93	0,07
Felt 246	**Felt 600**	3	1,00	0,93	0,07
Felt 130	**4 Forskjellige felt**	15	0,87	0,80	0,07
Felt 240	Felt 245	132	0,86	0,80	0,06
Felt 700	2 Forskjellige felt	125	0,94	0,88	0,06
Felt 246	2 Forskjellige felt	54	0,97	0,92	0,05
Felt 130	Felt 630	58	0,77	0,73	0,05

Tabell 6.4: Kombinasjonene av karakteristikker som har størst forbedring i precision

Kapittel 7

Konklusjon

Datakvalitet er et viktig konsept innenfor alle typer dataprosessering, også innenfor linked open data. I oppgaven har den varierende kvaliteten i data ved eksisterende systemer blitt vurdert. Flere og flere biblioteker publiserer de bibliografiske katalogene sine på nettet, men tiltakene som er gjort for å sikre kvaliteten på det som blir publisert er ofte ikke bra nok. Publisering av data som linked open data er et skritt i riktig retning for å gjøre dataene mer tilgjengelig, spesielt for brukerne som plutselig får tilgang til tidligere utilgjengelige data. Problemet er at dataene som blir publisert er fulle av støy, de er lite konsistente og ofte ufullstendige. Innenfor bibliografiske data er FRBR-modellen i hovedsak blitt brukt for dataene som blir publisert som linked open data. Den viktigste prosessen innenfor FRBRisering er identifisering av entiteter, og i første omgang identifisering av verk.

I denne oppgaven har det blitt undersøkt hvordan kvaliteten til bibliografiske data kan bedres, gjennom å finne en metode for å bedre kunne identifisere verk blant bibliografiske poster.

Forskningsspørsmålene for prosjektet er:

Q$_1$: Hvordan er tilstanden til kvaliteten på dataene i samlinger som er publisert som linked open data?

- **q**$_1$: Hvilke kvalitetsproblemer er det i dataene som er publisert som linked open data?

Q$_2$: Finnes det data fra verksidentifisering som kan brukes for å angi påliteligheten til genererte verk?

- **q**$_1$: Hvilke data fra verksidentifisering kan brukes for å angi påliteligheten til genererte verk?
- **q**$_2$: Hvordan kan verk identifiseres uten å skape feilidentifiserte verk?

Oppsummeringen av svarene på disse spørsmålene kan leses i 7.2, i dette kapittelet.

7.1 Oppsummering

I denne oppgaven har forskjellige måter for hvordan verk kan identifiseres i bibliografiske data blitt analysert. Først ble det gjort en studie på begrepet datakvalitet innenfor linked open data og andre lignende prosesser. Hvordan datakvalitet kan vurderes og evalueres i forskjellige datasystemer har blitt sett på, i hovedsak i forbindelse med publisering av biblioteksdata som linked open data.

FRBR-modellen og FRBRisering har blitt beskrevet, og fordi FRBR-modellen er mye brukt for data som publiseres som linked open data, kan mye av forbedringen til kvaliteten av dataene skje i FRBRiseringsprosessen.

Videre ble det gjennomført en analyse av utvalgte kjente systemer som har publisert bibliografiske data som linked open data, og som har identifisert verk. Bibliotekene og systemene som ble analysert var FictionFinder, den franske nasjonalbibliografien, den tyske nasjonalbibliografien, Biblioteca Virtual Cervantes, og BIBSYS semantisk web. Analysen gikk ut på å vurdere dataene som blir presentert for brukeren gjennom de forskjellige webgrensesnittene til bibliotekene og systemene. Denne analysen viser flere områder der systemene kan forbedre kvaliteten til de publiserte dataene.

Til slutt ble det sett på et utvalg bibliografiske poster hentet fra fire forskjellige kataloger. Dataene som har blitt sett på ble FRBRisert fra MARC21-formatet til FRBR som RDF. Deretter ble postene slått sammen med hverandre basert på en generert identifikatornøkkel. Karakteristikkene til datene ble deretter analysert. Analysen gikk ut på å undersøke hvilke karakteristikker som bestemmer om en verkskandidat kan representere et verk eller ikke.

Karakteristikkene som ble vurdert var antallet poster som var slått sammen til hver verkskandidat, hvilke MARC-felt verkskandidaten har opprinnelse i, og antallet forskjellige MARC-felt. Det ble også undersøkt hvilken katalog verkskandidaten har opprinnelse i, antallet forskjellige kataloger den har opprinnelse i, og om en verkskandidat står oppført med forfatter eller ikke. Til slutt ble forskjellige kombinasjoner av karakteristikker analysert.

7.2 Bidrag

I analysen av eksisterende systemer som har publisert bibliografiske data som linked open data, kom det fram tydelige tendenser i kvaliteten ved dataene som er blitt publisert. Gjennomgående bestod dataene av mye støy og generelt et lavt antall identifiserte verk. Verkene som var identifisert bestod, i flere av systemene, av feilidentifiserte poster og duplikater av allerede identifiserte verk. Denne analysen viste hvordan dagens situasjon lider av dårlig kvalitetssikring og evaluering av dataene før de blir publisert. Flere grep kan gjøres for å bedre dagens situasjon, og spesielt innenfor identifisering av verk er det mye som kan forbedres.

I analysen av karakteristikkene til verkskandidatene som ble sett på i denne oppgaven er det flere interessante funn som vil kunne være nyttige i videre arbeid med identifisering av verk i bibliografiske data. Flere av verkskandidatene som kunne representere et verk viste tydelige likheter i karakteristikker med hverandre. På denne måten kan disse karakteristikkene brukes videre til å identifisere verk.

Karakteristikker som hvilke MARC-felt en verkskandidat kommer fra, antallet forskjellige MARC-felt en verkskanididat kommer fra, antallet poster en verkskandidat har blitt slått sammen av, og antallet forskjellige kataloger en verkskandidat har opprinnelse i, er alle gode metrikker for å identifisere verk. Poster som har noen av disse karakteristikkene vil ha over 80 prosent sannsynlighet for å representere et verk og noen vil ha over 90 prosent sannsynlighet. Tallene viser at flere av karakteristikkene vil kunne være effektive metrikker for å si med høy sikkerhet hvilke poster som representerer et verk.

7.3 Evaluering

Oppgaven viser hvordan data som blir og har blitt publisert som linked open data trenger en opprydding. De spesifikke utfordringene er også presentert, og ved å sette fokus på disse utfordringene vil det være enklere å iverksette forbedringstiltak. Mer og mer data blir publisert som linked open data, men dataene som blir publisert er ikke mer nyttige enn kvaliteten de har. Denne oppgaven vil kunne brukes til enklere å identifisere verk med høyere riktighets-rate, som er det første steget mot å bedre kvaliteten til bibliografiske data som er publisert. Resultatene som er vist i denne oppgaven gir ikke en metode for å publisere riktig identifiserte verk, men bruk av resultatene vil kunne gjøre det mulig å publisere verk som man med en viss sikkerhet vet er et riktig identifisert verk.

Dette prosjektet har vist tydelige resultater som vil være nyttige i videre arbeid med identifisering av verk og andre entiteter i FRBR-modellen. Metoden som er presentert vil ha en verdi i videre arbeid rundt generering av riktig identifiserte poster. I løpet av prosjektet har det blitt tatt valg som har ført til fruktbare resultater, men noen valg kunne også ha blitt gjort annerledes for å få en enda dypere forståelse av problemet og resultatene.

Resultatene avhenger av mange faktorer, og én av disse faktorene er markeringen av de utvalgte postene som TPS, TPD og FP. Denne manuelle markeringen var en tidkrevende prosess, som avhenger av personen som gjennomfører den. I dette prosjektet ble verkskandidatene markert som TPS, TPD og FP i hovedsak basert på tittel og eventuelle undertitler. Inndelingen av TPS, TPD og FP er en streng inndeling der kun verkskandidater med riktig originaltittel regnes som et riktig identifisert verk. Man kan argumentere for at også verkskandidater som tydelig baserer seg på verk, blant annet oversettelser, kan være riktig identifisert. Verkskandidatene som da blir markert som FP vil være poster som helt tydelig ikke er et verk. Dette kan for eksempel være en samling av forskjellige historier, eller en *samlede-verker* bok, som inneholder verk, men som ikke er et verk i seg selv. Å markere oversettelser som FP gjør at det er mulig å lage en verksoversikt med høy pålitelighet for at verkene som er identifisert representerer et verk med originaltittel. Problemet med for eksempel FictionFinder var at flere av de identifiserte verkene var identifisert med engelsk tittel, men den oversatte versjonen av verket burde heller identifiseres som et uttrykk som realiserer verket.

Dataene som ble brukt kunne også med fordel ha blitt gjennomgått mer nøyaktig for å hindre opprettelsen av duplikater av verkskandidater. Hvis originaldataene hadde vært mer nøyaktig gjennomgått på forhånd ville kvaliteten på dataene som ble analysert vært bedre og resultatene ville blitt renere. Et eksempel på dette er at det oppstod duplikater av forfattere som gjorde at identiske verkskandidater ikke ble slått sammen fordi de var relatert til forskjellige instanser av samme forfatter. På den måten ville resultatene vært mindre preget av kvaliteten på dataene som ble analysert etter FRBRiseringen, og mer preget av dataene i seg selv.

7.4 Videre arbeid

Identifisering av verk i bibliografiske poster i FRBR-modellen er det første steget mot forbedring av kvaliteten på data som blant annet skal publiseres på nett som linked open data. Innenfor FRBR er det flere entiteter og relasjoner som må identifiseres. I første omgang vil det være interessant å se på identifiseringen av personer og relasjonene mellom personer og verk. Personer, manifestasjoner, uttrykk og så videre har alle forskjellige karakteristikker som vil kunne analyseres for enklere å identifisere dem. Hvis verk er identifisert vil det være enklere å for eksempel identifisere tilhørende uttrykk og manifestasjoner, samt relasjonen mellom dem.

Videre vil det også være interessant å teste resultatene fra denne oppgaven på et annet datasett for så å se på kvaliteten på verkene som blir identifisert. Det er forskjeller mellom hvordan kataloger bruker de forskjellige MARC-feltene, og det vil derfor være nyttig å se om resultatene i denne oppgaven kan være representative for flere samlinger enn bare samlingene som er brukt her.

Til slutt er det én forbedring innenfor dataprosesseringen som vil kunne øke kvaliteten på dataene betydelig. Sammenslåingsprosessen som er brukt i dette prosjektet er en enkel nøkkel-matching-prosess som ikke vil slå sammen like poster på grunn av små forskjeller i skrivemåte. En matching-prosess med en smartere strengmatching vil kunne gi betydelig bedre resultater i form av færre duplikater. For kataloger med mange millioner poster er dette en vanskelig og tidkrevende prosess, men forbedringen vil kunne være stor.

Bibliografi

[1] IFLA, *Funksjonskrav til bibliografiske poster*. Oslo: IFLA Universal Bibliographic Control and International MARC Programme, 2001.

[2] G. Candela, P. Escobar, R. Carrasco, and M. Marco-Such, "Migration of a library catalogue into RDA linked open data," *Semantic Web*, vol. Preprint, no. Preprint, pp. 1–11, 2017.

[3] O. Research, "OCLC research actifvities and IFLA's functional requirements for bibliographic records," 2002. Accessed: 28.03.2017.

[4] T. Aalberg and M. Žumer, "The value of MARC data, or, challenges of frbrisation," *Journal of Documentation*, vol. 69, no. 6, pp. 851–872, 2013.

[5] G. Weikum and M. Theobald, "From information to knowledge: harvesting entities and relationships from web sources," *PODS '10 Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 65–76, 2010.

[6] S. Auer *et al.*, "Dbpedia: A nucleus for a web of open data," *Proceedings of the 6th International The Semantic Web.*, pp. 722–735, 2007.

[7] O. Etzioni, M. Cafarella, D. Downey, and S. Kok, "Web-scale information extraction in knowitall:(preliminary results)," in *Proceedings of the 13th International Conference on World Wide Web*, pp. 100–110, 2004.

[8] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, pp. 39–41, Nov. 1995.

[9] J. Decourselle, F. Duchateau, and N. Lumineau, "A survey of frbrization techniques," in *TPDL*, 2015.

[10] A. Simon, D. A. Mascio, and V. Michel, "We grew up together: data.bnf.fr from the BnF and logilab perspectives," in *IFLA 2014 Satellite Meeting*, 2014.

[11] D. Vizine-Goetz, "FictionFinder: A FRBR-based prototype for fiction in WorldCat," 2009. Accessed: 5.5.2017.

[12] D. Vila-Suero, B. Villazón-Terrazas, and A. Gómez-Pérez, "datos.bne.es: A library linked dataset," *Semantic Web*, vol. 4, no. 3, 2013.

[13] V. Daniel and G. Asunción, "datos.bne.es and MARiMbA: an insight into library linked data," *Libr Hi Tech*, vol. 31, no. 4, pp. 575–601, 2013.

[14] A. Hogan *et al.*, "An empirical survey of linked data conformance," *Web Semant.*, vol. 14, pp. 14–44, 2012.

[15] ISO, "ISO9000: International standards for quality management," ISO 9000, International Organization for Standardization, Geneva, Switzerland, 2015.

[16] S. Schlobach and C. A. Knoblock, "Dealing with the messiness of the web of data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 14, no. 1, p. 2, 2014.

[17] B. Behkamal, M. Kahani, and E. Bagheri, "Quality metrics for linked open data," *In International Conference on Database ...*, pp. 144–152, 2015.

[18] B. Saha and D. Srivastava, "Data quality: The other face of big data," *Data Engineering (ICDE), 2014 IEEE 30th International Conference*, pp. 1294–1297, 2014.

[19] T. Nguyen, "The value of ETL and data quality," *Data Warehousing and Enterprise Solutions*, 2003.

[20] T. Baker *et al.*, "Library linked data incubator group final report," *W3C*, 2011.

[21] N. Takhirov, F. Duchateau, and T. Aalberg, *Linking FRBR Entities to LOD through Semantic Matching*, vol. 6966. springerlink, 2011.

[22] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for linked data: A survey," *Semantic Web, vol. 7*, 2016.

[23] A. Assaf and A. Senart, "Data quality principles in the semantic web," *IEEE Xplore Digital Library*, pp. 226–229, 2012.

[24] D. N. Bibliothek, "The linked data service of the german national library- modelling of bibliographic data," *The German National Library*, pp. 1–25, 2016.

[25] C. Mönch and T. Aalberg, "Automatic conversion from MARC to FRBR," *Research and Advanced Technology for Digital Libraries*, vol. 2769, pp. 405–411, 2003.