HØGSKOLEN I OSLO
OG AKERSHUS

Miriam Bakkeli

«God kveld og velkommen til Dagsrevyen»:

En studie av innholdsbeskrivelser for nyhetsinnslag i NRKs arkiv fra oktober 2012

Masteroppgave 2014

Master i bibliotek- og informasjonsvitenskap

Høgskolen i Oslo og Akershus, Institutt for arkiv- bibliotek- og informasjonsfag

Sammendrag

Denne studien undersøker indeksering av nyhetsinnslag fra NRK sitt produksjons- og arkivsystem Programbanken. Indekseringspraksisen er i dag desentralisert og metadata blir utarbeidet av produksjonsmedarbeidere. Datamaterialet er hentet fra Nyhetsredaksjonen og omfatter Dagsrevyen, Lørdagsrevyen og Søndagsrevyen fra oktober 2012. Det er utført undersøkelser av 1941 tagger og annen innholdsbeskrivelse. Formålet var å se etter spesielle egenskaper blant taggene, som kan fortelle om den nye praksisen. Taggene dekker i hovedsak emnet, men består også av andre innfallsvinkler. Taggene fordeler seg over flere ordklasser enn det man ser i et tradisjonelt kontrollert vokabular. Det var 1095 av 1941 tagger som bare var benyttet en gang. Taggene er sammenliknet med annen innholdsbeskrivelse fra metadatafeltene Tittel og Rubrikk. Det viste seg at det var 54 % ordlikhet, helt eller delvis, mellom Tagger og Tittel og/eller Rubrikk. Det ble undersøkt om det eksisterer relasjoner mellom metadatafeltene. Det ble oppdaget flest assosiative relasjoner. Semantiske fenomener som homonymi, synonymi, kvasisynomymi, forkortelser og skrivefeil ble kartlagt. Det finnes litt av alle disse fenomenene. Det var færre skrivefeil og mindre bruk av synonymer enn forventet. Homonymi er utbredt både hos taggekorpuset og hos taggene med kontekst. Dette kan skyldes at taggene er av en fasettert karakter og ofte av generelle betydning.

Høgskolen i Oslo og Akershus, Institutt for arkiv- bibliotek- og informasjonsfag

Oslo 2014

Innholdsfortegnelse

Forord

Nå setter jeg punktum for en fantastisk studenttilværelse på Høgskolen i Oslo (og etterhvert Akershus). Dette er veldig vemodig. Tenk på alle de fine studentene som jeg har møtt. Og alle de inspirerende og irriterende foreleser, som har satt sitt preg på meg som fag- og medmenneske. -Akk, for en tid vi har hatt sammen! Takk skal dere ha alle sammen.

Fagfellesskapet KBBMEBGF, skal ha stor takk for trampeklapp fra sidelinja. Nå sjåast vi snart til faglig spetakkel!

Takk til alle dere i NRK for alle råd og vink og all entusiasmen til metadata. Stå på! Det er et oppriktig ønske at oppgaven skal inspirere NRK til å tenke gjennom den nye indekseringspraksisen og foreta kontroller og utføre tiltak. Det er viktig å sikre optimal metadatakvalitet innenfor rammene av Metadatastandarden. Det er vår felles kulturarv!

En induktiv prosess er sannelig hektisk, så en varm takk går til alle dere som har sett over oppgaven.

Jeg takker hjertelig for all god hjelp og støtte fra veileder Ragnar Nordlie. Du skal vite det at uten din tålmodighet og noen små spark bak, så hadde jeg aldri blitt ferdig med dette livsverket *«God kveld og velkommen til Dagsrevyen»: En studie av innholdsbeskrivelser for nyhetsinnslag i NRKs arkiv fra oktober 2012*.

Til sist, dere der hjemme: Takk for all støtte. Dette hadde jeg ikke klart uten dere. Dessuten takk for; oppvask, klesvask, middag, rydding, lek og moro, kos og trøst.

Per, du er underbar! Og Trym, du ruler! Mitt hjerte det banker for dere!

-og størst av alt; I love metadata.

Oslo, juni 2014

Miriam Bakkeli

1 Innledning

1.1 Bakgrunn

NRK gjør mye for å holde tritt med den digitale samtida. Nye medier, nye tekniske muligheter, nye publiseringsplattformer og endrede krav fra brukeren gjør at NRK har endret seg. Fra å kringkaste én sending til ett stort publikum på et gitt tidspunkt, er sendinger i radio og tv i dag også tilgjengelig på nett og mobil. NRKs nye programspiller har slagordet «NRK – når du vil».

I kjølvannet av den teknologiske utviklingen har NRK endret arbeidsflyten for å kunne publisere på nett umiddelbart etter sending. Arkiveringen har gått fra å være sentralisert i arkivene til å bli desentralisert i redaksjonene. Metoden for indekseringen har endret seg fra detalj-beskrivelse av utvalgte programmer til en taggepraksis for alle programmer. Indeksering blir utført av flere ulike yrkesgrupper og ikke av rendyrkede eksperter som tidligere.

Tagging som forskningsområde har rukket å bli stort og differensiert siden fenomenet startet tidlig på totusentallet. Tagging er kjent for å kunne supplere et gjenfinningssystem på en fornuftig måte. Det enes om at tagging er brukergenererte termer som beskriver metadata om ressursen. Søbak har undersøkt 1800 tagger fra NRK og mener de i all hovedsak representerer emne (2013 s. 52-54; s. 81). Bibliotekarer er gjennom faglitteraturen kjent med kontroll på synonymi, homonymi, relasjoner og standardiserte former på emneord, navn og korporasjoner. Tagging kan være et annet ytterpunkt; helt fritt for kontroll og brukergenerert.

Den nye strategien til NRK innebærer at taggingen med liten grad av kontroll, skal beskrive innholdet i alle sendinger, både tv og radio. Tagging er en av de frieste former for indeksering. Den manglende vokabular kontrollen overlater mye ansvar til de som indekserer. Dessuten stiller det høye krav til et sofistikert søkesystem om gjenfinningen skal optimaliseres. Mindre kontroll med indeksering og flere som tagger, kan kanskje gjøre beskrivelsene lite konsistente, mangelfulle og mangefasetterte. Men man vet lite om taggingen hos NRK. Dette studiet skal undersøke taggene for å finne deres særegenheter både som taggekorpus og i lys av øvrige innholdsbeskrivelser.

1.1.1 Samfunnsoppdraget til NRK

Digitalisering av NRKs arkiv og bruk av nye publiseringsplattformer henger nøye sammen med samfunnsoppdraget til NRK. To stortingsmeldinger fra senere tid berører NRK sin satsing på digitalisering og publisering.

I *Kringkasting i en digital fremtid* (St.meld. nr. 30 (2006–2007)) diskuterer man hvorvidt de nye publiseringsplattformene, som mobil og internett, er en del av allmennkringkastingsoppdraget. Departementet anbefaler at internett blir omfattet av samfunnsoppdraget etter gratisprinsippet. De tar ikke stilling til nedlastning på mobil fordi markedet er nytt og uetablert (s. 106-109). *Nasjonal strategi for digital bevaring og formidling av kulturarven* (St. meld. nr.24 (2008-2009)) er fra stortingssesjonen etter. Den framhever og befester digitalisering som ett av NRKs allmennkringkastingsoppdrag (s. 96). Den skriver også at noe av det som har satt fart på digitaliseringa av arkivmateriale, er mulighetene for gjenbruk og tilgjengeliggjøring på nett og mobil (St. meld. nr.24 (2008-2009) s. 58).

Vedtekter for Norsk rikskringkasting AS (Kultur- og kirkedepartementet 2012) styrer driften av NRK. Kapittel II, også kalt NRK-plakaten, definerer samfunnsoppdraget til NRK. Vedtektene ble revidert senest i juni 2012. NRK-plakaten illustrerer rekkevidden av nåtidens samfunnsoppdrag: NRK skal styrke og understøtte demokratiet; være allment tilgjengelig; styrke norsk språk, identitet og kultur; etterstrebe kvalitet, mangfold og nyskapning; være et ikke-kommersielt tilbud (Kultur- og kirkedepartementet 2012 §12-§16). NRK-plakatens §17 handler om forholdet til såkalte nye medier. Publikumstilbudet på internett og mobil-tv skal være løpende oppdatert, attraktivt og i hovedsak gratis, og tjenestene skal blant annet bestå av internasjonale og nasjonale nyheter. De fleste programmer skal være tilgjengelig på nettet etter sending, minst i sju dager (§17).

NRK er en allmennkringkaster som betyr at tilgjengelighet og allmenn tilgang etter gratisprinsippet faller inn under samfunnsoppdraget, slik departementet uttrykkelig ønsker i disse stortingsmeldingene. NRK er en viktig del av vår felles historie og kultutarv. NRK-plakaten styrer driften av NRK og § 17 viser hvordan de må tilpasse seg den digitale hverdagen.

1.1.2 Ny arkiveringspraksis i NRK

NRK har vært igjennom en omstillingsprosess de kaller Morgendagens mediehus. Intensjonen og hensikten har vært å etablere et økonomisk handlingsrom ved å kutte i interne kostnader. Rapporten *Morgendagens arkiv* (Røed et al. 2011b) er en oppfølging av dette arbeidet og

berører arkivtjenestene til NRK. Ett hovedmål med omorganiseringen er ifølge denne rapporten, utfasing av arkivtjenestene. Rapportens mandat er å skissere en omorganisering av arkivenhetene og foreslå en ny arbeidsflyt for arkivering.

I kjølvannet av *Morgendagens arkiv* har NRK derfor endret arkiveringspraksis av radio- og tv-programmer og opprettet ny arbeidsfordeling (Røed et al. 2011b). Hovedansvaret for metadata er flyttet fra arkivene til de enkelte redaksjonene. I tillegg holder den nyopprettede avdelingen, Metadataseksjonen i Arkiv & Research, kurs i metadataføring og utfører veiledning ved behov. Det er ansatte med tilknytning til produksjon av tv og radio som utarbeider alle metadataene, og ikke egne arkivansatte som tidligere (Wettmark og Holgersen møte 20. september 2011; Wettmark e-post 14. mars 2013). Dagens arkivarer har en variert rolle som blant annet journalister, redaksjons- og produksjonsmedarbeidere. Nå blir alle programmer beskrevet utfra en ny metadatastandard, og ikke som før, hvor bare utvalgte programmer ble detalj-beskrevet av arkivarer (Wettmark og Holgersen møte 20. september 2011). Kort sagt; arkiveringen har gått fra å være sentralisert i arkivene til å bli desentralisert i redaksjonene.

Morgendagens arkiv (Røed et al.) vektlegger at den nye arkiveringspraksisen skal dekke tre delmål. Det første målet dreier seg om enkel og god gjenfinnbarhet, det andre skal sikre journalistisk medskapende research, og det siste delmålet framhever tilgjengeliggjøring [sic] av arkivinnhold (2011b s. 4). Rapporten *Tags i NRK* (2011) utdyper nødvendigheten av at metadataene også skal fungere i nett-TV, Programspilleren (Bakke & Fleicher 2011).

1.1.2.1 *Minimumsstandard for metadata*

Det er utarbeidet en minimumsstandard for metadata, her kalt Metadatastandarden. NRK har definert sju punkter for beskrivelse i den nye strategien, som skal sikre god arkivering og gjenfinning. Disse sju obligatoriske feltene er «Tittel», «Sendetid», «Programleder/team», «Medvirkende», «Rubrikk», «Tags» og «Rettigheter» (Røed et al. 2011a s. 3). Videre i oppgaven vil metadatafeltet *Tags* bli kalt «Tagger»

1.1.2.2 *Taggeregler og forventninger til taggene*

Metadatastandarden er supplert med retningslinjer for hva som er en «god» tagg. *Taggeregler i NRK* (Bakke 2012) bestod av ti punkter på det tidspunktet datamaterialet er hentet fra.

Retningslinjene til NRK er laget for å veilede de som fører metadataene, og de skaper visse forventninger til taggene. Forarbeidet til taggereglene, *Tags i NRK* (Bakke & Fleischer), gir at

taggene skal være innholdsbeskrivende, enkeltord eller begreper, frie og ikke-hierarkiske og de skal være et tillegg til andre metadata (2011 s. 3). *Taggeregler i NRK* (Bakke 2012) fremhever at det mest sentrale taggene skal dekke er «hva» som er innholdet. Om det er viktig for innholdet skal taggene dekke «hvem» eller «hvor».

Videre vil «gode» tagger i tillegg se slik ut: Taggene er på bokmål, og de er enkeltord og begreper, vanlig uttrykk. De er skrevet med små bokstaver. De gir kjente navn eller kallenavn på hendelser, og forkortelser når de er kjente. Taggene er presise og generelle. Taggene er også synonymer, men er ikke ord med dobbeltbetydning. Taggene er valgt med omhu når det gjelder betydning (Bakke 2012).

1.1.3 Bibliotekariske prinsipper

NRK har et relativt fritt indekseringsspråk, og indekseringspraksisen blir styrt med få og lettfattelige regler, som beskrevet over. Arkiverings- eller indekseringspraksisen til NRK følger ikke alltid faglige anbefalinger knyttet til gjenfinningssystemer og indeksering.

Indekseringspraksisen innebærer liten grad av kontroll, men avdelingen Metadataseksjonen i Arkiv & Research har en oppdragende effekt, med gulrot og pisk. De framhever redaksjoner med god praksis, og de viser til mindre heldige eksempler på uønsket tagging (Wettmark 6. november 2012). En forslagsliste med taggene gjør indekseringsspråket delvis kontrollert, i alle fall på papiret. Søbak (2013) kaller indekseringsspråket for et semi-kontrollert vokabular, men hennes studium avdekker ikke i hvilken grad eller om taggene fra forslagslista, faktisk blir valgt.

Flere av reglene *Taggeregler i NRK* (Bakke & Fleischer 2011) går på tvers av bibliotekariske konvensjoner. Retningslinjene strider mot den bibliotekariske «universalloven» spesifisitetsprinsippet. Lancaster forklarer spesifisitetsprinsippet som når indeksering skjer med den mest spesifikke termen (2003 s. 33-35). Regel 9 «bruk generelle tags i tillegg til de presise» bryter med dette (Bakke 2012).

Retningslinjene bryter også med innarbeidede prinsipper for å fremme gjenfinning. I *Emneordskatalogisering: Innholdsanalyse, emnerepresentasjon og lagring* (Hjortsæter 2005) blir det anbefalt å etablere kontroll over en rekke semantiske fenomener deriblant synonymi og homonymi. Hun anbefaler å praktisere fortrukket term og etablere relasjoner og henvisninger, slik at systemet sørger for best mulig gjenfinning. I den forbindelse skaper man en rollefordeling mellom gjenfinningssystem og indeksering (2005). NRK har ingen slike

semantiske kontrollfunksjoner i sitt system. De løser synonymproblematikkene ved å anbefale «legg til vanlige synonymer» (regel 7) og regel 11 sier «unngå ord med dobbelt betydning» for å unngå homonymi (Bakke 2012).

Ingen av Taggereglene sier noen om taggens ordklasse. Hjortsæter (2005) anbefaler som hovedregel substantiv i ubestemt form eller substantiviske uttrykk (s. 39). Videre gir hun råd om hvordan ulike begrepskategorier bør bli uttrykt. Blant annet bør Enheter, typer og deler av enheter som kan telles, settes i flertall ubestemt form (s. 41) og ting det finnes én av settes i bestemt form entall (s. 43). Hensikten med en standardisert framgangsmåte som det Hjortsæter skisserer, er å skape en konsekvent og forutsigelig indeksering (2005 s. 9).

1.2 Formål med masteroppgaven

Den opprinnelige forskningsinteresse var tilknyttet intern research ved samsøk på tvers av ny og gammel praksis. NRK har både intern og ekstern bruk som utgangspunkt for metadataproduksjonen. Men siden de nye retningslinjer for metadataføring er resultat av en omstillingsprosess, har intern bruk fått størst oppmerksomhet. Samfunnsoppdraget til NRK og endrede TV-vaner, gjør at publikums mulighet til å bruke nett-TV ble vel så viktig for studien. En samlet vurdering har gjort at denne oppgaven ser på metadataen fra den nye indekseringspraksisen uten en eksplisitt målgruppe i tankene. Et sentralt poeng med studien er å se potensialet i et utvalg av innholdsbeskrivende metadata.

Retningslinjene for metadataene er lettfattelige og fleksible, så de åpner for ulike tolkninger, se **Error! Reference source not found.**. De strider dessuten mot anbefalinger til indekseringspraksis og gjenfinnings-prinsipper. Dette kan være et problem for beskrivelsene av innhold fra radio og TV. En kjennskap til den nye praksisen kan hjelpe i framtidig indeksering eller ved søking. Men forhold rundt selve søkesystem, funksjonalitet og mulige gjenfinningsteknikker er utenfor denne studiens nedslagsfelt. Det er allikevel vanskelig å se på emnebeskrivelser uten å tenke på begreper som fullstendighet og presisjon, som er nært knyttet til forskning på effektivitet til gjenfinningssystemer.

Systemene hos NRK mangler en fastlagt semantisk struktur og kontroll. Taggene skal være innholdsbeskrivende, enkeltord eller begreper, frie og ikke-hierarkiske og et tillegg til andre metadata (Bakke & Fleischer 2011 s. 3). Retningslinjene for taggene gir at man kan uttrykke flere nivåer med taggene om nødvendig (Bakke 2012). Det er interessant å undersøke taggenes natur og eventuelle relasjoner mellom feltene Tittel, Rubrikk og Tagger.

Denne masteroppgaven tar utgangspunkt i den nye desentraliserte indekseringspraksisen. Undersøkelser av dagens taggepraksis skal vurdere egenarten til taggene og se dem i sammenheng med andre innholdsbeskrivelser. Det er et ønske at kartleggingen av taggene som verbale innganger og forholdet til konteksten de står i, skal avdekke hva taggene egentlig er.

Datamaterialet er hentet fra Nyhetsredaksjonen sin hovedsending, Dagsrevyen. Undersøkelsene har studert utvalgte felter som den nye praksisen, Metadatastandarden, omfatter. Disse feltene er Tagger, Tittel og Rubrikk samt feltet Medvirkende.

- Taggene skal angi innsalget «hva» (utdypet under).
- Tittel inneholder en lokkende og triggende overskrift (Liu samtale 24. mai 2013).
- Rubrikk gir en beskrivelse av tv-innslaget, men som oftest inneholder det nyhetsankerens introduksjon i studio (Wettmark møte og opplæring 20. september 2011).

Disse feltene blir her sett på som innholdsbeskrivelser. Denne studien ser både taggene som et taggekorpus og vurderer taggene i deres kontekst.

I mangel av autoritetsregister for navn, så blir også forholdet mellom Medvirkende og de andre feltene vurdert. Medvirkende blir undersøkt, fordi personer i nyhetsbildet kan være en naturlig søkeinngang. Medvirkende skal angi hvem som deltar i innslaget (Røed et al. 2011a s. 3). Taggene skal angi innslagets «hva», eventuelt «hvem» når det er vesentlig for innholdet (Bakke 2012). Det er forskjell mellom å være medvirkende og navngitt med tagg. En tagg skal altså være mer enn og synes i tv-ruta. Personer kan være registrert dobbelt, uten å være sakens egentlige kjerne.

1.3 Problemstillinger

Problemstillingene har vært under kontinuerlig utvikling i tråd med en induktiv prosess, men hovedspørsmålet har være uendret.

Hvilke karakteristikker -egenskaper, særegenheter, trekk eller kvaliteter- har taggene?

Det er to ulike innfallsvinkler til taggene i denne studien.

I. Først er taggene sett på som tagger. Deres opprinnelige kontekst kun medvirker til emnekategorisering. *Hva dekker egentlig en tagg?*

II. Deretter er taggene studert i forhold til andre innholdsbeskrivelser fra hvert enkelt innslag. *Hvilke egenskaper har taggene sett i lys av de øvrige innholdsbeskrivelsene?*

For å belyse dette tar undersøkelsene utgangspunkt i taggene.

Datamaterialet er i tillegg analysert med sikte på å gi kunnskap om frekvens, grundighet og utbredelse om taggene som en deskriptiv beskrivelse.

1.4 Oppgavens disposisjon

I kapittel 2 blir teori og litteratur gjennomgått. Kapittelet tar også for seg utvalgt litteratur som er bakgrunnen typologien undersøkelsene er tuftet på.

Kapittel 3 er metodekapittelet hvor arbeidsprosessen, metodiske overveielser og avgrensning. Kapittelet beskriver valg og bearbeidelse av datamaterialet, samt hvordan undersøkelsene er gjennomført. I kapittel 3.5 blir litteraturen som er vurdert i utviklingen av typologien gjennomgått, og typologien blir framstilt/vist fram.

Kapittel 4 omhandler resultater, analysen og diskusjon.

Kapittel 5 inneholder en oppsummering og diskusjon av konsekvenser indekseringspraksisen kan ha.

I kapittel 6 spinner tanken videre om flere interessante forskningsprosjekter knyttet til metadataproduksjon i NRK og indekseringspraksisen deres.

2 Litteratur og teori

Dette kapittelet tar for seg begrepsavklaring, teori, litteraturgjennomgang og litteratur som er brukt til inspirasjon for å lage en typologi til undersøkelsene.

De faglige referanserammene er fra indekseringsteori, emneordskatalogisering og tesaurus-relasjoner. Samtidig er det tverrfaglige fenomenet tagging av interesse. Gjenfinningssystemer hvor en fellesnevner handler om å oppnå kontroll over det språklige mangfoldet, står i kontrast til den nye indekseringspraksisen hos NRK med tagging.

2.1 Hedden og taksonomisten

Tradisjonelt har deskriptive metadata blitt laget av profesjoner tilknyttet bibliotekvesenet, og slike metadata er ofte forbundet med kvalitet. Metadata laget av opphavspersoner er ofte sett på som et alternativ. En annen tilnærming til produksjon av metadata er brukerne selv (Mathes 2004). En utvikling i tiden er at «hvem som helst» tagger. Hedden (2010) skriver i The accidental-taxonomist om nettopp «den tilfeldig 'taksonomist'». Arbeid tilknyttet taksonomi er i dag ofte blitt en mer eller mindre tilfeldig tildelt arbeidsoppgave, og er ikke et eget «yrke» som det var tidligere (Hedden 2010). Denne skildringen av tidsånden stemmer med utviklingen i NRK.

2.2 Beskrivelsesmetoder av emneinnhold

Det er flere beskrivelsesmetoder som har relevans for denne studien. Kategorisering, indeksering, klassifikasjon, sammendrag og tagging har på ulike måter betydning. Tagging vil bli kort omtalt her, men i kapittel 2.5 er tagging beskrevet ytterligere.

Beskrivelse av emneinnholdet kan skje med ulike nivåer av kontroll med vokabularet. Hedden beskriver fire nivåer med varierende grad av kontroll. Tagging i folksonomi er et ytterpunktet. Hun beskriver tagging, eller 'keywording', som å tilordne nye termer med liten eller ingen grad av kontroll. Taggingen blir ifølge Hedden normalt utført av ikke-proffe og er enkel, men graden av enkelhet blir bestemt av faktorene innhold, taksonomi og indekseringspolicyen (2010 s. 172). Det andre ytterpunktet hos Hedden er kontroller vokabular uten mulighet til å bruke frie nøkkelord (s. 190).

Kontrollert vokabular håndterer meningen med ord og fjerner tvetydigheten i naturlig språk. Taksonomier og kontrollert vokabular er typer klassifikasjonssystemer som definerer

relasjoner mellom termer. De hjelper til med å forstå, navigere og bruke fordi språket blir mindre tvetydig (Smith 2008 s. 67-68).

Broughton åpner sin bok *Essential Classification* (2004) med følgende påstand; «Classification is everywhere» (2004 s. 1). Det er en allmenn påstand at vi ustanselig klassifiserer og kategoriserer, helt fra vi er barn. Å putte ting i bås er en vanlig måte å forklare hvordan vi organiserer våre tanker og minner på eller hvordan vi tilegner oss kunnskap. Klassifikasjon og indeksering handler om å organisere og ordne verden etter det som likner hverandre.

2.2.1 Kategorisering

Kategorisering er en enkel måte å sortere verden på og ofte ser man etter felles egenskaper, men Lakoff mener at kategorisering viser seg langt mer komplisert i virkeligheten.

Lakoff skriver i *Women, fire and dangerous things* (1987) at den klassiske forståelsen av kategorier, arvet fra de greske filosofene, er mer en antakelse enn en teori (1987 s. 6). Dette var fram til 1980-åra et udiskutabelt faktum hevder Lakoff (1987 s. xii; s. 8). Innen den klassiske forståelsen deler kategoriene felles egenskaper. Det betyr at kategoriene relaterer seg til ting i virkeligheten og er adskilt fra det kroppslige (1987 s. 8). Lakoff mener denne oppfatningen bare er en liten del av bildet (s. 5), og han underbygger dette ved å sammenfatte andre forskere sitt arbeid.

Vi har i følge Lakoff (1998) kategorier for alt vi kan tenke på. Det meste av kategoriseringen skjer ubevisst og automatisk, i så vel konkrete som abstrakte kategorier. Denne nye forståelsen ser kategorisering i sammenheng med både erfaring og forestillingsevne. Det er en rekke menneskelige faktorer som spiller en vesentlig rolle blant annet nevropsykologi, menneskelig adferd og menneskets intellektuelle kapasitet. En innsikt i hvordan vi kategoriserer er sentralt i forståelsen av hvordan vi tenker og fungerer som mennesker, fordi all tankevirksomhet består av kategorier, hevder Lakoff (1998 s. 6-9).

Lakoff blir beskrevet som en av grunnleggerne for forskningsfeltet kognitiv lingvistikk i 1980-åra. I deres øyne refererer ord til det de kaller mentale konsepter som er basert på personlig forståelse, kroppslig og sosial erfaring. Ord har således ikke alltid referanse til ting i den fysiske verden (Simonsen 2012b).

Golder og Huberman (2006) skriver at tagging er «sensemaking» som handler om prosessen der informasjon blir kategorisert og merket, slik at meningen blir tydeliggjort. Alle har vi ulik bakgrunn, og det medfører forskjeller i tagging.

Jamfør Lakoff og Golder & Huberman så kommer kategorier og ord, eller tagger, med en bagasje som ikke er lik for alle. Dette er forhold som kan forsterke behovet for kontroll med et indekseringsspråk.

2.2.2 Indeksering

Hos Lancaster blir indeksering beskrevet som en aktivitet hvor man lager en representasjon av emneinnhold med en eller flere termer. Termene er verbale innganger som sørger for lokalisering, søkbarhet og gjenfinning ved emnesøk (2003 s.6). Broughton definerer emneordskatalogisering med den prosessen som behandler innholdet i dokumentet ved katalogisering (2004 s. 307). Men Broughton beskriver også indeksering som en prosess hvor man bestemmer innholdet og tilordner deskriptor (2004 s. 302). Hos Chowdhury (2010) er indeksering beskrevet som den prosessen der man lager et dokumentsurrogat ved å tilordne identifikator. Når indekseringen er basert på en konseptuell analyse av emnet, utføres det emneindeksering (2010 s. 77). Lancaster nevner at det er uenighet om terminologien, men det er enighet om at indeksering er prosessen der man tilordner termer (Lancaster 2003 s.6).

Det vanskeligste med emneindeksering er å oppsummere emneinnholdet i noen få ord. Uansett hva retningslinjene sier så vil det være en risiko for ulik beskrivelse. Dette er den største ulempen med manuell indeksering ifølge Chowdhury (Chowdhury 2010 s. 97).

Det er svakheter ved en manuell indeksering. Salton og McGill (ifølge Chowdhury 2010) beskriver to problemer. Der er ikke sikkert at innsatsen i indekseringen blir belønnet i vellykket gjenfinning. Det andre er at konsistens ved indeksering er vanskelig uansett fremgangsmåte (Chowdhury 2010 s. 112-113). Salton og McGill skriver også positivt om indeksering ved bruk av naturlig språk som en overgang fra kontrollert vokabular til automatisk indeksering (1983 s. 55-59).

2.2.3 Klassifikasjon

Klassifikasjon er grunnleggende for organisering av samlinger og til hjelp i prosessen med søk og gjenfinning (Broughton 2004 s. 5).

Kontrollert vokabular som klassifikasjon er hjelpemidler både i indekserings- og søkefasen (Chowdhury 2010 s. 77-78). I følge Broughton (2004) har begrepet klassifikasjon tre

betydninger. Den første handler om tilordning av objekter til klasser, den andre gjelder identifisering og organisering i et klassifikasjonsskjema og til sist betyr klassifikasjon selve tilordningen av klassesymbol eller emneord (2004 s. 296). Broughton ser også klassifikasjon som grunnleggende for organisering av samlinger og til hjelp i prosessen med søk og gjenfinning (2004 s. 5).

Chowdhury ser som nevnt, klassifikasjon som et hjelpemiddel hvor man tilordner klassebetegnelse for alle dokumenter tilhørende klassen og hvor klassebetegnelsen da speiler emnet for dokumentet. Han trekker fram fire hovedfunksjoner, og de er alle tilknyttet notasjonene som hyllesignatur. Han forklarer at klassifikasjonsskjemaet gir dokumenter hylleplassering som da igjen samler dokumenter med samme eller liknende emner. Oppslag i katalog gir referansen til hylla og til sist muliggjør de strukturelle egenskapene i klassifikasjonsskjemaene manøvrering ('browsing') i en samling (2010 s.78).

Klassifikasjonssystemer har problematiske sider som at endringer tar tid, at de gjenspeiler ett verdensbilde og er preget av den kulturen og tidsperioden de ble laget, og dessuten kan de stride mot enhver fornuft (Smith 2008 ss. 86).

Chowdhury trekker fram tre ulike varianter av klassifikasjonsskjemaer. Broughton supplerer Chowdhury i de tre tilnærmingsmåter som bli nevnt videre. Disse som står i kontrast til fri tagging.

2.2.3.1 *Enumerativ klassifikasjon*

Chowdhury sier enumerative skjemaer lister opp alle de mulige klassene (2010 s. 79). Broughton beskriver enumerativt klassifikasjonssystem som pre-koordinerte, sammensatte klassesymboler (2004 s. 31-32). Chowdhury sier at disse skjemaene er rigide og det kan være vanskelig å utvide skjemaet med nye klasser. Enumerative skjemaer repeterer samme emne under flere klasser som gjør skjemaet stort (2010 s. 79-80). Broughton trekker fram at klassene er underordnet med økende spesifisitet og derav hierarkisk (2004 s. 31-32).

2.2.3.2 *Analytisk-syntetisk klassifikasjon*

Analytisk-syntetisk klassifikasjonsskjema har hjelpetabeller for egenskaper som forekommer ofte, og klassenummer er sammensatt ved syntese (Broughton 2004 s. 33). Analytisk-syntetiske klassifikasjonsskjemaer overkommer noen av problemene med enumerative skjemaer. Chowdhury peker på at flere tabeller gjør skjemaene kortere og tilbyr mer

fleksibilitet, men samtidig blir klassifikasjon mer komplisert. Analytisk-syntetiske klassifikasjon har regler for notasjonen (2010 s. 80).

2.2.3.3 *Fasettert klassifikasjon*

Det er Ranganathan og hans Colon Classification som blir regnet for å være opphavet til fasettert klassifikasjon (Broughton 2004 s. 34). Ranganathan var i alle fall den første som utarbeidet en helhetlig modell for analyse (s. 259). Han utviklet de fundamentale kategoriene kjent som PMEST (s. 259). Alle bestanddeler av emnet blir gjenspeilet i en sammensatt klasse (s. 34; s. 258). Ranganathan sammenliknet selv fasettert klassifikasjon med byggesettet Meccano (2004 s. 258). P er `personality´ og angir det særegne studieobjekt for emnet. M står for `matter´ og beskriver ganske enkelt materialer av ulike slag. Termer som beskriver aktiviteter og hendelser skjuler seg bak E for `Energy´. Til sist står S for `Space´ og T for `Time´(2004 s. 264).

I tillegg introduserte Ranganathan APUPA-mønsteret. U (umbral) representerer det sentrale emnet, mens P (peumbral) gir nært beslektede emner. A (alien) er ikke relatert til hoved-emnet. På denne måten blir beslektede emner samlet i den hierarkiske strukturer til klassifikasjonsskjemaet. Både PMEST og APUPA er representer i Colon Classification (Chowdhury 2010 s. 80-81).

2.2.4 Tagging

Ifølge Smith (2008) mener flere at tagging er radikalt forskjellig fra disse overnevnte klassifikasjonssystemene, og han er enig (s. 68). Han skriver at tagging er bruker-generert og ressurser kan få flere plasseringer, men tagging er også et komplement til tradisjonell klassifikasjon (Smith 2008 s. 67-68).

Tagging blir sett på som «sensmaking» av Golder & Huberman (2006). Det medfører at menneskets måte å kategorisere og sortere på gjøres ut fra ulike referanserammer påvirket blant annet av kultur og kunnskapsnivå. Det betyr at vi har ulike utgangspunkt når man bestemmer tagg.

Det er skrevet mer om tagging og tagger i kapittel 2.2.4.

2.2.5 Sammendrag

Aktiviteten emneindekseringen er beslektet med ‘abstracting’. Begrepet abstracting er å lage et sammendrag, eller mer konkret en kortfattet oppsummering av faglig innhold. Lancaster skriver at begge metodene for beskrivelse skal lage en representasjon av dokumentets

emneinnhold (Lancaster 2003 s.6). Indeksering identifiserer emnet med en deskriptor fra et kontrollert vokabular eller frie ord (s. 1). Et sammendrag skal indikere hva dokumentet handler om eller oppsummere innholdet kortfattet som løpende tekst. Lancaster trekker fram at indeksereren kan tilordne termer som supplerer de verbale inngangene i sammendraget. Jo lengere en representasjon er, jo flere verbale innganger (Lancaster 2003 s. 6-9).

Denne studien har sett på innholdsbeskrivelser av nyhetsinnslag. Tagger fra NRK blir sett i lys av andre innholdsbeskrivende tekster, altså løpende tekster som representanter for innholdet i innslagene. Fra NRK sin nye indekseringspraksis kan man trekke paralleller til Feltet Rubrikk som et slags sammendrag av innslaget, og feltet Tagger beskriver innslagets «hva» ved indeksering med frie nøkkelord. Men innholdet i Rubrikk må ikke forveksles med et faglige sammendrag. Det vesentlige er at egenskapene av å være løpende tekst kan ha visse likheter. Lancaster skriver at indeksering og sammendrag kan være utfyllende for hverandre. Dessuten vil en lengere representasjon ha flere potensielle verbale innganger.

2.3 Indekseringsprosessen

Lancaster beskriver indeksering i to trinn. Først finner det sted en konseptuell analyse og deretter en oversettelse. En forenklet forklaring på disse begrepene er at den konseptuelle analysen handler om dokumentets emneinnhold (2003 s. 9) og at oversettelsen gjelder hvordan emne skal bli uttrykt, enten ved tilordning eller ved avledning (2003 s.18-19).

Chowdhury deler indeksering i flere trinn enn Lancaster. Han velger å se videre på selve gjenfinningssystemet og etableringen av et indekseringsspråk. Første punkt er analyse av emnet, deretter valg av nøkkelord. Det tredje punktet består i å standardisere nøkkelordene og dernest å velge enten post-koordinert eller pre-koordinert indekserings-system. Det siste punktet består i å arkivere de verbale inngangene (2010 s. 100-102).

2.3.1 Subjektiv og objektiv emnebeskrivelse

Emne kan være subjektivt eller objektivt beskrevet. Et objektivt bestemt emne kan flere enes om, mens subjektivt bestemt emne er av privat karakter (Bill Marton ifølge Svenonius 2000 s.46). Svenonius problematiserer prosessen med å bestemme emne. Istedenfor intellektuell innsats kan man bruke automatiske teknikker basert på grammatikk, statistikk eller lingvistikk, men uansett metode vil man støte på ulike problemer, og da særlig med hensyn til ikke-tekstlig materiale. Hun framhever clustering-algoritmer hvor likhet indikerer samme emne etter sammenlikning av beskrivelser for flere dokumenter. Dette er et imponerende arbeid med hensyn til suksess, effektivitet og kostnader (2000 s 46-50). Lancaster (2003 s. 82)

avslutter kapitelet som problematiserer konsistens i indeksering, med at arbeid med søkeprosessen kan kompensere for de menneskelige feil i indekseringen.

2.3.2 Aboutness

Det første punktet ved indeksering handler om å avdekke dokumentets «aboutness» eller dets «iboende aboutness» (Chowdhury 2010 s. 101). Lancaster, Elliker & Harkness (1989) skriver at begrepet aboutness er unnvikende og nærmest «glatt som såpe» (egen oversettelse). De lager et skille mellom iboende aboutness (intrinsic aboutness) og utenforliggende aboutness (extrinsic aboutness). Emneinnholdet i et dokument er beskrevet som iboende aboutness, mens bruksområder, anskaffelsesgrunner og andre ytre forhold, er karakterisert som det utenforliggende aboutness (Lancaster, Elliker & Colonell 1989 s. 36).

Broughton skriver at bestemmelse av emnet er en veldig subjektiv aktivitet, fordi den innebærer tolkning og det er en intuitiv prosess. Det er derimot noen teknikker for innholdsanalyse som kan hjelpe til med å etablere gode rutiner for en effektiv emneordskatalogisering (2004 s. 52-53).

2.3.3 Begrepsorientert eller spørsmålsorientert indeksering

Både Lancaster (2003) og Chowdhury (2010) understreker at det å bestemme emneinnholdet ikke er en enkel prosess, fordi *hvordan* man skal bestemme emnet er komplisert. Det er to tilnærminger som nærmest rivaliserer; begrepsorientert og spørsmålsorientert indeksering. Lancaster skriver at det er ved begrepsanalysen brukerens behov blir identifisert av indeksereren, og ved oversettelsen så velger indeksereren den mest passende uttrykksformen for brukerens behov (2003 s. 27).

Dagobert Soergel skilte i O*rganizing Information: principles of database and retrieval systems* (1985 ifølge Raieli 2013) mellom begrepsorientert indeksering (entity-oriented indexing) og spørsmålsorientert indeksering (request-oriented indexing). Den begrepsorientert prosessen velger emneord som passer best til innholdet i dokumentet. Mens den spørsmålsorientert prosessen tenker på hvilke deskriptorer som vil bli brukt i søk etter dokumentet (2013 s.70-71). Ifølge Svenonius argumenterte Soergel for den spørsmålsorienterte indeksering framfor den begrepsorienterte. Svenonius hevder derimot at man trenger dem begge fordi de supplerer hverandre. Begrepsorientering vil gi en plassering eller flere i et univers, mens en spørsmålsorientert tilnærming antakelig vil gi flere mulige innganger fordi et dokument kan være svar til flere spørsmål (Svenonius 2000 s. 135). Raieli (2013) mener at det ensidige fokuset på å avdekke emnet, hemmer bruk og favoriserer den

spørsmålsorienterte prosessen fordi aboutness-modellen er begrepsorientert og leter etter ett eller noen utsagn som skal dekke emnet i sin helhet. Da utelukker man ifølge Raieli antakeligvis flere potensielle muligheter for bruk av dokumentet (2013 s. 70).

Chowdhury (2010) skriver at hovedoppgaven til et gjenfinningssystem er å angi innholdet i et dokument som matcher brukerspørsmålene. Det kritiske punktet er å velge rett beskrivelse i forhold til potensielle spørsmål fra brukere. I den forbindelse legger Chowdhury vekt på alle formene for kontrollmekanismer man har utviklet som tesaurus, klassaurus og tesaurofassett. Disse hviler igjen på indeksereren både med hensyn til innsats og intellekt, men disse verktøyene har vist seg å være ineffektive i flere forsøk. Chowdhury framhever automatiske prosesser for indeksering som en løsning for å løsrive seg fra dette problemet (2010 s. 77-78).

Lancaster (2003) beskriver spørsmålsorientert indeksering i sin litteraturgjennomgang, men legger også vekt på aboutness som en del av indekseringsprosessen (s. 9-13;13-19). Han har ikke det todelte synet på tilnærmingsmåtene som Raieli.

2.3.4 Indeksering av bilder og levende bilder

Innen indeksering og gjenfinning av multimedia er det vanlig å skille mellom konseptbasert (consept-based) og innholdsbasert (content-based). Lancaster skriver at det innholdsbaserte refererer til automatisk indeksering og gjenfinning etter indre mønstre som farge, form og tekstur i bilder. Mens det konseptbasert sikter til tekstbaserte beskrivelser tilordnet av mennesker (2003 s. 215), og Chowdhury legger til at beskrivelsene kan være kontekstuell tekst eller tilordnede termer (2010 s. 350).

Tekst er essensielt ved beskrivelse av multimedia (Lancaster 2003 s. 235). Trant mener dette er selve nøkkelen til gjenfinning av bilder og at det er et sterkt behov for en standard (ifølge Lancaster 2003 s. 235). «Hva» noe handler om er vanskelig å avgjøre, men blir mer komplisert når man indekserer skjønnlitteratur, film eller bilder (Lancaster 2003 s. 18). Chowdhury trekker fram at katalogisering av bilder er mer komplekst enn tekstlig materiale (2010 s. 351). Lancaster skriver at skriftlig beskrivelse av bilder har sine begrensninger og at de er mer subjektive og inkonsistent beskrevet (2003 s. 217). Raieli har i *Multimedia information retrieval: theory and techniques* (2013) gjennomgått forskning på multimedia gjenfinning med hovedtyngden på innholdsbasert gjenfinning. Han understreker imidlertid at disse to teknikkene supplerer hverandre (2013 s. 3-12).

Lancaster trekker fram at forskning på gjenfinning av bilder og lyd har tiltrukket seg forskere fra andre fagfelt som er ukjent med tekstgjenfinning. Han mener flere forskere innen gjenfinning av multimedia overser forskning på tekstgjenfinning og oppfinner fenomener på nytt. Han nevner for eksempel at tekstlig beskrivelse av bilder blir kalt annotering, og det hevder Lancaster, per definisjon, er det samme som indeksering (2003 s. 246).

Men mye har hendt siden begynnelsen av 2000-tallet. *Introduction to modern information retrieval* (Chowdhury 2010) er en nyere litteraturgjennomgang hvor det står at bilde- og videogjenfinning har ofte tekstlig beskrivelse (s. 350). Innholdsbaserte systemer som gjenfinner etter tekstur, farge og form ble forsket på fra 1990-tallet (2010 s. 345-346), men pixler er ikke meningsbærende i samme utstrekning som tekst (s. 350). Innholdsbasert gjenfinning ser etter mønstre i selve bildet, og slike systemer blir kalt content-based image retrieval som forkortes CBIR. De fleste bildegjenfinningssystemer er derimot basert på gjenfinning i beskrivende tekst, skriver Chowdhury (2010 s. 350-351).

Mange av teknikkene fra bildegjenfinning kan brukes ved videogjenfinning, men prosessen er mer kompleks. Film kan brytes ned til hver scene som kan beskrives. Video-gjenfinning er på et tidlig forskningsstadium (Chowdhury 2010 s 356).

Selv om det arkiverte materialet hos NRK er levende bilder, så handler denne studien om den tekstlige beskrivelsen av tv-innslagene. Det har foregått en konseptbasert indeksering.

2.3.5 Deskriptiv indeksering og emneordskatalogisering

Noe litteratur er opptatt av skillet mellom deskriptiv indeksering og «emneordskatalogisering» (subject indexing) deriblant Soergel Dagobert (ifølge Kirkegaard 2008). Det deskriptive omhandler det håndfaste som tittel og opphavsmann, mens den emnebaserte indekseringen er et resultat av en intellektuell innsats for å avdekke emnet. Denne forskjellen blir opphevet i et elektronisk miljø forklarer Kirkegaard (2008 s. 86-87).

Materialet fra NRK er en slik blanding. Tagger og Rubrikk vil tradisjonelt bli plassert innen emneindeksering siden de beskriver innholdet. Tittel blir betraktet som en del av den deskriptive indekseringen, selv om den er innholdsbeskrivende i mange tilfeller.

2.4 Begrepsavklaring og forklaring av prinsipper

I dette kapittelet blir utvalgte begreper med betydning for studien forklart. Det er ikke alle disse begrepene som er ikke aktivt brukt i undersøkelser, diskusjon eller analyse, men de er viktige referanse for helheten til studien og det teoretiske rammeverket.

2.4.1 Metadata

Milstead og Feldman beskrev det mangfoldige begrepet metadata allerede i 1999.

> Whether you call it cataloging, indexing, or metadata, the concept is a familiar one for information professionals. Now the electronic world has finally discovered it. Until a few years ago, only a few philosophers had ever heard of the word "metadata." Today, it is hard to find a publication about electronic resources that ignores it. (Milstead og Feldman 1999)

Metadata er på manges lepper fordi flere profesjoner har eierskap til begrepet. Innen bibliotek- og informasjonsvitenskap er fenomenet selve kjernevirksomheten, men ofte under betegnelsen bibliografiske data[1]. I et digitalt arkiv blir metadata helt vesentlig. Man kan snuble over en bok på hylle, selv om den ikke er i bibliotekkatalogen. Men en fil uten metadata finner man ikke «tilfeldig» på samme vis. Morville (2005) beskriver metadata som et tverrfaglig fenomen med lange tradisjoner i bibliotek- og arkivmiljøene. Metadata har forskjellige former og ulike formål. De kan blant annet beskrive emnet for et dokument med tanke på gjenfinning. I Morville egen terminologi angir man dokumentets aboutness for å støtte findability (2005 s. 125).

Metadata blir ofte beskrevet som «data om data» eller «informasjon om informasjon» (NISO 2004 s. 1). Denne vide forklaringen bidrar til et vidt spekter av anvendelsesområder for begrepet. NISO har delt disse inn tre kategorier; administrative, strukturelle og deskriptive. I videste forstand omfatter bibliografiske data alle disse aspektene ved en ressurs.

Arkiv- og produksjonssystemet til NRK, kalt Programbanken, inneholder alle disse ulike metadatakategoriene. NISO (2004) skriver at deskriptive metadata skal bidra til å oppdage, finne og identifisere. Strukturelle metadata angår sammensetningen av en ressurs. Administrative metadata er av teknisk art og gjelder forvalting og drift av resursen (NISO 2004 s. 1). Denne oppgaven ser på innholdsbeskrivelsene og det er deskriptive metadata. Eksempel på strukturelle metadata i Programbanken kan være innslag og deres rekkefølge i en sending. Administrative metadata i Programbanken kan være sendedato og rettigheter.

Tagging: people-powered metadata for the social web (Smith 2008) handler om brukergenerert metadata. Smith foreslår at metadata bedre kan bli forklart ved å si

[1] Bibliografiske data forklart http://openbiblio.net/principles/

«dokumentasjon av dataene» (s. 65). Denne beskrivelsen passer godt til bruken av begrepet 'metadata' innen bibliotekvitenskap.

Smith understreker videre at tagger er metadata, men at det ikke alltid er lett å avgjøre om de er deskriptive, administrative eller strukturelle (2008 s. 66).

2.4.2 Prekoordinert og postkoordinert

Det er vanlig å dele mellom pre-koordinert og post-koordinert indeksering eller systemer. Lancaster (2003) skriver at post-koordinerte systemer tillater søk som kombinerer termer (Lancaster 2003 s. 38). Mens en pre-koordinert indeks viser multidimensjonale relasjoner i en betydningsfull eller uttrykksfull rekkefølge, som er vanskelig å kombinere i søk. Som eksempel på pre-koordinert system er det vanlig å forestille seg en gammeldags kortkatalog (s. 51-52). *NRK praktiserer en post-koordinert indekseringspraksis, men det er funnet pre-koordinerte termer i datamaterialet. Jeg er usikker på om dette er tiltenkt eller en brukerfeil.*

2.4.3 Grundighet og spesifisitet

Gjenfinningseffektiviteten (i forbindelse med indeksering) blir målt etter indekseringens grundighet og termenes spesifisitet. Grundighet i indekseringen gir at dokumentets emne(er) er fullt uttrykt og speiler seg i indekstermene (Chowdhury 2010 s. 99). Broughton skriver at grundighet refererer til bredden i beskrivelsen. Analysen av et 'dokument' identifiserer alle aspekter ved dokumentet, og klassifikasjon eller indeksering er grundig når disse aspektene blir uttrykt med `klassemerker´ (Broughton 2003 s. 73-74). Lancaster skriver at grundighet i indekseringen eksisterer når nok termer er tilordnet til å dekke emnene. Det er et mål på antall termer som er tilordnet gjennomsnittlig. Lancaster understreker at grundighet sikter til bredden i emnene som dekkes, mer enn gjennomsnittlig antall. Lancaster nevner samtidig selektiv indeksering som bare dekker sentrale deler av dokumentets emner (2003 s. 27-29).

Spesifisitet i indekseringen er isolert sett, det aller viktigste prinsippet innen indeksering og stammer fra Cutter. Spesifisitetsprinsippet gir at indekseringen skjer med den mest spesifikke termen (Lancaster 2003 s.33-35). Chowdhury skriver at spesifisiteten sikter til hvor spesifikt et emne er uttrykt i hvert tilfelle (2010 s. 99). Broughton beskriver begrepet spesifisitet i forhold til hierarkier man finner i klassifikasjonsskjema (2003 s. 70-71). Spesifisitet betegner presisjon av deskriptor i forhold til emnet, og hun beskriver spesifisitet som nøyaktigheten til plasseringen av emne (2003 s. 73).

2.4.4 Fullstendighet og presisjon

Et gjenfinningssystems effektivitet blir ofte målt gjennom fullstendighet og presisjon. Fullstendighet er forholdet mellom `antall relevante dokumenter gjenfunnet´ og `antall relevante dokumenter i samlingen´. Presisjonen er forholdet mellom `antall relevante dokumenter gjenfunnet´ og `totalt antall dokumenter gjenfunnet´ (Chowdhury 2010 s. 99). Dette er to forhold som altså går på bekostning av hverandre.

2.4.4.1 *Konsekvenser av grundighet og spesifisitet; sammenhengen*

Grundighet ved indeksering (flere verbale innganger) sikter mot bedre fullstendighet. En kan si at grundighet i indeksering sikrer høy fullstendighet. Men når en øker grundigheten så minker muligheten for presisjonen. Dette skjer fordi en grundig indeksering dekker mange mulige emner også de som så vidt er behandlet i dokumentet, og derav en lavere presisjon i de dokumentene som er gjenfunnet. Jo mer spesifikke termene er, jo bedre blir muligheten for presisjonen. Men dette går på bekostning av fullstendigheten, fordi økende presisjon krever nøyaktig beskrivelse for emner som hovedsakelig blir behandlet i dokumentet. Av dette ser man at det ikke er mulig å oppnå både fullstendighet og presisjon samtidig. Man må derfor avgjøre rett balanse ut fra hva som gagner brukerne av systemet og hvilke kostnader dette medfører (Chowdhury s. 99-100).

2.4.5 Konsistens i indeksering

Mange studier har vist at å oppnå høy grad av konsistens når mange indekserer, er vanskelig. Voss understreker at oversettelsesfasen ved indeksering er vanskelig og at kontroll med vokabularet er en viktig faktor for å oppnå høyest mulig konsistens (Voss 2007).

2.5 Tagging

Smith skriver om tagging. Boka *Tagging: people-powered metadata for the social web* (Smith 2008) handler om hvordan brukergenerert metadata har endret bruk, deling og gjenfinning av informasjon. Forfatteren beskriver boka som en guide til hva er tagging og hvordan det gjøres (s. vii-viii). Smith skriver at tagger er hovedsakelig metadata om ressursen (2008 s. 5). Det er imidlertid ikke alltid lett å avgjøre om de er deskriptive, administrative eller strukturelle metadata (Smith 2008 s. 66).

Tokin et al. (2008) skriver at tagging som regel er fritt valgte nøkkelord. Tagging i folksonomier gir vilkårlige og skjønnsmessige ord som metadata for systemets objekter. Taggesystemet organiserer informasjonen for gjenfinning og innholdsbeskrivelsene (2008). Hjortsæter (2005) kaller frie nøkkelord for frittstående emneord. Tagger har mye til felles

med frittstående emneord. Hjortsæter skriver at flere frittstående emneord er hver for seg for vide med hensyn til dokumentets emneinnhold. Det kan skje en overbefolkning av termer som gir mange treff. Dette indekseringsspråket mangler syntaks, og det er derfor stor fare for feilkoblinger. Postkoordinert søking stiller krav til brukerens faglige kunnskaper, fantasi og kreativitet. Frittstående emneord er tids- og arbeidsbesparende ved indeksering, men man flytter sannsynligvis tids- og arbeidsforbruket til utforming og gjennomføring av søk (Hjortsæter 2005 s. 58).

Golder og Huberman (2006) beskriver tagging som selve handlingen hvor man organiserer ved å merke med tagger for å uttrykke meningene. Tennis sammenlikner sosial tagging med tradisjonell indeksering. Han sier om tagging «It is a catalyst for improvement and innovation in indexing.» (2006).

Smith skriver at tagging alene er som en bibliotekkatalog uten data. Taggingen er avhengig av ressurser og brukere, for å være en suksess. Tagge-systemet er hvor bruker, tagg og ressurs virker sammen (Smith 2008 s. 6). Cutters prinsipper vedrørende funksjonaliteten til en katalog. Disse går også ut på at opplysningene skal bidra med å finne, samle, identifisere og lokalisere ressurser (Cutter 1904 s. 12). Morville (2005) skriver at folksonomier støtter serendipitet, men er dårlig på gjenfinning av eller på nettsider (hans begrep er findability) fordi de ikke takler ekvivalens, hierarkiske og andre semantiske relasjoner (Morville 2005 s. 139). Mathes (2004) framhever at tagging passer bedre til browsing enn å finne relevante dokumenter (s. 6).

Kipp & Campbell (2006) sier at relasjoner finnes blant taggene i folksonomier. De er blant annet mellom stavemåte, akronymer og synonymer. Som andre trekker også Smith fram lenkene mellom taggene som gir inngang til nye ressurser eller ideer (2008 s. 2). I følge Kipp (2005) beskriver Peter Morville i *Ambient Findability* at folksonomier har lenker mellom emner som ikke blir uttrykt i en tesaurus. Slik gir folksonomier en ny dimensjon innen organisering og den er ikke skapt av fagfolk, men folket (Kipp 2005). I NRK er ikke taggene lenket sammen.

Weinberger er kjent for å ha knyttet uttrykket «wisdom of the crowd» til sosiale taggesystemer. Tanken er at «alle» bidrar slik at ressurser blir beskrevet av flere. Men dette er bare mulig hvis folket er motivert for en innsats (2006). Men i NRK sitt tilfelle har ikke taggingen en sosial side, og dermed vil ikke den kritiske massen ha en eliminerende effekt på dårlige tagger eller bekrefte de gode taggene.

Vander Wal (2005) uttrykker tanker om et skille mellom en bred folksonomi hvor flere tagger samme ressurs, og en smal folksonomi hvor en eller få tagger samme ressurs. NRK likner den smale, siden det er en eller noen få som tagger. Prinsipper til Weinberger «wisdom of the crowd» knyttes til den brede typen folksonomier. Quintarelli (2005) skriver at det finnes folksonomier for ulike objekter og brukerprofiler, og det gjør at folksonomier kan opptre ganske forskjellig. Den smale eller den begrensede folksonomien (narrow) mister styrken som finnes når mange brukere tagger. Den har allikevel fordeler i forhold til tradisjonelle metoder som fulltekst søk eller i forhold til tekstbaserte programmer på webben.

Lois Rosenfeld (2005) stiller spørsmål om hva som skjer i folksonomier når datamengdene blir uoverkommelig stor. Han sikter til de altfor mange generelle termene som beskriver innholdet. Rosenfeld skriver at kontrollert vokabular og folksonomier bør fungere sammen, for å få bedre innganger til å finne informasjon.

Tagger trenger ikke skape problemer for konsistens i indekseringen. Konsistens er problematisk uansett beskrivelsesmetode. Selv intra-konsistens, overenstemmelse i egen indeksering, er et problem (Lancaster 2003 s. 68).

Taggesystemer danner et distribusjonsmønster man kaller «power law» . Fordelingen av tagger blir slik at noen få forekomster er representert med høy frekvens, mens resten har lav frekvens (Smith 2008 s. 52-53). Studien til Halpin, Robu & Shepherd (2007) undersøkte populære bokmerker fra Delicious og fant også dette, men kaller det «short head» og «long tail», og understreker at dette viser hva taggerne mener er viktig. Smith skriver at det antakeligvis blir slik fordi tagger som allerede er bruk, har større sjanse for å bli brukt på ny. Forslagslister og synlige tagger forsterker dette hevder Smith (2008 s. 53). Men Yi og Chan (2009) viser til i sin artikkel at det ikke er entydige forskningsresultater som viser at forslagsliste øker bruk av tagger.

Spiteri (2007 ifølge Yi & Chan 2009) fant uregelmessigheter i taggene med blanding av entall- og flertallsform av substantiver, forkortelser, akronymer og homonymer i en undesøkelse av Delicious og Furl. Kipp & Campbell (2006) fant inkonsistent bruk av synonymer, akrinymer og stavevarianter blant taggene fra Deliciuos.

Heckner et al. (2007 ifølge Yi & Chan 2009) undersøkte ordklassene til tagger fra den sosiale bokmerketjenesten Connotea. De fant ut at 72 % var substantiver, 15% var akronymer, 12% var adjektiver og 1% var tall. Blant taggene var det ingen verb eller adverb.

Spiteri (2007 ifølge Yi & Chan 2009) understreker behovet for regler om tagging og foreslår å bruke skrivestøtte som ordbok for å minske tvetydigheten særlig blant homonymi og forkortelser.

Heckner et al. (2007 ifølge Yi & Chan 2009) sammenlikner tagger og forfaller-nøkkelord med andre bibliografiske data som tittel, sammendrag og fulltekst. De finner at bare 54 % av taggene er nevnt i disse metadataene. Det var tittel som hadde flest overlappende ord. Yi & Chan framhever at det var 30% av taggene som ikke ble funnet i andre metadata.

Voss (2007) har laget en typologi for taggesystemer. Han beskriver også trender i forskningsområdet relatert til tagging, som han kaller manuell indeksering på webben. Voss (2007) hevder at ofte velger forskere å se på tagging i kontrast til tradisjonell organisering som et nytt fenomen. Tagging kan bedre bli framstilt som en moderne form for manuell indeksering.

Populariteten til tagging har gitt manuell indeksering en ny vår. Et godt brukergrensesnitt som støtter tagging, kan være mer verdifullt enn enhver algoritme, skriver Voss (2007).

2.6 Inspirasjonskilder til typologi

Det er utviklet en typologi for undersøkelsene i denne studien. I dette kapittelet blir de faglige forutsetningene for typologien gjennomgått. Kapittel 3.5 viser typologiene slik de er brukt i denne oppgaven.

2.6.1 Taggekategorier

Golder og Huberman har delt inn tagger hentet fra den sosiale nettressursen Delicious etter taggenes funksjoner. De fant sju kategorier for taggene. 1) Taggene kan beskrive hva eller hvem det handler om, som emner, person eller organisasjon. 2) Taggene kan angi hvilken form for ressurs det er, som bok eller blogg. 3) Taggene kan markere eierskap. 4) Taggene kan modifisere andre tagger, men selv være uten opplagt mening. 5) Taggene kan angi karakteristikker eller egenskaper og viser taggerens mening om ressursen. 6) Taggene kan fortelle om relasjon til tagger, som «mystuff». 7) Taggene kan organisere samlingen utfra ulike gjøremål, som «toread» (egen oversettelse Golder og Huberman 2006).

Marlow, Naaman, Boyd & Davis fant seks motivasjonsfaktorer som de mente influerte taggingen og deres funksjoner. Faktorene var 1) Framtidig gjenfinning. 2) Deltakelse og deling. 3) Oppmerksomhet og blest. 4) Spill og konkurranse. 5) Framstilling av egen identitet. 6) Meningsytring (Marlow et al.2006).

Table 4.1 Seven Kinds of Tags

Tag Type	Examples
Descriptive	css, webdesign, ajax, Minnesota, drama, gardening, zen, microfinance, music, halo3, networks, sushi, hibiscus
Resource	blog, book, video, photo
Ownership/Source	nytimes, genesmith (author), newriders
Opinion	cool, funny,*****, lame, beautiful, crap, defective by design
Self-reference	mystuff, mine, me
Task Organizing	toread, todo, work
Play and Performance	squaredcircle, seenlive, aka vogon poetry

Tabell 1 Seven Kinds of Tags (Smith 2008 s. 67)

Smith (2008) sine kategorier baserer på inndeling av taggenes funksjon av Golder og Huberman, samt faktorer fra Marlow et al. Smiths taggetyper er vist i Tabell 1. Smith kuttet ut modifiseringskategorien til Golder og Huberman og erstattet den med en spill og underholdningskategori inspirert av Marlow et al.

Smith understreker som tidligere nevnt, at tagger er metadata, men at det kan være vanskelig å avgjøre deres funksjon. Smith beskriver de tre første kategoriene som hovedsakelig deskriptive metadata (2008 s. 66-67). Taggene kan bære preg av formålet for taggingen (Marlow et al. ifølge Smith 2008 s. 67).

Smith har også beskrevet fem vanlige formål med tagger. Det første gjelder personlig administrering og organisering av informasjon etter ditt eget system, som g-mail. Det andre er sosiale bokmerker som organisere nettsider med tagger og deler dette med andre i nettsamfunnet, eksempelvis Delicious. Det tredje formålet er å samle, organisere, dele og tagge objekter i digitale samlinger, som Flickr. En annen oppgave er at taggene skal øke findability, som å forbedre nett-handel-opplevelser. Den siste gruppen samler andre formål (Smith 2008 s. 6-12).

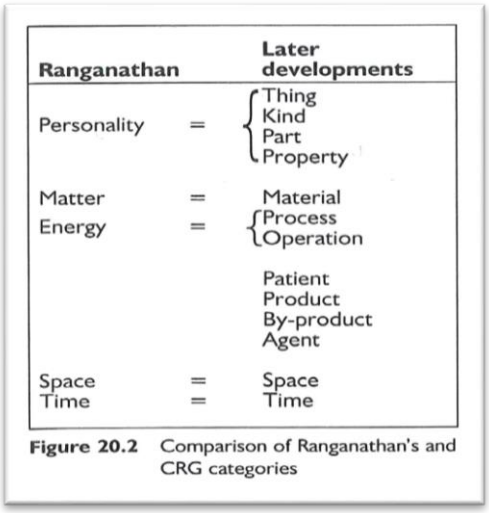
2.6.2 Emneordskategorier

Hjortsæter deler emne inn i begrepskategorier. De er: Enheter, typer og deler av enheter, materialer og råstoffer, egenskaper, prosesser, operasjoner og aktiviteter, fagområder og til sist begreper det finnes ett av. Hun hevder kjennskap til begrepskategoriene er nødvendig ved tilordning av de rette emneordene, jamfør hennes anbefalinger om emneordskatalogisering (2005 s. 18-19).

Hjortsæter (2005) deler også emneord inn i ulike typer utfra hva de beskriver ved ressursen. Det er tre grupper av emneord. Innholdsbeskrivende emneord påviser selve emnet for dokumentet. Neste gruppe er emneord som angir emner av mer allmenn karakter, men ikke form. Hun refererer i den sammenhengen til spesielle hjelpetabeller i klassifikasjonsskjemaer. Den siste typen beskrivelse gjelder indre form og har sin parallell til generelle hjelpetabeller (2005 s. 71-73).

Ideen om fasettanalyse er en videreutvikling av analytisk-syntetisk klassifikasjon hvor vanlige egenskaper blir utskilt i hjelpetabeller. Men det må ikke forveksles med analytisk-syntetiske skjemaer, fordi absolutt alt blir analysert og alle fasetter blir kombinert (Broughton 2004 s. 258-259). Ranganathan var altså ikke alene om den faseterte strukturen, men han var den første som utarbeidet en helhetlig modell for analyse med det fasetterte klassifikasjonsskjemaet Colon Classification publisert i 1933 (Broughton 2004 s. 259). Ranganathan la merke til at det var ulike typer av termer som gikk igjen ved beskrivelse av ulike emner (Broughton s. 264). Han utviklet de fundamentale kategoriene kjent som PMEST (2004 s. 259), tidligere er forklart på s. 12. I fasettert klassifikasjon vil hver fasett passe inn i en av PMEST-kategoriene (Chowdhury 2010 s. 80-81).

Classification Research Group (CRG) ble stiftet i 1952 (s. 259) og de arbeidet videre med Ranganathans ideer og valgte å utvide kategoriene (s. 260). Deres arbeid resulterte i utgaven Bliss's Bibliographic Classification 2 (s. 260). Utvidelsen av kategorien gjorde analysen enklere (s. 265) først og fremst fordi P ble delt i flere aspekter (som vist på Tabell 2).



Ranganathan		Later developments
Personality	=	Thing Kind Part Property
Matter	=	Material
Energy	=	Process Operation
		Patient Product By-product Agent
Space	=	Space
Time	=	Time

Figure 20.2 Comparison of Ranganathan's and CRG categories

Tabell 2 Kategoriene fra PMEST og Classification Research Group (Broughton 2004 s. 265)

2.6.3 Målestokk for relasjoner og semantiske forhold

Flere forskere som sammenlikner emneord, tagger og annen beskrivelse som titler, nevner skalaen til Voorbij (blant annet Kipp 2005; Kipp & Campbell 2006). Kipp (2005) undersøker om lag 1900 termer opp mot tittel til artiklene og plasserer dem i kategorier inspirert av Voorbij. Mange har også tilpasset skalaen til eget prosjekt. Kipp sammenlikner flere studier som ser på forholdet mellom tagging og kontrollert vokabular. Hun understreker at ulike definisjoner av nivåene gjør det vanskelig å sammenstille resultatene (2011).

Voorbijs skala er på sju nivåer. Han bruker nivåene for å vurdere verdien deskriptor fra kontrollert vokabular og Tittel som verbale innganger.

1. descriptor is exactly or almost the same as word from title;
2. descriptor is synonym of word from title;
3. descriptor is broader than word from title;
4. descriptor is narrower than word from title;
5. descriptor is related to word from title;
6. descriptor has a certain relation to word from title, but it is difficult to distinguish between 2, 3, 4 and 5;
7. descriptor does not appear in title at all.

(Voorbij 1998)

Voorbij (1998) deler skalaen i to (se ovenfor). Nivå 1 til og med nivå 3 tilfører ingen forbedring av beskrivelsen, mens nivå 4 til 7 potensielt forsterker beskrivelsen og dermed kan øke verdien av deskriptor.

Voorbij, og hans etterfølgere, sammenlikner deskriptor med et kontrollert vokabular (Voorbij 1998; Kipp 2005). Hos NRK mangler dette sammenlikningsgrunnlaget. Systemet har ikke et kontrollert vokabular og mangler derfor styring med ekvivalens, synonymi, homonymi og hierarkiske og sideordnede relasjoner. Siden NRK mangler den språklige kontrollen er ikke relasjonene eksplisitte hos NRK. Derfor vil denne undersøkelsen utvide flere av Voorbij nivåer med ytterligere detaljer som er av interesse.

Taggene hos NRK følger ingen streng indekseringspraksis som kan måle seg med tesaurus. Smith refererer til tesaurus som taksonomi på steroider. De kombinerer ekvivalens, partitive, generiske og assosiative relasjoner (2008 s. 72).

I emneordkatalogisering som hos Hjortsæter (2005) er et anbefalinger om hvilke termer som blir uttrykt og hvilke relasjoner som blir formaliserte. Det blir feil å sette NRKs tagger opp mot et kontrollert vokabular. Allikevel vil tankegangen fra en tesaurus hjelpe å fungere som

en målestokk i undersøkelsene av taggene og innholdsbeskrivelsen. Taggene kan sees i en tesaurusliknende kontekst, men da i en mer uformell variant hvor ordboka og allmenn språkkunnskap blir retningsgivende.

2.6.4 Semantikk og språkets mangfold

Språkets tvetydighet og kompleksitet er bakgrunnen for kontrollert vokabular i gjenfinningssystemer. Siden taggingen hos NRK er ukontrollert så er betydningen av taggene og relasjonene dem imellom og i forhold til rubrikk og tittel, av stor interesse. I sær for å få bedre kjennskap til hva den nye indekseringspraksisen innebærer og hvordan man kan optimalisere søketjenesten.

Semantikk blir i Store norske leksikon[2] beskrevet som læren om språkets innholdsside og betydning, enten av ord eller setninger (Simonsen 2012a). I denne studien er ord sentralt. Innen semantikken kan man skille mellom betydning eller mening og intensjon av et ord (*kvinne; hunkjønn*) og hvordan ord refererer til omverdenen (*kvinne; alle kvinner*). Ord kan altså ha en grunnleggende betydning (*kjerring; kvinne*) og en konnotasjon eller bibetydning (*kjerring; negativt ladet om kvinne*) (Simonsen 2012b). Denne undersøkelsen vil ikke gi en dybdeanalyse av disse betydningsforskjellene. Retningslinjene for tagger berører allikevel dette forholdet, når de oppfordrer til varsomhet ved ordvalg (Bakke 2012).

Videre forklarer Simonsen de betydningsmessige relasjoner mellom ord. De fleste av disse relasjonene er her gjenstand for undersøkelser. Antonymi har motsatt betydning (*liten* og *stor*) (2012a). Kvasisynonymer er brukt i emneordskatalogisering når begreper om ulike ting blir behandlet som synonymer. Disse er ofte ulike former for antonymi (Hjortsæter 2005 s. 35). Simonsen skriver at Synonymi har lik betydning (*fjernsyn* og *televisjon*) (2012a). Broughton definerer synonymi litt bredere som ord med lik eller hovedsakelig lik mening (2004 s. 308), og denne forståelsen må sees i forhold til et indekseringsspråk og gjenfinning. Videre forklarer Simonsen at homonymi har samme uttrykksform, men helt ulik betydning (*mark*) (2012a). Polysemi har også samme uttrykksform, men en beslektet betydning (*stjerne*) (2012a). Det kan være vanskelig å skille mellom homonymi og polysemi (Hjortsæter 2005-utg s. 36). Eventuelle tilfeller av polysemi vil bli behandlet som homonymi i denne oppgaven fordi de i gjenfinningssammenheng har samme konsekvens og problematikk.

[2] snl.no

Disse forholdene i språket er interessante for oppgaven, fordi systemet til NRK mangler et kontrollert vokabular. Det kan medføre at brukerne skaper relasjonen både fordi de blir oppfordret til det, men også fordi det kan falle naturlig å uttrykke seg forklarende når man tagger. Tagger og innholdsbeskrivelsene blir undersøkt med tanke på bruk av flertydighet/tvetydighet og homonymi, synonymi og antonymi.

2.7 Beslektet forskning

Det er utført ett forskningsarbeid som er beslektet med denne oppgaven. Søbak undersøkte taggene fra NRK i sin masteroppgave *Desentralisert indekseringspraksis: En studie av det semi-kontrollerte vokabularet i NRK* (2013). Hennes datamateriale var taggene fra samtlige sendinger på en vilkårlig dato i oktober 2012. Hun undersøkte taggene, indekseringskonsistensen og medarbeidernes holdninger til metadataproduksjon. Søbak fant en lav indekseringskonsistens og mange sendinger som ikke møter NRK sine krav til metadata. Det ble funnet lav indekseringskonsistens fordi et høyt antall tagger er brukt bare en gang (55 %). Hun antyder at de ansatte ikke forstår formålet med indekseringen fordi få sendinger hadde tagger, mange tagger var bare brukt en gang og indekseringen er ofte for spesifikk eller for generell. Taggene i studien hennes beskriver i hovedsak emnet for sendingene eller innslagene (Søbak 2013 s. 79-81).

3 Metodologi

Taggepraksisen er relativt ny for NRK og få har undersøkt tilsvarende praksiser tidligere, så man vet lite om de nye beskrivelsene av tv-innslagene. Målet med denne studien har vært å undersøke egenarten til taggene for nettopp å få bedre kjennskap til den nye praksisen. Det er utviklet en egen typologi til undersøkelsene av datamaterialet. Det er foretatt en grundig gjennomgang og analyse av NRK sine 19-nyheter fra oktober 2012.

Det har vært viktig at undersøkelsene blir forstått ut fra den organisasjonen som datamaterialet stammer fra, nemlig NRK. Derfor har arbeidsprosessen hatt behov for supplerende opplysninger om NRK generelt og deres indekseringspraksis spesielt. Slik informasjon er hentet fra flere møter og samtaler med nøkkelpersoner, som alle er tilknyttet metadataarbeidet i NRK. I tillegg har visning av søkesystemer og elementær innføring bidratt med en del opplysninger, som er viktige for denne oppgaven. Hovedsakelig er informasjonen brukt for å illustrere spesielle poenger. Dette gjelder blant annet uklarheter i hva taggene skal beskrive, eller hvordan de ulike tekniske systemene fungerer sammen.

Denne forskningen er både kvalitativ og kvantitativ. Det er utført en kvantitativ bearbeidelse og analyse. Dette gir et deskriptivt bilde av materialet som opptelling, frekvens og utbredelse. Det er verdt å merk seg at det ikke gir grunnlag for statistisk signifikans. Selve undersøkelsene er utført på et kvalitativt materiale. Datamaterialet er tekst og analysen ser blant annet etter tekstlige mønstre.

Forskningen har vært en induktiv prosess i den forstand at arbeidet med litteraturgjennomgang, teori, problemstilling, databehandling og analyse har foregått både vekselsvis og parallelt. Dette har vært nødvendig fordi kvalitetene ved datamaterialet har vært relativt ukjent. Framgangsmåten har ikke alltid gitt de antatte resultatene, så det har vært flere «bomskudd». Men framgangsmåten har også gitt uventede resultater og vist spennende tagging, som igjen har krevd nye undersøkelser og nye «verktøy». Drivkraften i arbeidet har vært nysgjerrighet, forventinger og antakelsene til datamaterialet, men forekomstene i materialet har styrt retningen for undersøkelsene.

3.1 Metodiske overveielser

3.1.1 Subjektivitet

Det har vært en utfordring å være varsom med subjektive vurderinger i bearbeidelsen og analysen av datamaterialet. Fire forhold har vært spesielt viktig for å etterstrebe en viss objektivitet. Det er grundig forberedelse, klare rammer for sammenlikninger og kategorisering, gjentakelse av forsøk og loggføring.

3.1.1.1 *Forprosjekt*

Arbeidet med masteroppgaven startet med få preferanser i tråd med en induktiv prosess. Hovedintensjonen var å undersøke taggene, fordi disse er nye med Metadatastandarden. Pilotundersøkelser ble gjennomført på et mindre datamateriale for å finne gode løsninger og interessante fenomener. I en induktiv prosess har jeg stadig kommet nærmere den metoden som ble benyttet i selve forskningen. Flere møter og ulik opplæring hos NRK har gjort meg skikket til å foreta datainnsamlingen og undersøkelsene. Mer om forarbeidet blir forklart i kapittel 3.2 og 0.

3.1.1.2 *Typologi og kategorisering*

Faste likhetsmål forebygger subjektiv vurdering. Ett eksempel på målestokk er nivåene inspirert av Voorbij (1998). I tillegg var andre størrelser emnekategorier, semantiske fenomener og ordklasser, blant annet.

Alle undersøkelser er gjennomført flere ganger, og resultatene av undersøkelsene er i etterkant blitt sammenliknet. Eventuelle avvik har medført ny gjennomgang. Disse gjentatte forsøkene og loggføring har motvirket en vilkårlig og skjønnsmessig behandling. Subjektiviteten har vært spesielt utfordrende i vurderingen av relasjoner, homonymi, kvasisynonymi og synonymi De assosiative relasjoner som ofte kan være svært så subjektive skal vurderes, er vanskelig å vurdere objektiv.

3.1.1.3 *Undersøkelsesguide og loggføring*

Datamaterialet er omfattende. Det består av store mengder tekst og flere av undersøkelsene er komplekse. Det ble derfor viktig at undersøkelsene er utført på en konsekvent måte. Forstudiene viste at gjennomgang av datamaterialet i enkelte tilfeller, gikk over flere dager. Dette ble utarbeidet en mal for hver undersøkelse, som skulle gi en mer konsekvent gjennomføring.

Disse undesøkelsesguidene inneholder nøyaktige beskrivelser av hvordan bearbeidelsen skal utføres. De ble fulgt i alle faser av arbeidet med datamaterialet. Hver arbeidsøkt startet med å lese malen. I forbindelse med de mest omfattende undersøkelsene ble en kortfattet sjekkliste benyttet, som var undersøkelsesguiden i stikkordsform.

Det har vært viktig å dokumentere prosessen for å opprettholde vitenskapeligheten. Alle undersøkelsene ble loggført og beskrevet, for å sikre en kontinuitet og konsistens i arbeidet. Dette har gitt god oversikt i mine vurderinger og begrunnelser. I tillegg ble undersøkelsene utført flere ganger og resultatene er sammenliknet, slik at tvilstilfeller ble avgjort utfra et samlet vurderingsgrunnlag.

3.1.2 Andre overveielser

Undersøkelsene kunne vært utført av flere testpersoner enn forskeren selv. Det kunne sikret en mer pålitelig behandling sett fra et vitenskapelig ståsted. I en induktiv prosess som denne studien har vært, ville det imidlertid vært vanskelig å involvere andre. Det ville vært en sjanse for frafall av testpersoner, fordi arbeidet har gått over flere semestre. Det ville også vært en risiko for feilbedømmelse etterhvert som analyseredskapene har utviklet og endret seg. Språk er mangfoldig, så mennesker kan ha ulik forståelse av enkelte beskrivelser. Denne studien ser på egenarten til taggene og deres forhold til konteksten. Min oppfatning av taggene vil illustrere viktige poenger like godt, som hvis flere hadde undersøkt de samme fenomenene.

«Uformell» informasjonsinnhentingen i form av enkel opplæring, samtaler og møter kan være en svakhet ved metoden. Det viser seg at det er flere ulike oppfatninger blant annet om forståelsen av hva taggene skal dekke. Intervju eller spørreundersøkelse kunne vært utført, for å avdekke forhold rundt forståelsen av praksisen på en mer systematisk måte. Men forståelsen av indekseringspraksis angår ikke denne studien, selv om indekseringsredskapet, taggene, er vesentlig. Disse ulike oppfatningene, eller uenighetene, illustrerer derimot poenget med en god indekseringspolicy som medarbeiderne har forståelse for, kjennskap til og innsikt i.

Sammenstillingen av tagger og innholdsbeskrivelsene kunne vært utført automatisk i et dataprogram. Dette ville sørget for en konsistent behandling av data med tanke på ordlikhet. En automatisk sammenlikning ville ikke kunne dekke det språklige mangfoldet slik denne studien gjør. Det er imidlertid flere forhold undersøkelsen omfatter som ikke kan bli fanget opp av en maskinell behandling, som for eksempel synonymi, homonymi, kvasisynonymi. Det er per i dag ikke utviklet et godt norsk vokabular for automatisk sammenlikning. En

studie av egenarten til taggene og deres kontekst slik det er utført her, er en manuell tilnærming trolig det beste.

3.2 Datamaterialet

Det ble foretatt en grundig vurdering av omstillingsprosessen, gammel og ny arkiveringspraksis og de tekniske systemene til NRK, før valget av et datamateriale ble tatt. Det er flere resonnementer som har foregått parallelt i lang tid. Dette ble gjort for å skaffe et helhetsinntrykk og oversikt. Denne innsikten medvirket til avgjørelser som sikret et datamateriale uten for mange forbehold.

Valg av dataleverandør, datakilde og datamateriale vil bli beskrevet nærmere i dette kapittelet.

3.2.1 Valg av dataleverandør

Arbeidet med denne oppgaven har, som nevnt, vært en induktiv prosess. I startfasen hadde prosjektet et internt fokus med interesse for utfordringer rundt research i arkivet med ny og gammel indekseringspraksis (samsøk). Selv om dette perspektivet etter hvert ble endret så har disse interessene hatt betydning for utvalg av datamaterialet.

Nyhetsredaksjonen er anbefalt av Metadataseksjonens gruppeleder, Maja Wettmark, som god datakilde siden de har tagget lengst og driver research i og gjenbruk av eget arkivmateriale (Wettmark, samtale 19. april 2013). Det er derfor spesielt viktig at metadatakvaliteten er tilfredsstillende. I tillegg har TV de beste tekniske mulighetene for å hente ut metadata (Johnsen samtale 18. april 2013; Wettmark19. april 2013). Siden TV-redaksjonene også startet først med endringsprosessene, og Nyhetsredaksjonen har praktisert den nye taggepraksisen lengst, så er forutsetningene for å studere den nye praksisen tilstede. Nyhetsredaksjonens tv-sendinger ble valgt som case og dataleverandør, representert ved hovedsendingen Dagsrevyen.

3.2.2 Avgrensning og begrunnelse for valg av dataleverandør

3.2.2.1 *Tekniske forhold*

NRK har av historiske, praktiske og tekniske grunner flere systemer som i løpet av en produksjon for tilordnet metadata. Ved research og søking har redaksjonene tilgang alle disse systemene. Det er avdekket i en spørreundersøkelse utført av NRK (Tremoen, Günther, Engan, Johnsen og Howlid 2013) at de ansatte bruker disse tekniske plattformene i svært varierende grad. Rapporten *Evaluering av metadataregistrering i NRK* viser at 58,8 % gjør

sin research i Programbanken og 26,8 % angir kategorien annet system (Tremoen et al. NRK *Evaluering av metadataregistrering i NRK* s. 8; s. 18).

Det er altså ikke ensartet bruk av ett arkiv eller én database hvor det er tilgang til alle metadata. En slik begrensning gjør sitt til at undersøkelsene ikke dekker research og indekseringspraksis helhetlig, og det interne fokuset for studien blekner.

På lengre sikt ønsker NRK å utarbeide et nytt system for metadata som høster alle relevante kilder. Dette systemet, Metadatabanken, har som hensikt å egne seg bedre til søking og research, enn dagens mange systemer (Holgersen møte 20. september 2011).

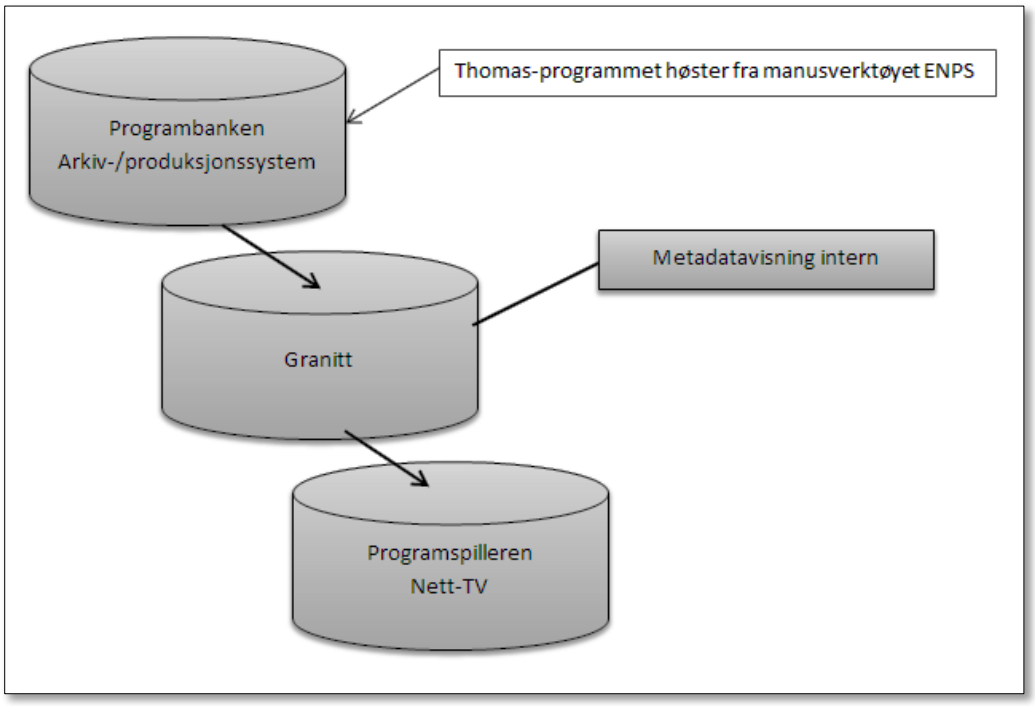
3.2.2.2 *Omfang av innholdsbeskrivelser*

Den nye Metadatastandarden ble etter hvert toneangivende for studien. Ifølge Engan inneholder Programbanken også to andre søkbare metadatafelt med innholdsbeskrivelser (Engan opplæring 18. april 2013). Disse feltene, Beskrivelse og Begivenhet, er utelatt fra datautvalg hovedsakelig av tre grunner. Først og fremst er feltene tidkrevende å hente ut fra Programbanken (mer om tekniske forhold i 3.2.3). Dernest er feltene ikke konsekvent benyttet. Til sist, er ikke feltene omfattet av den nye Metadatastandarden, som er bakgrunnen for denne studien.

Denne studien har undersøkt taggene, og de innholdsbeskrivende metadatafeltene Tagger (frie nøkkelord), Rubrikk (om innslaget) og Tittel (overskrift), samt «persongalleriet» i feltet Medvirkende. Alle metadataene er omfattet av den nye metadatastandarden. Gjennom hele prosessen har det vært et ønske å se etter potensialet blant disse metadataene.

3.2.3 Bakgrunn for valg av tekniske systemer og metadatafelter

Tv-redaksjonene har altså de beste tekniske løsningene for å hente ut egnet datamateriale i et håndterlig format.



Figur 1 De tekniske systemene til NRK

Arkiv- og produksjonssystemet for tv heter Programbanken. Dette systemet er ikke egnet til å hente ut de store mengdene metadata som dette studiet trenger. Datamaterialet blir derfor hentet fra programmet 'Metadatavisning intern' som gjør oppslag med databasen Granitt. Granitt inneholder de metadata som blir overført fra Programbanken til Programspilleren (Nett-TV). Disse metadataene er de sju feltene som blir omtalt som obligatoriske i Metadatastandarden (se 1.1.2.1 Minimumsstandard for metadata s. 3).

Rapporten *Evaluering av metadataregistrering i NRK* (2013 s. 4; s. 7) viser til at det er mer metadata registrert i Programbanken (PB) enn i den alternative kilden ENPS (Electronic News Production System - manusverktøy). Ifølge Wettmark er forklaringen til dette helt «naturlig». Programbanken samler all metadata som anses for å være korrekt. Hvor metadataene blir lagret avhenger av hvilken yrkesgruppe som registrerer det (Wettmark, møte 14. mars 2013). Engan understreker at Programbanken har «alt», og det hender at korrekt metadata bare blir registrert i PB og ikke ENPS (Engan, samtale 19. april 2013).

Men ved å velge Granitt som eneste dataleverandør går studien glipp av Beskrivelse og Begivenhet, som tidligere nevnt. Disse feltene er ikke i konsekvent bruk, og de er ikke omfattet av Metadatastandarden. NRK har ikke programvare som høster disse feltene og datainnhentingen må i tilfelle skje manuelt. Det vil innebære søk på hvert enkelt innslag og flere klikk for å finne rett visning. Deretter må eventuelle metadata med «klipp og lim» bli overført til et lagringsmedium.

3.2.4 Oppsummering for valg av dataleverandør

Det ble vurdert som hensiktsmessig å arbeide med dataene fra Granitt. Undersøkelsene vil da vedrøre Metadatastandarden og den nye indekseringspraksisen. Datamaterialet inneholder ikke beskrivelser utenfor den nye metadatastandarden som feltene Begivenhet og Beskrivelse. Hovedfokuset vil, som tidligere nevnt, være metadatafeltene Tagger, Rubrikk og Tittel.

Den nye metadataføringen vil også ha betydning for brukerne av NRK sitt nye nett-TV, Programspilleren (Bakke & Fleicher 2011 s. 3; s. 5). Det betyr at metadataene skal sørge for gjenfinning til flere målgrupper både internt og eksternt. I denne studien vil undersøkelser og analyse foregå uten en eksplisitt målgruppe.

Interessen for å søke på nyheter i Programspilleren (nett-TV) vil kanskje være begrenset for publikum. Det aller vanligste for nett-TV-tittere er nok å finne nyheter man har gått glipp av. Da vil tv-guiden, dato eller automatiske anbefalinger av de siste sendinger være en aktuell inngang. Emnebasert søk vil antakeligvis være basert på navn (korporasjon eller person), sak og hendelse eller kanskje sted.

3.2.5 Begrunnelse for utvalg og avgrensninger Dagsrevyen

Datautvalget er etter anbefaling hentet fra Nyhetsredaksjonens portefølje. Dagsrevyen er redaksjonens hovedsending og er prioritert med hensyn til metadataføring (Tremoen 2013 s. 3). Men valget av Dagsrevyen ble også basert på andre forhold.

Nyhetene på NRK er en veletablert sjanger. Dagsrevyen er normalt en sending som inneholder flere innslag (kalles indekspunkter i NRK). Hver sending inneholder vanligvis flere innenriksnyheter, flere utenriksnyheter, flere sportsinnsalg og avsluttes med en værmelding. I varierende grad inneholder sendingene direktesendte innslag og intervju direkte i studio. Alle sendingene i datamaterialet er ordinære nyhetssendinger som beskrevet over, altså ikke ekstrasendinger.

En velkjent sjanger gjør at alle har noen forkunnskaper som bidrar til færre forklaringer. Et homogent materiale bidrar også til en ensartet fremgangsmåte ved bearbeidelse og databehandling, som fremmer likebehandling. Selv om sendingene er skåret over samme lest, inneholder nyhetene et varierende saksinnhold, som skaper muligheter for bredde i taggingen.

Det var hensiktsmessig å finne et datamateriale hvor det var muligheter for å skalere ved behov uten store endringer. Nyhetssendingene viste seg å egne seg godt, fordi det er daglige sendinger. Dessuten har de en relativt konsistent praksis med utfylling av alle metadatafeltene som er av interesse for oppgaven. Nyhetsredaksjonen har som nevnt også utført den nye indekseringspraksisen i lengre tid, og de forholder seg til den nye pålagte metadatastandarden, ifølge evalueringsrapporten til Tremoen et al. (2013).

3.2.5.1 *Tidsavgrensning*

En tidsavgrensning for datamaterialet var nødvending for å sikre undersøkelser av den nye praksisen for metadataføringen.

NRK har vært i en omstillingsprosess over tid, og det har vært en blanding av ny og gammel praksis mellom 2009 og 2012 (John Arne Johnsen opplæring og visning 25. januar 2013; Tremoen et al. 2013 s. 3). Nyhetsredaksjonen overtok ansvaret for metadataføringen fra 01.01.2012, og metadataproduksjonen er helt desentralisert fra denne datoen (Tremoen et al. 2013 s. 4). Men ved spesielle omstendigheter eller behov kan redaksjonene be om assistanse til metadataproduksjonen fra Arkiv & research. Dette har ikke vært tilfelle i forbindelse med datamaterialet til denne studien (Wettmark e-post 14. mars 2013).

To forhold hadde betydning for tidsavgrensninga. For der første var det viktig at datamaterialet var fra etter innkjøringsperioden for Nyhetsredaksjonen og for det andre at den nye praksisen hadde pågått en viss tid. For å innfri disse forutsetningene anbefalte Maja Wettmark, seksjonsleder i Metadataseksjonen, sendinger etter september 2012 (møte 25. januar 2013).

3.2.6 Oppsummering av utvalget

Vurdering av omstillingsprosessen, gammel og ny arkiveringspraksis og de tekniske systemene til NRK, har ført fram til valget av datamaterialet.

Datamaterialet er hentet fra Nyhetsredaksjonen sine hovedsendinger klokka 19 og omfatter Dagsrevyen, Lørdagsrevyen og Søndagsrevyen i oktober 2012. Dette er Nyhetsredaksjonens hovedsendinger og blir derfor prioritert med hensyn til metadataføringen (Tremoen et al. 2013

s. 3). Nyhetsredaksjonen driver både med research i og gjenbruk av arkivert materiale. Derfor er det viktig at metadata-kvaliteten er tilfredsstillende.

3.2.7 Eksempel innslag fra datamaterialet

Norges Bank mangler 640 millioner
Medvirkende:
Trond Eklund (Medvirkende) , Carl Espensen (Medvirkende) , Gunn Hagenborg (Medvirkende)
Omtalt:
N/A
Sted
Norges Bank
Land:
Norge
Tags:
pengesedler, innlevering, norges bank, gjemt unna, verdiløs, destruering
Rubrikk:
JARLE: Om seks dager blir en rekke norske pengesedler ikke noe verdt. LISBETH: Og det finnes sedler for 640 millioner kroner der ute et sted.

Over er et eksempel på ett innslag fra Dagsrevyen 25. oktober 2012. Det er innslag 11 med Tittelen «Norges Bank mangler 640 millioner». Rubrikk er nederst i ruta. Taggen «pengesedler» har iden #25_11_01 (forklart i kapittel 3.3.1.1).

3.3 Framgangsmåte

Programbanken er, som tidligere nevnt, et produksjons- og arkivsystem for tv. Det er for tungvint å bruke til datauthenting. Derfor ble datamaterialet søkt opp i det interne systemet «Metadatavisning intern», som gjør oppslag i databasen Granitt. Granitt «leverer» metadata fra Programbanken til Programspilleren (Nett-TV). Granitt er ikke et arkiv, men et «reservoar av metadata» som inneholder de metadatafeltene som er interessante for studien. Se Figur 1 De tekniske systemene til NRK s.33.

I Granitt kan man gjøre oppslag på enten serie eller progid (id for sending). «Dagsrevyen 19» er lagt inn som serie sammen med Lørdags- og Søndagsrevyen. For å få tilgang til den aktuelle tidsperioden, ble progid for en tilfeldig valgt sending funnet i programmet Pi (*Programinformasjonen*, internt program). Deretter valgte jeg hele serien «Dagsrevyen» i grensesnittet til «Metadatavisning intern» for oktober 2012.

I «Metadatavisning intern» forekommer hver sending som eget punkt i trefflista. Disse ble slått opp i nettleseren. Visningen av sendingene med metadata for hvert innslag ble kopiert og lagret som PDF-filer, for å sikre at de ikke ble skadet eller forandret under bearbeidelsen.

Uthentingen av datamaterialet foregikk på det NRK kaller sendingsnivå. NRK kaller som nevnt, innslagene i en sending for indekspunkt. Det er utfylling av feltet Tittel som lager indekspunkt i systemet, som her kalles innslag. Uten Tittel-feltet eksisterer altså ikke innslaget som eget «objekt» i systemet, og det er da ikke lagringssted for metadata.

3.3.1 Bearbeidelse av datamaterialet

Bearbeidelsen av datamaterialet er av praktiske grunner gjort i flere faser. Grovt skissert ble det først gjennomført en kvalitetssikring av datamaterialet, og deretter ble datamaterialet undersøkt fra et induktivt perspektiv.

Forarbeidet ga en deskriptiv beskrivelse av datamaterialet og statistikk, men dette er uten signifikans siden dette datamaterialet er forholdsvis lite. Det ble utført ulike opptellinger for å si noe om generelt om utbredelse av innslagene, nyhetskategorier og tagging. Forarbeidet ble gjennomført manuelt og bearbeidelsen ble utført i Excel.

Kvalitetssikringen av datamaterialet ble gjort for å få et det homogent og sammenliknbart. Det innebar blant annet en grundig gjennomgang, hvor det datamaterialet som ikke egnet seg for undersøkelse og analyse, ble eliminert.

3.3.1.1 *Sporbarhet og ID*

Hver individuelle tagg ble tildelt en sporbar forbindelse, eller lenke, til både sending og innslag. Det ble laget en id til hver enkelt tagg som reflekterer nummer på sending, innslag og tagg. I teksten blir taggene referert til ved bruk av denne id-en slik (#SS_II_TT). «SS» angir nummeret på sending og er det samme som datoen i oktober 2012. «II» er innslagets nummer i sendingen før bearbeidelse og klargjøring. «TT» sikter til rekkefølgen av taggene for det konkrete innslaget.

3.3.1.2 *Kategorisering av innslag*

Innledningsvis ble det foretatt en kategorisering av innslagene etter funksjon og type nyhetsinnslag. Nyhetene er inndelt etter egenkomponerte nyhetskategorier. Saksopplysninger i feltene Tagg, Tittel, Rubrikk, Medvirkende og Sted var avgjørende for plasseringen.

Nyhetskategori

- Nyheter -innenriks
- Nyheter -utenriks
- Sport
- Gjest (ofte bortfall II og III)
- Direkte (ofte bortfall II og III)

Strukturell funksjon (bortfall I)

- Åpning
- Heading (overskrift til deler av sendingen)
- Værmelding

Informativ funksjon (bortfall I)

- Trailer

Ved inndelingen etter kategoriene og funksjonene nevnt over, ble det tydelig hvilke innslag som blir omfattet av Metadatastandarden, se Tabell 3.

3.3.1.3 *Bortfall av innslag*

Datamaterialet inneholdt enkelte innslag som er utenfor Metadatastandarden og derfor ikke egnet seg for undersøkelse og analyse. Datamateriale ble «vasket» for å fjerne støy.

Feilkilder ble fjernet av hovedsakelig tre grunner:

1. Innslag med strukturell eller informativ funksjon ble fjernet fordi de ikke er omfattet av Metadatastandarden.
2. Innslag uten tagg ble utelatt fordi sammenlikningsgrunnlaget er borte.
3. Innslag med ufullstendig beskrivelse eller innslag med opplagte feil ble fjernet for å unngå støy.

3.3.1.3.1 Funksjonell rolle

Først ble innslag som ikke er omfattet av Metadatastandarden utelatt. Dette gjelder innslag som etter kategoriseringen fikk rollene strukturell eller informativ funksjon, se oversikten over. Alle sendingene inneholdt ‘Værmelding’, ‘Åpning’, og ofte ulike headinger til deler av sendingen som ‘Sporten’. Disse organiserer sendingene og er tillagt en strukturell funksjon. Mens trailere for andre program som ‘Forsmak på kveldens Aktuelt’, ble regnet for å ha en informativ funksjon. Det var til sammen 91 innslag (av 722) som ble utelatt fra datamaterialet.

3.3.1.3.2 Manglede sammenlikningsgrunnlag

I neste omgang ble det stilt som krav til datamaterialet at innslagene måtte ha mimimum én tagg. Innslag uten tagger falt utenfor datamaterialet. Deretter ble det stilt krav om at innslagene skulle ha en Tittel og en Rubrikk med løpende tekst. Dette skulle sikre et utgangspunkt for sammenlikning i undersøkelsen mellom feltene Tagg, Tittel og Rubrikk. Alle innslag har Tittel siden dette feltet etablerer innslag i sendingen. Men enkelte innslag ble utelatt på grunn av tom Rubrikk.

3.3.1.3.3 Utilfredsstillende beskrivelse

De siste feilkilder ble oppdaget ved nærlesingen av alle innslagene. Det ble gjort en totalbedømming av beskrivelsene (Tagger, Rubrikk og Tittel) for å vurdere om de representerte en sammenhengende helhet og om beskrivelsene var tilfredsstillende. Det ble ikke tatt stilling til aboutness, men det ble sannsynliggjort at innholdet kan svare til en beskrivelse av innslaget.

Det ble foretatt en nærlesing av alle sendingene og hvert enkelt innslag, for å avgjøre om det var sannsynlig at Tagger, Tittel og Rubrikk var en sannsynlig beskrivelse av det konkrete innslaget. Noen av innslagene ble fjernet fra undersøkelsen fordi de opplagt inneholdt feil. Dette kunne innebære at rubrikk var fra et annet innslag eller at innslaget var et duplikat i sendingen (samme innslag har fått to av feltene Tittel). Når datamaterialet skapte mistanke om feil av denne typen, ble sendingene sett i Nett-TV (Programspilleren). Denne kontrollen ble bare systematisk utført ved rariteter. Det er ikke utført en systematisk og konsekvent titting på alle sendingene som undersøkelsen omfatter. Sendingen den 21.10.2012 er ikke tilgjengelig i nett-TV grunnet utløpte rettigheter selv om alle nyhetssendinger skal være tilgjengelig i ubegrenset tid. Men siden dette har vært en induktiv prosess så har alle sendingen blitt sett på ett eller annet tidspunkt. Denne tv-tittingen skjedde ikke som et konkret forskningsgjøremål, men mer av nysgjerrighet.

3.3.2 Undersøkelsene og analysefasen

Etter bearbeidelsen av datamaterialet som beskrevet ovenfor, ble det endelige datamaterialet undersøkt med hensyn til taggene og de utvalgte beskrivelsene. I forbindelse med undersøkelsene av taggene ble PDF-filene fra «Metadatavisning intern» kopiert, og sendingene ble lagret i Excel, for videre bearbeidelse.

Taggene ble studert og kategorisert som et korpus. Taggekorpuset på 1941 tagger ble kategorisert i flere omganger. Blant annet etter ordklasser og emnekategorier.

Taggene var springbrettet og utgangspunktet for undersøkelsene tilknyttet taggenes kontekst. Feltet Tagger (frie nøkkelord) ble vurdert opp imot metadataenefeltene Tittel (overskriften) og Rubrikk (innholdsbeskrivelsen). Det ble foretatt en kategorisering for å avgjøre eventuelle semantiske fenomenene og relasjoner som tidligere er gjennomgått. Feltet Medvirkende innebar mer registrering og liten grad av vurdering eller bedømmelse. Sted er tidligere i teksten antatt å være en interessant søkeinngang for brukere av nett-TV. Men feltet Sted er ikke omfattet av metadatastandarden og derfor er ikke feltet undersøkt.

3.4 Deskriptiv beskrivelse av datamaterialet

Datamaterialet består, som nevnt, av Nyhetsredaksjonens hovedsending klokka 19, «Dagsrevyen», fra oktober 2012. Det ble hentet ut 31 sendinger, men to sendinger var ikke overført korrekt og manglet innslag (indekspunkter) og metadata. Datasett I fra Programbanken består av 29 sendinger med til sammen 722 registrerte innslag og 2126 tagger.

3.4.1 Bortfall av innslag

En forutsetning i undersøkelsene var at datamaterialet måtte være omfattet av Metadatastandarden. Innslag av strukturell eller informativ art skal i henhold til Metadatastandarden, ikke bli beskrevet. Dette gjelder 91 innslag (Bortfall I). Det gir et potensielt matamateriale på 722 innslag (Datasett I) - 91 innslag = **631 innslag** (Datasett II).

Alle innslag i undersøkelsen måtte ha minimum én tagg. Det er 161 innslag i Datasett II som ikke har tagger. 631 innslag (Datasett II) - 161 uten tagger = 470 innslag med tagger (Datasett III).

Fordeling av innslag på Nyhetskategorier

Nyhetskategori	Programbanken	Datamaterialet	Bortfall %
Nyheter innenriks	307	235	23
Nyheter utenriks	113	75	34
Sport	131	86	34
Gjest	43	19	56
Direkte	37	17	54
SUM	**631**	**432**	**-**

Tabell 3 Fordeling av innslag på Nyhetskategorier

Det ble videre stilt som krav at innslagene i datamaterialet måtte ha tilfredsstillende Tittel, sammenhengende tekst i Rubrikk og mininum én Tagg. Disse skal samlet sett representere en sannsynlig beskrivelse av et innslag, men uten å ta stilling til aboutness. I datasett III var det 4 innslag som ikke har tilfredsstillende Rubrikk. En typisk feil er at tekst i Rubrikk er lik som forrige innslag, men Tittel og Tagger antyder på et helt annet saksforhold. Det var i tillegg 34 tilfeller hvor Rubrikk-feltet mangler innhold og det står tomt.

Dette medfører at 432 innslag som detalj-leses (Datasett IV).

Dekningsgraden for tagging vil mest rettferdig bli utregnet ved å eliminere innslag med strukturell og informativ funksjon. Det medfører en «tagge-dekningsgrad» på 76 % (470 innsalg av 631 har tagger).

3.4.2 Bortfall av tagger

Datamaterialet hentet fra «Metadatavisning intern» omfatter i utgangspunktet 2126 tagger fordelt på 722 innslag. Etter kvalitetssikring og nærlesing ble 185 tagger diskvalifisert. Disse er diskvalifisert fra undersøkelsen på grunn av sviktende sammenlikningsgrunnlag som f. eks manglende Rubrikk eller opplagte feil i Rubrikk. Ved ett tilfelle gjelder diskvalifiseringen tagging av en trailer altså et innslag utenfor Metadatastandarden (sending#29).

Det endelige datamaterialet inneholder 1941 tagger (2126-185) fordelt på 432 innslag. I tillegg har alle de 432 innslagene Tittel med mer enn ett ord og Rubrikk med en løpende forståelig tekst. Søbak undersøkte omtrent 1800 tagger fra NRK i *Desentralisert indekseringspraksis: En studie av det semi-kontrollerte vokabularet i NRK* (2013).

Diskvalifisering av tagger

	Tagger fra PB	Diskvalifiserte tagger	Tagger i datamaterialet
SUM	**2126**	**185**	**1941**

Tabell 4 Diskvalifiserte tagger

Bortfall av innslag, etter begrunnelse

		Bortfall I	Bortfall II	Bortfall III	Bortfall IV	
Dato «SS»	Antall innslag fra PB	Utenfor standarden	Uten tagger	Uten Rubrikk	Manglende sammenheng	Antall innslag
SUM	**722**	**91**	**161**	**34**	**4**	**432**

Tabell 5 Bortfall av innslag fordelt på begrunnelse. Vasking har skjedd kronologisk Bortfall I til Bortfall IV

3.4.3 Innfallsvinkel til datamaterialet

Det er hovedsakelig to innfallsvinkler til datamaterialet som begge er relevante for studiens problemstillinger og omfang. Man kan velge en periodisk eller saksorientert tilnærming, eller begge. I startfasen var det usikkert om datamaterialet skulle være fra en sammenhengende periode eller et utvalg av liknende saker over tid. Forstudiet viset at det var mest hensiktsmessig å samle inn innslag fra en gitt periode.

3.5 Typologien

3.5.1 Semantiske kategorier til denne studien

Inspirert av Voorbijs nivåer har denne studien utvidet kategoriene på grunn av manglende kontrollert vokabular. Jamfør Voorbijs todeling av skalaen hvor enkelte av hans nivåer forsterker beskrivelses, så vil alle kategoriene her være en forsterkning, foruten nivået «helt lik».

Nivå 1 Tagger er enten «Helt lik» eller «Delvis lik» Rubrikk og/eller Tittel. Dette refererer til ordlikhet, men «Helt lik» omfatter bøyingsformer. Tagger kan også være «Oppsplitta ord» sett i forhold til Rubrikk og/eller Tittel. Tagger representer også «Skrivefeil». Disse fire fenomenene blir registrert hver for seg.

Merkelappen «Oppsplitta ord» betyr her at Tagger er oppsplitta i forhold til Tittel og/eller Rubrikk. Dette blir sett på som en spesiell form av «Delvis lik» og er derfor plassert på Nivå 1. Kipp (2005) har Like termer og Oppsplitta termer i samme gruppe.

Nivå 2 Tagger er **synonymer** med Rubrikk og/eller Tittel. Synonymi er også kalt ekvivalens relasjon. I denne studien er målformvarianter (nynorsk og bokmål) blitt regnet som synonymer til tross for at retningslinjene krever at taggene er bokmål (Bakke 2012). Hjortseter anbefaler i *Emneordskatalogisering: Innholdsanalyse, emnepresentasjon og lagring* at ulike skrivemåter blir behandlet som synonymer (Hjortsæter 2005 s. 40; s. 33). Akronymer og andre forkortelser vil bli behandlet under synonymer, som også er i tråd med anbefalingene til Hjortsæter (2005 utg s. 32). Tagger blir sammenliknet med Tittel og Rubrikk og blir separat registrert som synonymi, målformvariant eller akronym/forkortelse.

Nivå 3 Tagger er «Bredere» enn Rubrikk og/eller Tittel det vil si at taggen hierarkisk sett er overordna.

Nivå 4 Tagger er «Mer snever» enn Rubrikk og/eller Tittel. Taggen vil være underordna i et hierarkisk perspektiv.

Nivå5 Tagger er «Relatert» til Rubrikk og/eller Tittel. Dette gjelder for assosiative relasjoner som ofte blir formalisert som sideordna henvisninger. I tillegg ble kategorien eksemplarrelasjon registrert.

Hjortseter (2005) skriver at assosiative relasjoner er subjektive og derfor vanskelig å avdekke. Om begreper er klart tilknyttet hverandre og de hverken er partitive, attributive eller generiske relasjoner som regnes de som assosiative (2005 s. 87).

Nivå6 Tagger har «Ubestemmelig» relasjon som er vanskelig å plassere i nivå 2, 3, 4 eller 5. Denne kategorien er uforandret fra Voorbij.

Nivå7 Tagger som er «Ulike» Rubrikk og/eller Tittel er ikke «Helt lik» eller «Delvis lik». De vil si at de ikke opptre annet sted i innholdsbeskrivelsene i samme innslag.

Det var vanskelig å plasser kvasisynomymi og homonymi inn i denne typologien. Kvasisynonymi ble antatt å passe best på nivået «Relatert». Dette er et interessant fenomenet på tvers av Tagger, Tittel og Rubrikk, men kvasisynonymi burde også bli undersøkt i forbindelse med tagge-korpuset. Det ble besluttet å undersøke Kvasisynonymi på taggene.

Det ble besluttet at Homonymi som fenomen best blir forklart i forbindelse med tagge-korpuset. Det er undersøkelse av flertydige tagger som er interessant og gjennomførbar. Om homonymi blir forklart og entydiggjort i forhold til feltene Tittel og Rubrikk er også interessant, men ressurs- og tidkrevende.

Typologi				
Nivå	Navn	Tagg	Tittel utdrag	Rubrikk utdrag
Nivå 1a	Eksakt match	'ekskjæreste'	…ekskjæreste	…ekskjæresten
Nivå 1b	Delvis match	'sjøtransport'	sjøtransportnæringen	sjøtransportnæring
Nivå 1c	Skrivefeil	'musemumsbygg'	-	-
Nivå 1d	Oppsplitta ord	'arbeidere'	utenlandske arbeidere	utenlandske arbeidere
Nivå 2a	Synonymer	'romfolk'	-	…sigøynere
Nivå 2b	Målformvariant	'vedlikehald'	-	vedlikehold
Nivå 2c	Forkortelse	'ap'	-	-
Nivå 3	Tagg Bredere	overgrep	-	voldtekt
Nivå 4	Tagg Mer snever	hjertemedisin	-	… medisinen
Nivå 5a	Relatert	-	-	-
Nivå 5b	Eks.-relasjon	'trener'	-	'vålerengatrener'
Nivå 6	Ubestemmelig	INGEN	INGEN	INGEN
Nivå 7	Ulike	'utrykningstid'	-	-

3.5.1.1 *Enkle og sammensatte termer*

På verbalplanet skiller man mellom enkelt og sammensatte termer. Sammensatte termer er flere enkelttermer i ett ord eller flere enkelttermer (Hjortsæter 2005). Emneord kan både være enkelttermer og sammensatte termer. Alternativt kan man splitte termene opp til enkelttermer, og da blir vokabularet lite (s. 57). Det er vanlig å splitte de fleste sammensatte termer ved postkoordinert indeksering (Hjortsæter 2005 s. 60).

3.5.1.2 *Ordlikheten*

I typologien er «helt lik» og «delvis lik» to begreper som handler om ordlikhet mellom feltene Tagger og Rubrikk og/eller Tittel. Disse begrepene sjeler til ideen om ordlikhet inspirert av Voorbij. En tagg er «helt lik» hvis utrykkene er eksakt like eller består av en bøyningsform av hverandre. Taggene er «delvis lik» bare hvis bestanddeler av ordet har eksakt likhet. Dette kan forklares med at når synonymi ser på matching av begrep eller selve ideen, men bruker ulike termer på verbalplanet, så er ordlikheten matching av uttrykk på verbalplan. Ordlikheten undersøker på verbalplanet, og undersøkelse av synonymien angår begrepsplanet.

Hjortsæter skriver at begreper tilhørende ideplanet blir uttrykt på verbalplanet med ett eller flere termer. Målet med kontrollert vokabular er å klargjøre denne sammenhengen. Det vil si å lage én-til-én korrespondanse (2005 s. 29).

Ved flertydighet er det en flere-til-én korrespondanse mellom begrep og ord/uttrykk (Hjortsæter 2005). For å oppnå en kontroll gjelder det å skaffe en én-til-én korrespondanse mellom verbalplan og ideplan i indekseringen. Hjortsæter skriver at termer/uttrykk hvor sammenhengen framgår eksplisitt som forklarende parentes, er en anbefalt løsning (2005 s. 35-36).

Studiet til Yi, K. & Chan, L.M. (2009) ser på ordlikhet mellom tagg og kontrollert vokabular, men ikke synonymi. De refererer som nevnt, til andre undersøkelser som har inspirert denne studien.

3.5.1.3 *Bestemmelse av relasjoner*

Bestemmelsen av hierarkiske relasjoner er hovedsakelig basert på hvordan Hjortsæter beskriver partitive, generiske og attributive relasjoner (2005). I tillegg har allmenne språkkunnskaper og ordbok vært til hjelp i selve beslutningen.

Hjortsæter skriver at partitiv relasjon mellom en helhet og en del. Hun eksemplifiserer det med at Nord-Norge er en del av Norge. De generiske relasjonene er mellom en klasse og dens medlemmer som også blir kalt en genus-species relasjon. Eksemplarrelasjoner er også en type generisk relasjon hvor klassen er uttrykt ved et fellesnavn og enkelttilfellet blir angitt med et egennavn (2005 s. 85-86).

I typologien er eksempelrelasjonene skilt ut som egen egenskap, fordi det er interessant om fenomenet er representert mellom Tagger og Rubrikk og/eller Tittel.

Attributiv relasjon er mellom begreper der selve sammenstillingen av dem, innsnevrer det enkelte begrep. Men en attributiv relasjon kan sees på som en slags generisk relasjon. Eksempelvis vil marmor være et attributt til bord, fordi marmor innsnevrer betydningen av bord (Hjortsæter 2005 s. 86-87)

Hjortseter gir anbefalinger om hvordan man kan etablere et henvisningsapparat mellom termer ved å formalisere relasjoner mellom begreper på idéplanet. De generiske, partitive og attributive relasjonene blir formalisert som hierarkisk relasjoner når de tilhører samme begrepskategori (2005 s. 85-92).

Broughton definerer bredere term når termen har en mer generell betydning enn termen den relaterer seg til (2004 s. 296). En mer snever term har mer spesifikk betydning enn termen den relateres til (2004 s. 303). I forbindelse med disse definisjonene refererer Broughton til et kontrollert vokabular.

3.5.2 Emneordskategorier til denne studien

Analysen av taggene ble utført flere ganger og i starten uten spesielle kategorier. Dette var for å finne kategoriene som best passer datamaterialet i denne undersøkelsen. Studien til Søbak (2013) var inspirert av Ranganathan og PMEST i sin kategorisering.

Etter inspirasjon av Golder og Huberman var utgangspunktet for kategorien emne; hva eller hvem det handler. Dette ble videre inndelt i kategorier som ikke ble benyttet videre i studien, men de baserte seg på et blanding av de overnevnte kategoriseringen. Det var behov for en spesialtilpasset emnekategori for ‘Hendelse eller nyhetssak’. I emnekategorien ble ‘Egenskaper’ og ‘Aktivitet, operasjon og prosesser’ underkategorier. ‘Person og navn’ ble en kategori, og det sammen med ‘Korporasjon’. I tillegg ble ‘Sted’ og ‘Tid’ egne kategorier. ‘Tid’ ble videre inndelt i ‘Periode’, ‘Tidspunkt’ og ‘Årstall’.

I tillegg har det blitt utført undersøkelser som har sett på taggenes ordklasser.

I prosessen med å bestemme taggenes kategorier var begrepskategoriene til Hjortsæter som nevnt i 2.6.2, en viktig målestokk.

3.6 Utdyping av problemstillingene

I. Taggekorpus

Taggene blir undersøkt som et taggekorpus for å se på «taggene som tagger». Oppgaven vil prøve å finne taggenes egenskaper, som listet opp nedenfor:

Hva dekker egentlig en tagg?

I hvilken grad har taggene en personlig karakter eller «sjargong» eller nyord?

Forkortelser; hvor utbredt er de kan man si de er «forklart» med andre tagger?

Praktiseres det overordninger og underordninger blant taggene?

I hvilken utstrekning forekommer det homonymer, synonymer eller kvasisynonymer blant taggene?

Hvordan fortoner faser seg som tagg?

I hvilken utstrekning blir taggene gjenbrukt?

II. Tagger i konteksten (skal flyttes til senere)

Taggene blir sammenliknet med de øvrige innholdsbeskrivelsene, altså feltene Tittel og Rubrikk, samt Medvirkende. Dette vil si noe om taggene i deres kontekst. Det ble utført en sammenlikning mellom tagg og de utvalgte metadataene. Til disse undersøkelsen ble det utviklet en typologi, se kapittel 3.5.

I hvilken grad er taggene «helt lik» eller «delvis lik» annen innholdsbeskrivelse?

I hvilken grad forekommer synonymi, homonymi eller kvasisynonymi mellom taggene og annen innholdsbeskrivelse?

Hvilken strukturell rolle spiller taggene i beskrivelsen? Finnes det relasjoner?

Analysen av materialet viser at orddeling er vanlig mellom tagg og øvrig innholdsbeskrivelse.

Hvordan er det med de oppsplitta taggene?

Personer i nyhetsbildet er viktig og feltet Medvirkende er omfattet av Metadatastandarden. Taggene og feltet Medvirkende blir sammenstilt for å undersøke «persongalleriet».

Hvordan blir personer tagget? Er feltet Medvirkende i aktiv bruk?

4 Resultat, analyse og diskusjon

Datamaterialet var opprinnelig på 722 innslag som inneholdt 2126 tagger fra Dagsrevyen. Bearbeidelsen medførte at det endelige datamaterialet består av 432 innslag. Det var 185 tagger som ble diskvalifisert for mangelfullt sammenlikningsgrunnlag, og i datamaterialet er det 1941 tagger som er undersøkt.

I dette kapittelet blir resultatene hovedsakelig presentert todelt i tråd med problemstillingene. Enkelte forekomster er allikevel behandlet samlet for å vise et helhetlig perspektiv. Først følger en presentasjon, analyse og diskusjon av fenomener tilknyttet problemstillingen rundt taggene som tagger. Deretter følger en presentasjon, analyse og diskusjon av taggene sett i lys av andre innholdsbeskrivelser, samt de fenomener som er felles for begge innfallsvinklene.

Det har vært to ulike innfallsvinkler til datamaterialet som har forskjellig perspektiv til taggene. Taggene ble undersøk som et taggekorpus hvor taggene blir sett på som tagger og undersøkt for å finne deres egenskaper eller særpreg. Hver individuelle tagg ble kategorisert etter egendefinerte kategorier og ordklasser.

Med den andre innfallsvinkelen ble taggene sett i lys av andre innholdsbeskrivelser med metadatafeltene Tittel og Rubrikk. Taggene ble kategorisert etter inspirasjon av Voorbijs nivåer (1998) og undersøkt for semantiske forhold, som tidligere er omtalt. Utgangspunktet og springbrettet for analysene av taggene og konteksten deres var metadatafeltet Tagger. Hver enkelt tagg ble vurdert opp imot resten av de innholdsbeskrivelsene i metadatafeltene. Feltet Tittel er innslagets overskrift og som teknisk sett etablerer et innslag i sendingen. Rubrikk inneholder som oftest innslagets introduksjon i studio eller reporterens opplesning i innslaget.

Felt Medvirkende ble undersøkt for å se på «persongalleriet», men det innebar mer en registrering og liten grad av vurdering.

Datamaterialet er hentet fra Nyhetsredaksjonen og omfatter Dagsrevyen, Lørdagsrevyen og Søndagsrevyen i oktober 2012. De er hovedsendingene og blir prioritert med hensyn til metadataføringen. Nyhetsredaksjonen driver både med research i og gjenbruk av arkivert materiale i tillegg til at sendingen er tilgjengelig på Nett-TV. Derfor er det viktig at metadatakvaliteten er tilfredsstillende. Nyhetsredaksjonens sendinger blir lagt ut i nett-TV med ubegrensede rettigheter.

Det er tre forhold som spiller inn i diskusjonen av resultatene. Litteraturen om tagging beskriver hovedsakelig særegenheter ved tagger i folksomomier. Dette står i kontrast til teori og litteratur om kontrollerte vokabularer. Til sist spiller NRK sin Metadatastandarden og egne retningslinjer for gode tagger inn.

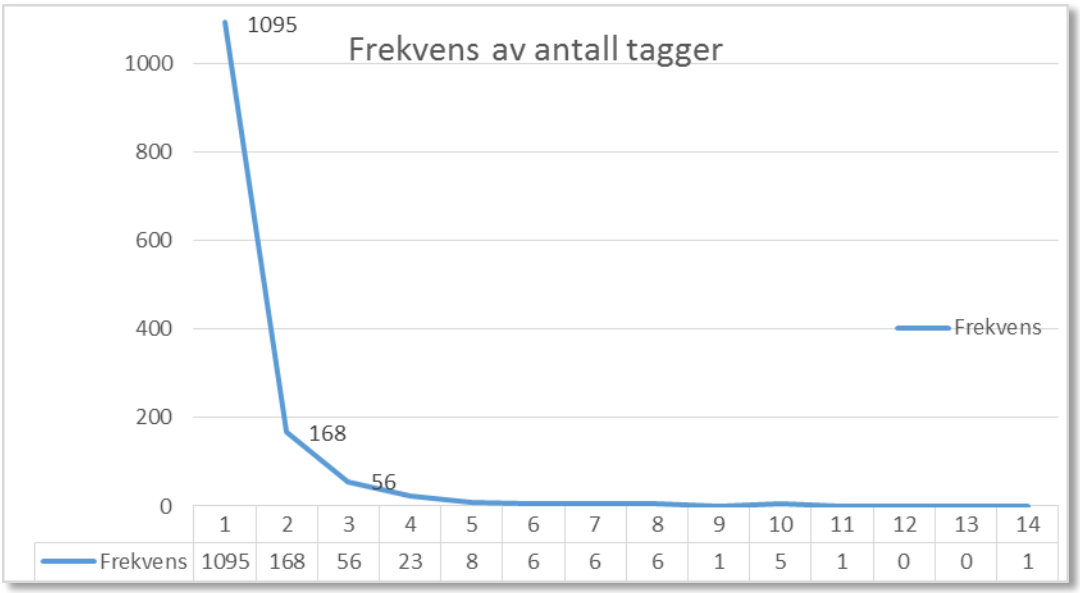
I denne studien blir ett eller flere ord fra metadatafeltet «Tags» sett på som en tagg. De blir gjennomgående kalt «Tagger». I datamaterialet er ny tagg markert med komma.

4.1 Tagger

4.1.1 Gjenbruk av tagger

Det var mange tagger som bare brukt én gang. Det gjelder 1095 av 1941 som tilsvarer 56%. Taggekorpuset følger power law-prinsippet (se Figur 2). Det etablerer seg i taggesystemer over tid. Det er et mønstre som utvikler seg hvor noen tagger har høy frekvens, mens de aller fleste taggene er brukt en enkelt gang (Smith 2008 s. 52-53). Halpin, Robu & Shepherd (2007) fant den samme tendensen, men beskriver den som «short head» og «long tail». Det er vært å merke seg at en distribuering av taggene etter dette prinsippet, er funnet i sosiale taggesystemer og etter en viss tid. NRK har ikke sosial tagging, og taggekorpuset i denne studien har ikke utviklet seg over tid på samme måte som ved sosial tagging. Allikevel viser frekvensen til taggene dette mønsteret. Golder & Huberman (2006) skriver at i Delicious tok et slikt mønster form etter omtrent 100 bokmerker, og taggefrekvensen står i forhold til det totale antallet tagger.

Kipp & Campbell (2006) hadde 30% av taggene fra Delicious som bare var brukt en gang i datamaterialet. Men igjen, dette er ikke direkte overførbart. Denne studien ser bare på ett lite utvalg av taggene fra NRK som er utvalgt etter spesielle kriterier, og NRKs taggesystem er ikke en folksonomi. Søbak fant tilsvarende tendens, og i hennes datamateriale er litt over halvparten av taggene brukt én gang (2013 s. 72).



Figur 2 Taggene etter frekvens

Topp-taggene viser i grove trekk hvilke saker som hadde bred dekning i Dagsrevyen i oktober 2012. Dette er på sett og vis et tegn på at taggingen reflekterer innslagenes innhold. Men i dette datamaterialet er det 30 % av innslagene som mangler en fullverdig beskrivelse i henhold til Metadatastandarden. Dette forholdet kan selvsagt påvirke topp-taggenes karakter. Det er en større andelen innslag som ikke følger Metadatastandarden. Det ble oppdaget flere tilfeller av tomme Medvirkende-felt ved kartleggingen av persongalleriet. Denne studien har bare registrert bortfall i forhold til Tagger og Rubrikk som er analyseenheter.

ANTALL	TAGGER
4	anke, blackburn, bombe, bortført, drillo, eliteserien, manager, new york, ny bok, presidentvalg, påkjørsel, romney, rune olsø, samferdsel, sigrid-saken, skatt, trener, trondheim, tyrkia, valgkamp, veiing, vekt, voldtekt
5	demonstrasjon, frp, krise, oslo, oslo tingrett, stortinget, trusler, usa
6	barn, beredskap, entra, eu, kampfiksing, overgrepstiltalt
7	jens stoltenberg, landslaget, orkan, statsbudsjett, sykkel, øygard-saken
8	arbeiderpartiet, drap, obama, ordfører, rettssak, uvær
9	overgrep
10	fotball, rune øygard, sandy, syria, vågå
11	politi
14	doping

Tabell 6 Tagger med frekvens fire eller mer

I oktober 2012 ble det avslørt to store dopingsaker i sykkelsporten; en internasjonal med Lance Armstorng og en nasjonal med Steffen Kjærgaard. Tabell 6 viser at omfanget av disse

sakene gjør taggene «doping» til den mest utbredte og «sykkel» til en av de mest utbredte. Frekvensen for taggen 'doping' er 14, mens taggen 'sykkel' er brukt 7 ganger (se s. 49).

Orkanen Sandy var mye i nyhetsbildet oktober 2012. Det er ingen av innslagene som har laget en samlende tagg for hendelsen. Taggene «sandy» og «orkan» er benyttet hver for seg som oppsplitta tagger. Disse to taggene er blant taggene med frekvens høyere enn fire i datamaterialet som vist i Tabell 6. Taggereglene i NRK tilsier at en kjent sak eller hendelse bør tagges med «kjente navn på hendelser» (Bakke 2012). I datamaterialet er det kun ett innsalg med taggen «orkan» som ikke handler om orkanen Sandy. Det er i tillegg flere innslag med taggekombinasjonene «sandy» + «uvær» eller «sandy» + «flom» ('flom' er brukt tre ganger). I denne forbindelse blir 'sandy' den taggen som sammenstiller innslagene som «handler om det samme», eller er innslagets «hva». Innslagene har ikke konsistent bruk av tagger. Det medfører blant annet at et kombinert søk på 'orkan' og 'sandy' ikke vil gi en uttømmende treffliste, altså gi fullstendighet.

I tidens hete er dette muligens en fruktbar løsning, men det er det kanskje ikke om ti eller tjuve år. I dette tilfellet er ikke taggen så spesifikk som mulig. Taggen kunne være indeksere pre-koordinert som for eksempel «orkanen sandy». Taggene 'orkan' og 'sandy' og de andre kombinasjonene nevnt over, kan fungere som syntetiske synonymer i et gjenfinningssystem.

Innslagene i datamaterialet har mellom én tagg og opp til 14 tagger. Tabell 6 viser taggene som er brukt fire ganger eller flere (se s. 58). Disse taggene står for 342 tilfeller av de 1941 tagger i datamaterialet og det svarer til 18%. 1599 tagger er brukt en, to eller tre ganger.

4.1.2 Kategorier og ordklasser

I kategoriseringen av taggene som et taggekorpus var konteksten avgjørende i selve kategoriseringen. Det er opplagt at en del tagger er tvetydige og vanskelig å plassere, når konteksten er utelatt. Taggen «rødt» (#02_03_06) er et eksempel på at kontekst er avgjørende for betydningen. Taggen «bjørnar moxnes» (#02_03_07) fra samme innslag sannsynliggjør at det er snakk om partiet Rødt. Rubrikken bekrefter med teksten «Rødt-lederen».

Tid (periode-«fire uker» og år-2012 tidspunkt-22.jul). «Når» kan være årsaken til en tagg og taggereglene eksemplifiserer tidspunkt med «idrettsgallaen 2011». I denne analysen av datamaterialet har slike «når» antakeligvis blitt kategorisert som EMNE-hendelse/sak. Egenarten til taggene av nyhetsinnslag tøyer her grensen for hva som tradisjonelt blir regnet som tid.

Taggen representerer hovedsakelig emne for innslaget. Dette sammenfaller med Søbaks undersøkelse av alle tagger fra en dag i oktober. To tredjedeler av taggene var relatert til emnet (2013 s. 52).

Gjennomgang av taggenes grammatikalske form viste at taggene hovedsakelig er substantiv, men taggene fordeler seg også over andre ordklasser enn det normalt er i et kontrollert vokabular. Heckner et al. (2007 ifølge Yi & Chan 2009) undersøkte tagger fra Connotea og fant flest substantiv blant bokmerkene.

Taggen «rives» (#23_17_01) er et eksempel på bruk av et passivt verb. Når regler for indeksering ikke anbefaler beskrivelser med substantiver og substantiviske uttrykk slik Hjortsæter gjør (2005 s. 39) eller gir andre råd om ordklasse, så blir det potensielle vokabularet så uendelig stor, og indekseringen kan virke uoverkommelig. Det vil være tilsvarende problemer ved søk.

I datamaterialet er det ingen personlig tagger som «mystuff» eller «toread». Det er enkelte kryptiske tagger i datamaterialet som antakeligvis er for intern eller personlig bruk. Det er ett spesielt tilfelle med «abb» (#31_09_10). Taggen er tvetydig. Taggen er i orden i henhold til taggereglen om å bruke kjente navn på hendelser. Men 'abb' for konsekvenser i nett-TV. Firmaet ABB blir neppe fornøyd av å bli gjenfunnet sammen med saker tilknyttet Anders Behring Breivik. Taggen 'abb' kan derfor stride mot regelen om å velge tagg med omhu. Taggen er benyttet i ett innslag i datamaterialet og der finnes også taggen «anders behring breivik» (#31_09_02). Disse taggene er også registret som synonym, og sammenstillingen av taggene virker oppklarende og fjerner tvetydigheten. Taggen «abb» (#31_09_10) er også et eksempel på hvordan en tagg kan ha flere av egenskapene som undersøkes. Foruten synonymi og homonymi som er nevnt, er taggen også betraktet som en forkortelse.

4.1.3 Taggenes form

Automatiske gjenfinningsteknikker vil kunne løse problemene med at taggene har forskjellig form, men i enkelte tilfeller så er formen avgjørende for betydning. I henhold til anbefalinger i emneordskatalogisering så mister man betydning som ligger i entalls- og flertallsformene.

Taggene «lastebil» (#09_05_02) og «lastebiler» (#22_14_03) utgjør ingen vesensforskjell i betydning, men det hender at entalls og flertallsform utgjør en forskjell.

Svenonius nevner at uansett beskrivelsesmetode så er det å bestemme emnet vanskelig. Dette gjelder også ved bruk av ulike automatiske gjenfinningsteknikker, særlig i forbindelse med

ikke-tekstlig materiale (2000 s. 46-50). I denne studien undersøkes de tekstlige representasjonene, men beskrivelsen av dem byr også på spesielle utfordringer.

4.1.4 Nyord og slang eller sjargong

Tagging kan sikre at nyord blir en del av indekseringsspråket, og det kan være fornuftig i et raskt skiftende nyhetsbilde.

Taggen «abortskip» (#04_14_01) og taggen «valgbudsjett» (#08_03_04) har det til felles at man kan anta hva de dreier seg om. Men har en søker, intern eller ekstern, fantasi nok til å søke på disse taggene? Det er ingen god søketerm. Innslaget med «valgbudsjett» har også «statsbudsjett» som tagg. Disse taggene er betraktet som synonymet. Denne taggingen forklarer taggen 'valgbudsjett', noen også Rubrikk gjør i samme innslag.

Ved to innslag er taggen «demo» brukt (#22_11_02 og #26_11_01). I tillegg er «demo» også benyttet i en tagg som er en frase (#28_15_05). Dette sikter antakeligvis til 'demonstrasjon'. Taggen «demonstrasjon» er brukt i fem andre innslag og i en frase, samt at det er to forekomster av taggen «demonstrasjoner». I tillegg er taggen «demonstranter» brukt i ett tilfelle. Denne praksisen samler ikke innslag på en entydig måte.

Taggen «justis» (#07_07_03 og #11_13_04) er benyttet ved to innslag fra forskjellige sendinger. Slike forkortede ord som 'justis' og 'demo' vil med gode gjenfinningsteknikker bli funnet, men praksisen går på bekostning av spesifisiteten i indekseringen og presisjonen i gjenfinningen.

4.1.5 Nivåer

Taggereglene oppfordrer til å beskrive presist og generelt (Bakke 2012). Denne praksisen blir fulgt. Et utvalg av datamaterialet viser noen av disse;

- «barnehage» (#10_12_01) er overorda «barnehageplass» (#10_12_04)
- «vold» (#11_12_03) er overordna «seksuelle overgrep» (#11_12_02) og «incest» (#11_12_04)
- «militæreffekter» (#15_15_08) er overordna «militæruniformer» (#15_15_07)
- «samferdsel» (#20_12_03) er overordna «kollektivtransport» (#20_12_01) og «biltrafikk» (#20_12_07)
- «kriminalitet» (#21_02_01) er overordna «drap» (#21_02_04) og «vold» (#21_02_05)
- «avis» (#26_17_05) er overordna «papiravis» (#26_17_06)
- «bombe» (#31_09_01) er overordna «terrorbombe» (#31_09_06)

Det forekommer at tagger til samme innslag omfatter flere nivåer, og nivåene i eksemplene er hierarkiske. Flere nivåer i beskrivelsen sikrer fullstendigheten siden systemet ikke har noen form for kontrollert vokabular, men for presisjonen i gjenfinningen blir denne praksisen et problem og det det grenser til støy. Men det står ikke eksplisitt at dette gjelder hierarkier.

I oversikten på forrige side forekommer taggen «vold» to ganger i forskjellig innslag. Dette viser hvordan taggene sammen utvikler et hierarki med taggene; «kriminalitet» -> «vold» -> «seksuelle overgrep» og «incest». I tillegg til de hierarkiske nivåene ble det funnet flere tilfeller hvor tagger var både oppsplitta ord og hele ordet (mer i kapittel 4.1.8). Disse fenomenene skaper mønstre som kan bli utnyttet i et gjenfinningssystem.

4.1.6 Spesifikke tagger

Taggen «presidentvalg 2012» (#23_07_01) er en spesifikk tagg. Tagger av denne typen, som betegner en sak eller hendelse, motvirker mange innslag på generelle tagger som «presidentvalg» og «2012». En slik indekseringspraksis er innenfor retningslinjene og en større utbredelse vil øke spesifisiteten til taggen. Presidentvalgkampen i USA er hyppig omtalt i Dagsrevyen fra oktober 2012, men det er kun i ett tilfelle denne spesifikke taggen er benyttet. Det er rundt 20 innslag som er relatert til denne presidentvalgkampen. Dette vitner om en lav konsistens i indekseringen.

En forekomst, «islamtrusler» (#27_02_03), er veldig spesifikk. Denne grenser mot hva som er tenkelig at noen vil søke på. Til sammenlikning blir upward-posting anbefalt når emneordene i et kontrollert vokabular blir veldig spesifikke (Hjortsæter 2005).

4.1.7 Synonymer, kvasisynonymi, homonymer og forkortelser

I datamaterialet ble det funnet synonymer, kvasisynonymer, homonymer og forkortelser.

Det ble undersøkt om tagger var synonymer med andre tagger i ett og samme innslag. Det ble funnet 55 par av synonym blant de 1941 taggene. Det totale antallet forekomster blir 110 synonymer siden synonymene henviser begge veier. Det er mindre enn forventet siden retningslinjene, som nevnt, oppfordrer til bruk av synonymer. Vurderingen av synonymi var vanskelig, men det ble forsøkt å stille strenge krav.

Et eksempel som også ble beskrevet i forbindelse med nivåer mellom taggene, ble også betraktet som synonymer; «seksuelle overgrep» (#11_12_02) og «incest» (#11_12_04). Dette var et av tvilstilfellene fordi incest også kan ses på som en type av seksuelle overgrep. Taggene «frafall» (#31_10_05) og «dropper ut» (#31_10_06) er blitt betraktet som

synonymer. Det skjer til tross for at de er fra forskjellig ordklasser, men i det konkrete innslaget er betydningen av taggene den samme fordi de sikter til det samme på idéplanet.

Det var også tilfeller av synonymi som vanligvis er mer utradisjonell i kontrollerte vokabularer. Taggen «gir seg» (#15_24_04) og taggen «slutter» (#15_24_03) opptrer i samme innslag, og de ble betraktet som synonymer.

I datamaterialet var det også synonymer som ble avgjort uten store overveielser, som taggen «sms» (#30_06_02) og taggen «tekstmeldinger» (#30_06_03).

Synonymien er ikke formalisert eller markert på noen vis i systemet, og det gjør det vanskelig å utnytte dette i gjenfinningssammenheng. Men synonymer øker antall verbale innganger til sitt innslag. Dessuten kan synonymer i gitte tilfeller klargjøre taggenes betydning i en relevansvurdering som er tilfelle med taggene «slutter» og «dropper ut».

Ved indeksering betyr synonymer merarbeid, fordi det blir flere tagger å skrive. Det kan være grunnen til at synonymi er mindre utbredt enn antatt. Søbak finner svært varierende holdninger blant informantene når de vurderer terminologiske utfordringer til taggene, deriblant synonymi og flertydighet. En informant er veldig observant på å finne synonymer og splitte opp tagger, mens to andre var mest opptatt av skilletegnet mellom taggene (2013 s.74).

Det ble funnet 11 tilfeller av kvasisynonymi. I innslagene ble det sett etter motsetninger som ble koplet til hverandre. Et eksempel på kvasisynonymi er hentet fra et innslag om at forlagsbransjens ønske om fastpris på bøker. Det er tagget med «fripris» (#10_16_03) og «fastpris» («10_16_04).

I ett innslag ble tre tagger vurdert til kvasisynonymi, fordi det også var benyttet synonymi mellom to av taggene. Innslaget er om frafall i den videregående skolen, som er nevnt tidligere. Taggene «frafall» (#31_10_05) og «dropper ut» (#31_10_06) ble begge vurdert til å være kvasisynonymer til taggen «fullfører» (#31_10_07). Et annet innslag om samme tema hadde taggene «lærer» (#31_13_10» og «elever» (#31_13_11)

Bruk av kvasisynonymer kan gi grundighet i indekseringen hvis de faktisk sier noe om innholdet. Men i tilfeller hvor kvasisynonymer blir tagget fordi taggen er en motsetning til innslaget innhold, så medfører denne indekseringspraksisen overfylte klasser som kan gi støy.

Om man ser taggene med konteksten som i systemet «Metadatavisning intern» er tilfellene av homonymi relativt få, men uten disse opplysningen er det mange potensielle homonymer. Dette forholdet må man ta i betraktning når man lager tagger.

Homonymi er tilstede «overalt» fordi taggene bærer preg av å være postkoordinert og fasetert. Det fører til mange unødvendige tilfeller av homonymi. Taggene «ullevål» «master», «berg», «tønsberg», «hammer» og «brann» er eksempler på tvetydige tagger. Konteksten gjør eksempelvis at taggen «hammer» (#19_05_02) får mening som et etternavn og ikke en sang av Bob Marley eller et verktøy.

Et annet eksempel på homonymi er taggen «estelle» (#21_08_01 og 20_02_05) som henviser til navnet på båten som deltok i aksjonen Ship to Gaza. Dette ser man ut fra konteksten. Men tronarvingen i Sverige heter også Estelle. Hun har riktig nok flere navn, men det er ingen garanti for å bli presist tagget. Kong Harald er eksempelvis tagget «harald» (#10_17_06).

Det er registrert 34 forekomster av forkortelser som tagger. De fleste av dem er allment kjent som «fn», «eu», «vif» «frp», «ap» og «lo». Men ett tilfelle skiller seg ut. Forkortelsen 'dnt' er brukt og det er en kjent forkortelse for Den norske turistforening. I det tilfellet taggen «dnt» (#30_28_06) sikter den til Det norske travselskap som blir delvis forklart av tittelen. Men det er ingen egen tagg som forklare dette. Forkortelser er en av de tilfellene som skaper tvetydighet i en folksomomi (Spiteri 2007 ifølge Yi & Chan 2009).

4.1.8 Orddeling og oppsplitta tagger

Noen av taggene i ett og samme innslag, kan være et resultat av en annen oppsplitta tagg.

- Taggen «nobels fredspris» (#13_04_04) er oppsplitta i taggene «nobel» (#13_04_02) og «fredspris» (#13_04_03).
- Taggen «abortklinikk» (#21_16_03) er oppsplitta i taggene «abort» (#21_16_01) og «klinikk» (#21_16_02)
- Taggen «oslopolitiet» (#27_18_05) er oppsplitta i taggene «oslo» (#27_18_01) og politi (#27_18_02)

En slik praksis vil kunne overfylle og skape støy ved søk.

4.2 Tagger sett i lys av konteksten

4.2.1 Overlappende tagger Helt lik delvis likhet

I typologien er «helt lik» og «delvis lik» to begreper som handler om ordlikhet mellom feltene Tagger og Rubrikk og/eller Tittel. Disse begrepene sjeler til ideen om ordlikhet inspirert av Voorbij. En tagg er «helt lik» hvis utrykkene er eksakt like eller består av en bøyningsform av hverandre. Taggene er «delvis lik» bare hvis bestanddeler av ordet er eksakt likt. Dette kan illustreres med at synonymi ser på matching av begrep eller selve ideen, men bruker ulike termer på verbalplanet. Da ordlikheten matching av uttrykk på verbalplan. Ordlikheten undersøker på verbalplanet, og undersøkelse av synonymien angår begrepsplanet.

Fordeling av taggene på «Helt lik» og «Delvis lik»		
	TITTEL	RUBRIKK
TAGGER «HELT LIK»	412	713
TAGGER «DELVIS LIK»	130	245

Tabell 7 Antall tilfeller av "Helt lik" og "Delvis lik" for Tagger, Tittel og Rubrikk

Taggenes fordeling på «Helt lik» og «Delvis lik» jamfør Nivå 1, er vist på Tabell 7. Taggene kan ha flere av egenskapene, for grupperingene er ikke gjensidig utelukkende. Det er et høyere antall for både «Helt lik» og «Delvis lik» i Rubrikk. Dette har en naturlig årsak, siden Rubrikk er en lengre løpende tekst enn Tittel, og det er da normalt at Rubrikk i større grad overlapper med taggene. Jamfør Lancaster som skriver at jo lengere en representasjon er, jo flere verbale innganger er det (2003 s. 7-9). Den løpende teksten gir flere muligheter for ordlikhet med taggene, «helt lik» eller «delvis lik».

Det er 888 tagger som ikke er kategorisert som «Helt lik» eller «Delvis lik». Disse taggene er kategorisert som «Ulike» jamfør Nivå 7 i typologien for relasjonene. Tagger som er «Ulike» De vil si at de ikke opptre i innholdsbeskrivelsene i ett og samme innslag som taggen.

Omfanget av «Delvis lik» skaper et slags hierarki. Ut fra denne studien er det vanskelig å si om dette er ordsplitting som også kunne blitt formalisert i henhold til Hjortsæters anbefalinger som tidligere nevnt.

4.2.2 Orddeling og oppsplitta ord

I denne studien blir «Oppsplitta termer» registrert separat som en variant av «Delvis lik». Når man tilordner frie nøkkelord uten støtte fra et kontrollert vokabular, så vil de oppsplitta termene kunne utgjøre en forskjell. Dette kan gjøre at taggene fraviker spesifisitetsprinsippet og gir mulige overfylte klasser, eller tagger. Denne problemstillingen er aktuell selv om norsk

språk ofte har sammensatte ord for sammensatte begreper. Det ser nemlig ut til at taggene splittes uansett om det sammensatt begrepet er flere ord eller ett.

Når et ord blir delt opp til flere tagger, så får taggene en fasetert struktur. Det er mange eksempler på at sammensatte begreper blitt uttrykt i oppsplitta ord. Det gjør at det ikke alltid er taggen som er mest spesifikke uttrykket, jamfør spesifisitetsprinsippet. Det hender at taggene er overordna, mens Tittel og Rubrikk har begrenser betydningen av begrepet med bruk av adjektiver, som vist på Tabell 8 og Tabell 9 under. I datamaterialet er det også tilfeller hvor taggene ikke er oppsplitta og inneholder adjektiver som «utenlandske fanger» (#29_19_07). I tabellen under er til sammenlikning taggene «arbeidere» og «utenlandske» oppsplitta. I datamaterialet er det også noen forekomster av sammensatte tagger som «ny bok» (#24_19_03). I dette tilfelle ville man normalt anbefalt dele eller omformulere uttrykket.

ID	TAGG	TITTEL	RUBRIKK
07_12_01	arbeidere	utenlandske arbeidere	utenlandske arbeidstakere
07_12_02	utenlandske	utenlandske arbeidere	utenlandske arbeidstakere

Tabell 8 Eksempler på oppsplitta tagger

ID	TAGG	TITTEL	RUBRIKK
30_17_02	vogntog	utenlandske vogntog	-
30_17_03	utenlandske	utenlandske vogntog	-

Tabell 9 Eksempler på oppsplitta tagger

Et fenomen er at ord forekommer som sammensatte uttrykk i Rubrikk og Tittel, men som oppsplitta i Tagger. I undersøkelsen av sammensatte uttrykk er ord som kommer etter hverandre i løpende tekst betraktet som sammensatte. Tabell 10 viser et eksempel på to tagger som var «Delvis lik» Tittel.

ID	TAGG	TITTEL	RUBRIKK
13_10_02	kronjuveler	kronjuveltyveri	-
13_10_03	tyveri	kronjuveltyveri	-

Tabell 10 Eksempel på sammensatt utrykk i Tittel og oppsplitta tagger

Denne indekseringen som ikke følger spesifisitetsprinsippet, vil sørge for at taggene blir overfylt innslag med en vag tilknytning til emnet. I en søke situasjon vil det medføre til støy og muligheter for feilkoplinger. Hjortsæter trekker inne feilkoplinger i forbindelse med grundighet og dypindeksering hvor treff stemmer med søkeprofil, men ikke faglig sett har relevans for brukeren (2005 s. 25). Det er noe tilsvarende som skjer her når det er en mer eller mindre relevant tilknytning. Ved orddeling så mister taggene de avgrensningene som er i det sammensatte begrepet.

Golder og Huberman (2006) forklarer at det er en utfordring å velge et basisnivå for beskrivelsene. De forklarer dette med begrepet «sensemaking» som betyr at mennesker kategoriserer alt utfra egne preferanser, og dette skjer derfor på ulike basisnivåer. Alle har ikke samme forkunnskap og referanserammer og derfor finnes det ikke alltid et naturlig nivå på beskrivelsene. Lakoff (1987) er også inne på hvor komplekst menneskers kategorisering er.

I datamaterialet er det funnet tagger som er kategorisert som «Delvis lik», men som til sammen dekker et sammensatte uttrykk i Tittel. Det er også funnet samme fenomen i Rubrikk.

4.2.3 Relasjoner mellom Tagger, Rubrikk og Tittel

Fenomenet med orddeling blant taggene, som er beskrevet over, viser at det er mulige relasjoner mellom metadatafeltene som er gjenstand for undersøkelse. Men det er funnet få relasjoner som ville blitt formalisert i henholdt til anbefalingene hos Hjortsæter. Den relasjonen som forekommer oftest er den assosiative. De vil i en tesaurus bli formalisert med sideordna henvisning.

Tabell 11 viser hvordan ordet «mishandling» i Rubrikk har en relasjon til et utvalg av taggene til innslaget. Tittelen for innslaget «Regjeringa vil endre loven for å styrke barns rettigheter» har derimot ingen relasjoner til disse taggene.

ID	TAGG	TITTEL	RUBRIKK
11_12_03	vold	-	mishandling
11_12_02	seksuelle overgrep	-	mishandling
11_12_04	incest	-	mishandling

Tabell 11 Eksempel på relasjoner

Noen få relasjoner ble betraktet som eksempelrelasjoner. Taggen «trener» (#15_24_02) ble funnet «Delvis lik» Vålerengatrener i Rubrikk.

ID	TAGG	TITTEL	RUBRIKK
15_24_02	trener		vålerengatrener

Tabell 12 Eksempel på relasjon

I mange tilfeller ser det ut som om taggene er en rekke med assosiasjoner. Et eksempel er et innslag om utfordringene utrykningspersonell har når de skal finne fram i fjellet. Samtlige tagger til innslag #05_15 er: «redningsmannskaper», «amk», «luftambulanse», «sykebil», «hytte», «nød», «sjusjøen», «brann». Et annet tilfelle er i innslag #09_15 om Egil Olsen som har blitt frastjålet lommeboka si hvor et utvalg av taggene er slik; «ran», «lommetyver»», «robbet», «lommebok». Mens en nyhetssak (25_19) om vær er tagget slik; «uvær», «polart lavtrykk» og «dritvær».

Et innslag (#07_03) med Tittel «Opposisjonen kritisk til samferdselsbudsjett» bærer også preg av en assosiasjonsrekke med taggene «samferdsel, statsbudsjett, tog, vei, utbygging, hoksrud, hareide, opposisjonen».

4.2.3.1 *Nivåforskjeller spesifisitet*

Det forekommer at det er forskjell på spesifisitetsnivået mellom Tagger og Rubrikk og Tittel. Dette i seg selv er ikke spesielt, men det ser ikke ut til å være noen etablert mønster på hvilket felt som er mest spesifikt. Spesifisitetsprinsippet er klar på at emneord skal være uttrykt så spesifikt som mulig (Lancaster 2003 s. 33-35). Der er en liten overvekt av at taggen er mest spesifikt uttrykt, men datamaterialet er for lite til å trekke en bombastisk slutning.

4.2.4 Flere ord i taggene

Frasene blir er sett på som sammensatte termer, pre-koordinert eller frase-setninger.

Enkelte tagger er blitt kategorisert som fraser. Med fraser menes her tagger som inneholder flere ord, men som ikke er sammensatte begrep eller uttrykk. Det er 61 forekomster i materialet. Flere av skrivefeilene hadde klare likhetstrekk. De ble delt inn i undergruppene; pre-koordinert, bildebeskrivelser, brukerfeil i systemet, hendelse, saksopplysninger eller hendelsesforløp, og rene tastefeil.

Eksempel på en tagg som ikke blir regnet for å være et sammensatt uttrykk og derfor ble registrert som frase er «arbeidsformidling malmø» (#21_15_07). Det er også andre tilfeller av frase-tagger som har et pre-koordinert preg, det betyr at de har tagg som avgrenser betydning som «antoine sollacaro drept» (#17_07_01). Dette er også eksempel setninger blant taggene. Innslag #09_13 har enda flere pre-koordinerte frase-tagger; «solberg vingård», «solberg kjører ut», «solberg utfor».

I flere tilfeller er frasene klart bildebeskrivende. Det var seks tilfeller hvor taggen kunne svare til en bildebeskrivelse av innslaget. Eksempler på bildebeskrivelser er vist på Tabell 13.

ID	TAGG
16_18_01	gudmund skjeldal ved minnestatue
16_18_02	gudmund skjeldal i nasjonalbiblioteket
16_18_03	ting som nordahl grieg har eid
25_04_01	diverse bilder av skadde dyr

Tabell 13 Utvalg av tagger som er bildebeskrivende

De bildebeskrivende taggene er alle «Ulike» de andre feltene. Dette fenomenet med så deskriptive beskrivelser kan vise et behov for bildebeskrivelse i gitte tilfeller. Det har ikke

vært vanlig indekseringspraksis siden i 2011 (Engan 18. april 2013). Programbanken har som nevnt, feltet Beskrivelse som kan brukes til dette formålet.

Taggen «bilbelte fritak legeerklæring dødsulykke» (#16_16_01) og taggen «usa valg romney obama» (#04_05_01) kan virke som brukerfeil. Taggene er ikke blitt adskilt med komma ved indekseringen. Det er funnet minst ni forekomster av denne typen.

Det ligger sikkert også en brukerfeil bak taggen «vm-kval fotball polen england wayne rooney kamil glik avlyst håpløse forhold» (#17_23_01). Om det eksemplet hadde vært individuelle tagger ville taggen «håpløse» vært relativt meningsløs, i en gjenfinningssituasjon. Derimot er «håpløse» veldig beskrivende og avklarende i den konteksten det står. Taggene har to funksjoner. De skal funke i søk, men også hjelpe med relevansvurdering. Dette er også en forekomst som viser objektiv og subjektiv emnebeskrivelse.

Taggen «beretningen om rikets tilstand» (#02_06_04) sikter til en årlig hendelse. Det kan være vanskelig å beskrive dette på en annen måte enn nettopp slik. Derimot er taggen «castro lever» (#22_11_04) en fullverdig setning, og som beskrivelse for innslaget er den sikkert i overensstemmelse. Men i ettertid vil den ha behov for en nærmere forklaring og derfor er den mindre god som tagg i et lengre tidsperspektiv.

ID	TAGG
22_03_01	sedelighetssaker tar lang tid å etterforske
22_03_02	hastemøte justisdepartementet
22_03_04	konkrete tiltak fra riksadvoakten

Tabell 14 Utvalg av tagger som er saks- eller hendelsesbeskrivende fraser

I enkelte tilfeller refererer taggene til en beskrivelse av saksforhold eller hendelsesforløp på en annen måte. Tabellen over viser et utvalg av tagging som med stor sannsynlighet viser til saksforhold i et innslag. Disse er ikke i henhold til taggereglene, men det kan være at disse saksforholdene er vanskelig å forklare. Det å bestemme emne og uttrykke det i et par ord er vanskelig (Chowdhury 2010 s. 97).

Tagger som beskriver hendelsesforløp er også funnet. Taggen «lek med våpen» (#22_07_03; #22_08_01) er antakelig beskrivelse av foranledningen til et drap. En sak hvor en mann skal ha bedt en jente sette seg inn i en bil er tagget «lokket inn i bil» (#09_08_03) som sikter til hendelsesforløpet. Taggen «eu reaksjoner på nobels fredspris» (#12_05_01) er nok et tilfelle på en beskrivelse av hva som konkret skjer i innslaget, men måten «hva» er beskrevet på er ikke i form av tagger jamfør *Taggeregler i NRK* (Bakke 2012).

Det er flere tilfeller hvor verb er tagger også etterfulgt av objekt, preposisjon eller annet. Taggen «gjemmer unna» (#25_11_04) antyder hva folk har gjort med pengene som Norges Bank ikke har mottatt til destruering.

Fire innslag med tilsvarende foranledning har veldig ulike tagger. Taggen «gir seg» (#15_24_04) refererer til at en fotballtrener slutter, og landslagskeeper Jon Knudsen får taggen «legger opp» (#24_26_02). Mens Øystein Mæland seg trekker seg som politidirektør, og innslaget er tagget «fratre» (#08_02_04). Et innslag om frafall i videregående skole har taggene «frafall» (#31_10_05) og «dropper ut» (#31_10_06).

En slik praksis er intet godt tegn for framtidig gjenfinning, fordi det stiller store krav til oppfinnsomhet hos søkeren. Liknende saker blir ikke samlet ved tagging. Dette vitner kanskje om at disse sakene har ulike uttrykksmåter som kan stamme fra saksfeltet.

Et fotballag blir trukket poeng i Eliteserien på grunn av økonomisk rot. Rubrikk skriver «Det kan sende laget ute av gullkampen.» En av taggene er «gullet ryker» (#05_19_05). Dette er en metaforisk beskrivelse av et saksforhold.

Det er lett å forstå hva slike frase-tagger egentlig beskriver, men uansett målgruppe så kan man si at valg av frase-tagger gjør søk og gjenfinning vanskelig. Denne indekseringen er kanskje verken spørsmåls- eller begrepsorientert, men mer en privat notis. En samlet gjenfinning av innslagene med denne typen tagger er vanskelig å forestille seg. Mathes (2004) trekker fram tagger som gode til nettopp browsing (s. 6). Det er klart at slike tagger vanskeliggjør browsing. NRK-taggene er ikke lenket sammen, men om taggene skal bli det bør man se på automatiske løsninger, som kan samle taggene med for eksempel syntetiske synonymer. Dette vil ikke løse problemet med frase-taggene, men kanskje lette gjenfinningen.

Det er tre tilfeller av tagger som viser til titler. Taggen «drømmen om amerika» (#01_16_08) er tittelen til en bok, taggen «tomorrow nevner dies» (#05_17_02) er tittel på en film, og «torsdag kveld fra nydalen» (#06_10_01) er tittel på et TV-program. Det er vanskelig å beskrive disse taggene på noen annen måte enn med deres tittel. Tittelen kan være den rette representasjonen av innslaget og således viktige i både indeksering og gjenfinning. Det er derfor et poeng at tagger i frase, bør være reservert de tilfellene hvor det er nødvendig.

4.2.5 Skrivefeil i tagger

At det forekommer skrivefeil er opplagt. Det er registrert 44 tagger med skrivefeil. Dette er mindre enn forventet.

Det er to tagger som er medregnet til tross for at de ikke er egentlige skrivefeil. Men de er feil i henhold til taggeregelen om bokmålsform. Dette gjelder Djurpolisen (#29_16_03) som er svensk og Vedlikehald (#10_15_02) som er nynorsk. I begge tilfellene har Rubrikk ordene rett stavet med «dyrepoliti» og «vedlikehold».

De aller fleste skrivefeil i datamaterialet opptrer som blant taggene som er ULIKE sett i forhold til Tittel og Rubrikk (jamfør nivå7 i typologi). Det vil si at taggene ikke blir gjentatt i Tittel eller Rubrikk, men det er også tilfeller hvor skrivefeilene forplanter seg (omtalt i 4.2.6).

Enkelte av skrivefeilene er vurdert som regelrette tastefeil. Bokstaver som har byttet plass «persidentvalg» (#28_14_02) eller «antidpoing» (#27_15_05). Ord som inneholder bokstaven ved siden av på tastaturet «efts-domstol» (#08_06_03). Lengre tastetrykk som har gitt ekstra bokstaver «unngdom» (#27_15_02) eller omkringliggende bokstaver «vinterstporm» (#25_19_02).

Utbredelsen av skrivefeil var mindre enn forventet, og det harmoniserer med Søbak (2013) som også fant færre skrivefeil en forventet (s 74-75). Men skrivefeilene er alvorlige nok, særlig med tanke på at taggingen hos NRK ikke blir eliminert av andres tagging som i en folksonomi. Taggene forblir stående som representanter for innholdet med skrivefeil. Hos NRK er det ingen «the wisdom of the crowd»-effekt.

4.2.6 Skrivefeil i tagger og forholdet til konteksten

Skrivefeil havner ofte som nevnt over i kategorien ULIKE jamfør typologien inspirert av Voorbij. Dette viser at skrivefeilene ikke blir forsterket av Rubrikk og Tittel, om man ser for seg et vektingsprinsipp ved gjenfinning.

Men det er også eksempel på at skrivefeil ikke rettes opp av Tittel og/eller Rubrikk, og skrivefeil forplanter seg. Når skrivefeil blir gjentatt i Tittel og Rubrikk, blir det vanskelig å se for seg at feilen kan elimineres ved søk. Dette gjelder; barnevektsstudien (to tilfeller #29_20_01 og #29_07_06), aleksander kristoff (#23_27_01), rettsak (#23_12_03), guiness (#09_10_08), efts-domstol (#08_06_03). I forbindelse med «Øygaard-saken» er det tilfeller der Feltet Tagger, Tittel og Rubrikk har samme skrivefeil.

Feilstavete Tagger som er skrevet riktig i Tittel eller Rubrikk får flere verbale innganger. Ved ett eksempel forekommer det skrivefeil i tagg og overskrift «trond birkeland» (#17_04_01), mens navnet er stavet rett i Rubrikk (‘trond birkedal’). Beskrivelse i flere metadatafelt kan dermed eliminere en skrivefeil. ?

4.2.7 Homonymi

Det er tidligere skrevet om problematikken rundt tvetydighet med hensyn til taggene (se kapittel 4.1.6). Siden taggene ikke alltid er presise eller spesifikke, skaper dette problemer når generelle termer blir benyttet i ulike sammenhenger. Taggen «brann» forekommer ved tre innslag. To av dem er i forbindelse med at det brenner ulike steder, mens den siste forekomsten dekker Sportsklubben Brann. Det er flere innslag som antyder at det handler om brann på et mer spesifikt nivå; «husbrann» (#27_08_01) og «brannslokking» (#27_08_02). Det er også en tagg av pre-koordinert karakter «brann sykehus» (#23_16_04).

Et annet tilfelle hvor taggene mangler spesifisitet og blir tvetydig er taggen «ullevål». Taggen er benyttet ved to innslag, den ene sikter til sykehuset (#28_07_05) og den andre sikter til stedet (#13_13_04).

4.2.8 Persongalleriet

Feltene Medvirkende og Tagger har blitt undersøkt for å se på persongalleriet. Medvirkende og Tagger sier noe om hvordan personer i innslaget blir betraktet. Medvirkende skal angi hvem som deltar i innslaget. Personer skal bli tagget når de er sentralt i innholdet, i henhold til taggereglene. Personer er tagget i flere innslag og i henhold til retningslinjene skal de da være vesentlig for innholdet. Det er ikke nok å være synlig i tv-ruta. Det er mistanke om at ikke alle er så vesentlig for innholdet, at de egentlig ikke skulle vært tagget.

Kategoriseringen av taggene oppdaget både virkelige («lance armstrong» #22_02_01) og fiktive personer («james bond» #05_17_01).

Det er flere innslag om sport og især fotball som ramser opp etternavn på mange av spillerne. Dette er antakelig gjort fordi de er med, men det er uvisst i hvilken rolle. Et eksempel er innslag #02_20 «utenlandsproffer», «landslaget», «moa», «tettey», «demidov», «høgli» og «braaten». En slik bruk av etternavn som tagg er utbredt.

Det ble ikke funnet noe systematikk i hvordan feltet Medvirkende og Tagger med personnavn. Det ble oppdaget at en del av Medvirkende feltene var tomme til tross for at de er obligatoriske i henhold til Metadatastandarden.

Det kan altså være personer er registrert dobbelt, uten å være sakens egentlige kjerne. En tagg skal antyde at personen er vesentlig for innholdet, men etter nærlesning av alle innslagene i datamaterialet, så er det sannsynlig at mange av taggene er i strid med reglene. Denne praksisen medfører at persongalleriet blir utydelig. Når personer er tagget uten å være viktig for innholdet, så fører dette til mye støy.

NRK har heller ikke et autorisasjonsregister for personer, så det blir viktig å stave personnavn korrekt. Taggen «trond birkeland» (#17_04_01) er stavet feil, men i feltet Medvirkende er navnet stavet rett 'trond birkedal'.

5 Oppsummering og diskusjon

Taggene er alt mulig; enkelt ord og begreper, kjente navn på hendelser, bildebeskrivelser, saksopplysninger, setninger. Taggenes betydning og syntaks er også mangfoldig. En faseterte tilnærming som gir relativt generelle tagger og fenomenet med oppsplitta tagger gjør at det skjer en opphopning av innslag på generelle termer. Med en slik praksis skulle en tro at flere innslag for de samme taggene. Hovedtendensen i datamaterialet er at litt over halvparten av taggene bare er benyttet en gang. Det gjeler 1095 av 1941 tagger som tilsvarer er 56 %. Det er 846 tagger som er benyttet flere ganger altså 44 %. Tagger som er brukt fire ganger eller flere tilsvarer 18%. Power law-prinsippet gjør det vanskelig å finne krystallklare tendenser for mange av problemstillingene i studien

Det er opplagt at en taggepraksis som NRK har, fører til større spredning enn et kontrollert vokabular. Det ligger i sakens natur at kontrollert vokabular ved å samle om emnet og derfor sprer andre emner. Taggene skal, forenklet sagt, beskrive innslagets «hva». Hos NRK mangler kontrollen over vokabularet og da vil noen beskrivelser av «hva» bli uttrykt slik at liknende innhold ikke samles, men blir spredt.

Det er forskjell på tagging og de andre beskrivelsesmetodene for kontrollert vokabular omtalt i kapittel 2.2. Chowdhury skriver at selv om presentasjonen av innholdet varierer, så er selve den konseptuelle analysen den samme, uansett system (2010 s. 77). Det er i denne analysefasen våre ulike ‘mentale konsepter’ og ‘begrepsmessige systemer’ bidrar til vår forståelse av en ting, jamfør Lakoff i kapittel 2.2.1.

5.1 Uenighet om hva som er tagg

I løpet av prosjektet har flere ansatte i NRK hjulpet med opplysning, veiledning og opplæring, og det har vist seg at det er uenighet innad i NRK om hva en tagg dekker.

De fleste oppgir at taggene skal supplere programomtalen og gi merverdi. *Tags i NRK* (Bakke & Fleischer 2011) skriver at tagger er et tillegg til de andre metadataene (s. 3).

Taggene skal ifølge Holgersen (foredrag KORG-dagene 2013 8. februar 2013) dekke det journalistiske «hvem, hva, hvor». De nye retningslinjene *Taggeregler i NRK* (Bakke 2012) gir at tagger skal dekke «hva». Men i tillegg *kan* taggene dekke «hvem», «hvor» eller «når» hvis det er sentralt i innholdet. Tagger blir eksemplifisert med «askesky el. trafikkstøy el. trålefiske el. samtidskunst» og ikke utdypet videre. Wettmark (samtale 19. mars 2013) bekrefter

uenigheten og understreker at det ikke er eksplisitt definert hva taggene skal dekke. Dette fører til tolkningsmuligheter for de som tagger.

Rapporten *Evaluering av metadataregistrering i NRK* (Tremoen et al. 2013) bekrefter også dette. Medarbeiderne er usikre på hva som er en god tagg. NRK mangler verktøy for å avhjelpe usikkerheten i indeksering og tagging (s. 10). Denne rapporten er basert på spørreundersøkelse, test-søk og en supplerende spørreundersøkelse til fem utvalgte som driver research i utstrakt grad. Men rapporten konkluderer med at de sju metadatakravene er tilstrekkelig for gjenfinning, men det forutsetter også at metadataføringen prioriteres i redaksjonene (s. 3). Rapporten beskriver samtidig taggingen som tilfeldig og mangelfull (2013 s. 10).

5.2 Endringer i arkiveringspraksis

Endringer av indekseringspraksis bør skje fra tid til annen. Fra 21. mai 2013 ble taggereglene på NRK sine intranettsider endret. Den nye veilederen er redusert fra 10 til 5 punkter. Endringene handler i hovedsak om å forklare enklere og bruke færre eksempler. De nye reglene sløyfet rådene om å unngå homonymi og om å legge til synonymer (Liu 2013). Endringer påvirker ikke mitt datamateriale siden NRK ikke endrer beskrivelsene retrospektivt i allerede arkivert materiale (Engan opplæring 24. mai 2013; Wettmark møte 19. april 2013). Men disse nye reglene fjerner to viktige komponenter. Homonymi og synonymi er to faktorer, som om de blir brukt, kan hjelpe både presisjon og fullstendighet.

I samme periode ble det også lansert skriftlige anbefalinger om innholdet i feltene Overskrift og Rubrikk (Liu samtale 24. mai 2013). Disse anbefalingene sammenfaller med praksisen som NRK-medarbeiderne allerede har i nyhetsinnslag.

Søbak (2013) undersøker alle taggene i Programbanken i løpet av én uke, og hun beskriver taggepraksisen som lite konsistent, som tidligere omtalt. Chowdhury sier at det er vanskelig å oppsummere emnet i få ord, og uansett hvilke retningslinjer man har så er det en risiko for ulik beskrivelse (2010 s. 97). Vanskelighetsgraden og ulikhetene blir kanskje større, når reglene er for enkle og tvetydige forklart. Nye indekseringspraksiser kan medføre endringer, som får konsekvenser for indekseringspraksisen. Men om det stemmer som Søbak (2013) beskriver, at de ansatte har vanskelig for å fylle rollen som metadataprodusent, så vil endrede regler antakelig ikke styrke motivasjonen og ei heller konsistensen i indekseringspraksisen.

Smith hevder at noen mener vanlige metadataproblemer ikke bare skyldes klassifikasjonssystemer, men også selve metadataene. Smith retter kritikken mot de som utarbeider metadataene, som per i dag er hvem som helst. Han mener «vi» er late, dumme og uærlige (Smith 2008 s. 87). Han refererer til et essay av Doctorow.

Doctorow beskriver flere hindringer mellom dagens situasjon og det han kaller et meta-utopia, hvor det er pålitelige metadata. Disse hindringen – løgn, latskap, sløvhet, uærlighet– finnes i oss alle. Det er flere menneskelige trekk han understreker. Vi har vanskelig for å innordne oss et forhåndsbestemt system. Vi har vår egen oppfatning av informasjonen og derfor følger vi ikke fastsatte reglene. Alt har flere perspektiver, og vi vil gjerne uttrykke vårt syn. Resultatet vi får er at alle er like misfornøyd med situasjonen. Doctorow skriver at Meta-utopia aldri vil bli en realitet, og metadata må tas med en klype salt og sees på som antakelser og ikke fakta (Doctorow 2001).

Tagger kan sikre at nyord blir en del av indekseringsspråket, og det kan være fornuftig i et raskt skiftende nyhetsbilde. Spesifisitet og grundighet i indekseringen legger grunnlaget for fullstendighet og presisjon. En konsistens i indekseringen er viktig i denne sammenhengen.

5.3 Fordeler med innholdsbeskrivelse av nyheter

Indeksering av film og bilder har andre utfordringer enn tekstlig materiale. Denne studien har valgt å se på nyhetssendinger, nettopp fordi det ble antatt at indeksering av nyheter gir mindre rom for tolkning. Nyheter er dokumentariske og handler om saksforhold ved samfunnslivet. NRK oppgir at beskrivelsen i Programbanken skal angi det journalistiske «hvem, hva, hvor» og taggene skal dekke hva. Den nye praksisen gir at beskrivelsene skal være saksorientert og ikke tolkende eller bildebeskrivende. Dette betyr ikke at det er lett å bestemme innslagets «hva», men noe lettere enn ved tv-sendinger som gir rom for mer tolkning av innholdet. Denne studien har sett på potensialet i de foreliggende metadata og har ikke tatt stilling til selve beskrivelsen av emnene og «aboutness» i innslaget.

5.4 Taggepraksisen til NRK

Mitt hovedinntrykk av datamaterialet er at det finnes litt av alt. Undersøkelsene avdekker noen få tilfeller av hver egenskap. I denne sammenhengen er 1941 tagger for lite, men dette kan være et tegn på at indekseringsreglene gir mange ulike tagge-stiler. Enkle taggeregler gir jo rom for tolkning og dette kan dermed gjenspeile seg i beskrivelsen av innslagene. Dette er i alle fall et bevis på at det er uforutsigbart hva man kan søke på.

Taggene har to funksjoner. De skal funke i søk, men også hjelpe med relevansvurdering. Dette er et viktig perspektiv ved vurdering av taggene, fordi det er lett å tenke at taggene skal fungere som verbale innganger. Denne studien startet også med bibliotekariske referanserammer, hvor kontroll på vokabularet blir sett på som viktig. I slike systemer blir relevansvurdering gjerne overlatt til andre kilder, enn selve notasjonen. Organiseringen i klassifikasjonssystem sier ikke nødvendigvis «korrekt» utfra en ordbok-definisjon. Smith påpeker at klassifikasjonssystemer for eksempel har problematiske sider, blant annet kan de stride mot enhver fornuft (2008 s. 87)

I undersøkelsene er taggene sett i den konteksten de er arkivert og det er et perspektiv som ikke eksisterer ved søking. Dette kan påvirke forståelseshorisont og helt sikkert oppfattelse av taggene. Denne studien yter derfor ikke alltid rettferdighet til hvordan taggene fungerer i søk. Dette er viktig å merke seg, om man prøver å dra nytte av resultatene fra undersøkelsene i den praktiske søke-verdenen.

Alle sendinger med innslag har et minimum av verbale innganger fordi det er feltet Tittel som etablerer et innslag i systemet. Det er sannsynlig at Tagger og Rubrikk øker antall verbale innganger, men hva om de ikke tilfører ord som er dekkende for innslagets innhold? Rubrikk skal ifølge *Språktips til gode rubrikker* (Bakke 2013) appellere til hjerte og hjerne og vekke nysgjerrighet og forventning. Det er ikke sikkert Rubrikk gir beskrivelser som gir mening uten å se annen kontekst eller selve innslaget.

Taggene kan være et produkt laget med bakgrunn i Tittel og Rubrikk. Det ble funnet relativt høy forekomst av «helt lik» eller «delvis lik» og få uavhengige tagger. Det er kun 46 % av taggene som skiller seg ut og er «Ulike» teksten i feltene Tittel og Rubrikk. Dette kan skyldes at taggerne lar seg inspirere av disse feltene, istedenfor å beskrive innslagets faktiske innhold.

5.5 Sammenlikningsgrunnlaget

Nærlesning av alle innslagene i datamaterialet ga en fornemmelse av at Tittel, Rubrikk og Tagger, har hver sin sjanger med veldig forskjellig framtoning. Tidligere i teksten ble det skrevet at Tagger skal angi innsalgets «hva»; Tittel inneholder en lokkende og triggende overskrift og at Rubrikk gir en beskrivelse av tv-innslaget, men som oftest inneholder det nyhetsankerens introduksjon i studio. Det viser seg at forskjellene er langt mer enn det. Rubrikk har ofte en lokkende omtale av innslaget som skal friste publikum til å følge med videre. Den skal ifølge *Språktips til gode rubrikker* (Bakke 2011) appellere til intellekt og følelser og skape nysgjerrighet og forventning. Titlene har derimot en veldig varierende form, og av og til så kort at den likner en tagg som «Helse Bergen» (#27_17) eller mer en arbeidstittel som «tommel opp eller ned for budsjett» (#08_05). I tillegg kan Tittel også gi en klar oppfatning av hva et innslag dreier seg om som «Mener støtte til Hurtigruten er i strid med EØS-avalen» (#08_06) eller «Danskene leser like mye etter at de fikk fri priskonkurranse på bøker» (#10_16).

Slike sprikende kvaliteter ved innholdsbeskrivelsene kan være grunnen til at sammenlikning av dem er komplisert. Deres ulike formål gir svært ulikt språk. Dette kan også være årsaken til at det ikke finnes mange hierarkiske relasjoner imellom dem. Undersøkelsene av helt eller delvis match viser på samme tid at taggene likner Tittel og Rubrikk, 54 %.

Innledningsvis ble det trukket opp en parallell mellom den løpende teksten i Rubrikk og faglige sammendrag (abstract). Det er stor forskjell på et sammendrag av en faglig artikkel og Rubrikk, og deres formålet er ganske ulike. Rubrikk skal skape nysgjerrighet og appell, mens et sammendrag skal deskriptivt gjengi et studium. Sammendrag er nevnt som en beskrivelsesmetode fra bibliotek- og informasjonsvitenskapen. Den var ment å illustrere poenget med hvordan ulike metadata kan supplere hverandre, og at en lengre tekst øker antall verbale innganger.

5.6 Indekseringen med tagger

Om NRK praktiserer en spørsmålsorientert eller begrepsorientert indeksering er ikke klart. Taggene skal dekke «hva» ifølge retningslinjene, og samtaler om taggepraksisen med medarbeiderne viser at et vanlig råd for tagging er å tenke på hva man kan søke på. Gjennomgangen av det skriftlige materialet om taggeregimet avdekker at NRK kan prøve å dekke begge indekseringsvariantene i deres policy. Bakgrunnsmaterialet for taggereglene beskriver taggene som innholdsbeskrivende, enkeltord eller begreper, frie og ikke-hierarkiske

og et tillegg til andre metadata (Bakke & Fleischer 2011 s. 3). Taggereglene sier at taggene skal dekke «hva», eventuelt «hvem», «hvor», eller «når» om det er sentralt i innholdet (Bakke 2012 regel 1). Begge disse forklaringene legger opp til en begrepsorientert indeksering. I begrunnelsen for hvorfor man tagger vektlegger man gjenfinning av materialet internt og blant publikum, både i dag og om ti år (Bakke & Fleischer 2011 s. 3). Denne argumentasjonen vitner om et mer spørsmålsorientert syn på indekseringen. I tillegg oppfordrer taggereglene til bruk av synonymer (NRK 2012 regel 7), flere hierarkiske nivåer (NRK 2012 regel 8) og kjente uttrykk som folk flest bruker (NRK 2012 regel 4). Dette er selvfølgelig for å oppveie manglede vokabularkontroll, men kan også være tegn på en spørsmålsorientert tilnærming i indeksering.

En fri tagging basert på enkle regler er ingen garanti for at jobben blir gjort. Søbak påpeker at de ansatte hadde svært ulike holdninger til metadataproduksjon (2013). Men NRK er kanskje nødt til å stramme inn regelverket. Med en større kontroll på vokabularet vil NRK ha muligheter til å oppnå god gjenfinning som satser på fullstendighet og presisjon.

En varierende utførelse av metadataproduksjonen fører til lav indekseringskonsistens. Antakelser om lav indekseringskonsistens samsvarer med Søbak (2013). Bortfall i datamaterialet med 159 innslag uten tagger tyder på dette. I tillegg to sendinger (#14 og #18) metadata og innslag. Det gjør at de ikke har verbale innganger utover det som er på sendingsnivå som programtittel og sjanger. Det er tydelig at Metadatastandarden med sju obligatoriske krav, blir ikke fulgt. 161 innslag manglet utfylling av Tagger og 34 innslag manglet utfylling av Rubrikk. Dette gir en utfyllingsgrad etter denne studiens perspektiv på rundt 70% for oktober 2012.

Topp-taggene viser i grove trekk hvilke saker som hadde bred dekning i oktober 2012. Dette er på sett og vis et tegn på at taggingen reflekterer innslagenes innhold. Men det er 30 % av innslagene som mangler en fullverdig beskrivelse i henhold til Metadatastandarden. Dette forholdet kan selvsagt påvirke topp-taggenes karakter.

NRK sine egne undersøkelser av metadatakvaliteten (Tremoen 2013) konkluderte med at gjenfinningen var god nok. Mine undersøkelsen gir grunn til å være bekymret for enkelte aspekter ved indeksering med tagger. Denne studien har vist manglede konsistens i indekseringen.

5.7 En bedre framtid

Med enkle grep kan likevel NRK skaffe seg ytterligere kontroll over hva taggene egentlig beskriver. Autorisasjonsregister for navn, korporasjoner og sted eller geotags eller regel om at stedsnavnet skal skrives fullt ut som for eksempel Brann stadion, vil bøte på noen tvetydigheter. Det kan være lurt å lage standardiserte måter å angi hendelser som skjer ofte som fotballkamper (hjemmelag-bortelag). Instruks om å bruke offisielle forkortelser og fullstendige navn. Etablering av tagger for saker og hendelser i nyhetsbildet, vil kunne øke fullstendigheten (presidentvalg 2012). Det kan fungere som hashtags som oppstår når en ny sak, og etter kort tid har flere forskjellige brukt den. Mulighetene for tvetydighet i beskrivelsene vil kunne minske. Eksempelvis for en tagg som 'vålerenga'; den vil angi stedet i Oslo, ikke en fotballklubb, ikke en treningsarene heller ikke en kirke. I tillegg bør indekseringen støttes med både ordbok og begrensninger for antall tegn i hver tagg. Det vil avhjelpe skrivefeil og brukerfeil om man skulle glemme komma.

Et mer krevende grep vil være å etablere et 'rett nivå for taggene'. Golder og Huberman skriver at det er vanskelig å etablere et slikt basisnivå for spesifisitet, fordi vi alle har ulik faglig, sosial og kulturell bakgrunn. De argumenterer for at disse faktorene spiller inn ved valg av tagg (2006). Lakoff er den som forkastet kategorisering som «ting med felles egenskaper», og han viste hvor komplisert synet på kategorisering er. Kategorisering er avgjørende i alle syn på resonering (1987 s. 8). De begrepsmessige systemene kalt konseptuelle metaforer (egen oversettelse av conceptual systems), er organisert i kategorier. All tenking involverer disse kategoriene (1987 s. xvii). Til syvende og sist så viser dette hvor komplisert fri tagging egentlig er.

Jamfør Lakoff syn på kategorier og Golder & Huberman tanker om «sensemaking» og basisnivå beskrevet i kapittel 2.2.1. Man kan tenke seg at kategorisering likner tagging og at erfaring og referanserammer spiller inn, slik at taggingen får et personlig tilsnitt. Dette vil sannsynligvis skje når taggingen i liten grad er styrt, i alle fall sammenliknet med kontrollerte vokabularer. I tillegg kan den usikkerheten som er oppdaget blant de ansatte om hva som er en god tagg, spille inn. Dette er forhold som kan forsterke behovet for mer kontroll med indekseringsspråket.

Chowdhury setter indeksering inn i en større kontekst hvor også målet, gjenfinningssystemet, blir en viktig del (2010 s. 100-102). Lancaster deler indekseringen i to hovedprosesser (2003 s. 9; s. 18-19). NRK passer best inn i en to-trinns prosess i synet på indeksering; først en

analyse av ‘hva’ og deretter oversettelsen med tilordning av tagger. Men om man ønsker å etablere en strengere kontroll med indekseringsprosessen, så vil et utvidet syn kunne hjelpe.

6 Videre forskning

Det er interessant om taggen i Programbanken er hentet fra forslagslisten som er basert på tidligere tildelte tagger. Tagger som er valgt fra forslagslista vil ha en understreking i Programbanken (Engan 18. april 2013). I denne studien ble det for krevende å undersøke alle taggene i Programbanken. En interessant problemstilling for videre undersøkelser er hvilken påvirkning en forslagsliste til tagger har. Programbanken har en forslagsliste utfra ordlikhet, når man skriver tagger. Forslagslista er redaksjonelt utarbeidet av Metadataseksjonen basert på tidligere tildelte tagger. Per 26.10.12 var denne lista på 6524 tagger (Wettmark, e-post 26/10-12). En tagg er ett eller flere ord som ved indekseringen blir adskilt med komma. På samme tidspunkt var det totale innholdet av tagger i DIGAS (produksjonssystemet for radio) og Programbanken 52 042 (Wettmark, e-post 26. oktober 2012). En undersøkelse av denne fullstendige taggekorpusen er for øvrig også interessant.

Yi og Chan (2009) viser til forskningsresultater hvor forslagsliste øker bruk av tagger, og hvor forslagslister ikke har særlig innvirkning på valg av tagg. Det er uvisst hvordan dette stiller seg i NRK. Søbak (2013) undersøkte holdningene til metadataproduksjonen blant et utvalg metadata-medarbeidere. De ansatte benytter forslagslista i varierende grad, alt fra aktiv bruk av lista til bevisst ignorering av forslagene (2013 s. 61).

Konsistensproblematikken er veldig utfordrende for NRK. En studie som så på konsistensen i beskrivelsen blant liknende saker eller samme hendelse, ville vært fruktbart for utviklingen til metadataarbeidet i NRK. Ut fra sammen perspektiv vil en studie som ser på konsekvenser med samsøk på tvers av paradigmene sentralisert og desentralisert beskrivelser også være av verdi.

En vurdering av den faktiske aboutness-en og beskrivelsene av innslagene. Dette kan gi kunnskap som kan lage opplæring som øker forståelsen av indekseringens hensikt. I datamaterialet for denne studien var det mange innslag uten metadata. Derfor kan økt forståelse for metadataproduksjonens formål gi spennende og fruktbart forskning.

De interne feltene Beskrivelse og Begivenhet fra Programbanken (PB) er av interesse. Disse feltene kan være innholdsbeskrivende, men de er ikke en del av den nye Metadatastandarden (Tremoen et al. 2013 s. 3). Feltet Beskrivelse ble tidligere brukt til bildebeskrivelse dvs. det som skjer i bildet, hvem som er med, livemusikk, arkivklipp, rettigheter, blant annet. Feltet Begivenhet gir emneord til store nyhetssaker som «22. juli-rettsaken» eller «terroraksjonen

22.07.11». I testsøkene er det flere innslag som har kommentarer i dette feltet. Det minner ifølge Engan, om bildebeskrivelser som arkivarene gjorde i gammel praksis. I ny praksis er det ingen retningslinjer for hva dette feltet skal brukes til. PB har også feltet «Begivenhet» som gir emneord til store nyhetssaker som nevnt over (Engan opplæring og møte 18. april 2013).

En innfallsvinkel til datamaterialet i denne studien som kunne vært studert er de taggene som er «Ulike» Rubrikk og/eller Tittel. Disse taggene kan ha spesielle kjennetegn siden de ikke er representert andre steder i beskrivelsen. De kan være viktige fordi de skiller seg ut. Det kan være de utgjør en merverdi til beskrivelsen slik ønsket er i *Tags i NRK* (Bakke & Fleicher 2011), forarbeidet til taggereglene NRK.

Det hadde vært interessant å undersøke om taggenes karakter endrer seg med frekvensen. Det var et varierende antall tagger til hvert innslag i datamaterialet.

Hva endre mediebedrifter gjør med metadata er spennende, og det kan være en kilde til inspirasjon og nytekning. En annen mediebedrift, Associated press (AP), har utviklet en taksonomi som automatisk tildeler tagger til nyhetsinnhold. AP's News Taxonomy er maskinelt laget med hjelp av teknikkene Machine-learning (maskinlæring) og Natural Language Processing (språkteknologi). Det er utviklet som semantisk teknologi og tilgjengelig gjennom en API. Formålet er å beskrive alt nyhetsinnhold som bilder, bildetekst og artikler for å ivareta en god effektivitet i gjenfinningen, både for presisjon og fullstendighet. Selv om det er arbeidskrevende, så blir taksonomien kontinuerlig vedlikehold, for å sikre at gjenfinningen blir opprettholde. Daglig blir nyhetsbildet kartlagt og effektiviteten sjekket. Andre innen mediebransjen har også satset på tilsvarende taksonomier og teknologi, blant andre Thomson Reuters, New York Times og BBC (Li 2013).

7 Litteratur

Bakke, A.K. (2011). *Språktips til gode rubrikker*. Internt dokument fra 8. november 2011.

Bakke, A.K. (2012). *Taggeregler i NRK*. Intern dokument fra 4. juli 2012.

Bakke, A.K. & Fleicher, N.B. (2011). *Tags i NRK*. Internt dokument fra 15. januar 2011.

Broughton, V. (2004). *Essential classication*. London: Facet.

Chowdhury, G.G. (2010). *Introduction to modern information retrieval* (3. utg.). London: Facet.

Cutter, C.A. (1904). *Rules for a Dictionary Catalog*. Hentet fra https://archive.org/stream/rulesforadictio06cuttgoog#page/n4/mode/2up

Golder, S.A. & Huberman, B.A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32*(2), 198–208. Hentet fra http://jis.sagepub.com/content/32/2/198.full.pdf

Doctorow , C. (August 26, 2001). *Metacrap: Putting the torch to seven straw_men of the meta_utopia*. Hentet fra http://www.well.com/~doctorow/metacrap.htm

Halpin, H., Robu, V., Shepherd, H. (2007). *The complex dynamics of collaborative tagging*. I *WWW 2007, May 8-12, 2007, Banff, Alberta, Canada*. Hentet fra http://www2007.org/papers/paper635.pdf

Hedden, H. (2010). *The accidental-taxonomist*. Medford, New Jersey; Information today inc.

Hjortsæter, E. (2005). *Emneordskatalogisering: Innholdsanalyse, emnerepresentasjon og lagring*. Oslo: Høgskolen i Oslo.

Kipp, M.E.I. (2005). Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator, and Intermediary Keywords. *Canadian Journal of Information and Library Science 29*(4), 419-436.

Kipp, M.E.I. (2011). Controlled vocabularies and tags: An analysis of research methods. I Smiraglia, R.P. (Red.), *North American Symposium on Knowledge Organization (NASKO), June 15-16 2011, Toronto, Canada* (s. 23-32). Hentet fra https://journals.lib.washington.edu/index.php/nasko/article/view/12787

Kipp, M.E.I. & Campbell, D.G. (2006). Patterns and Inconsistencies in Collaborative. *Proceedings of the American Society for Information Science and Technology 43*(1), 1-18. doi: 10.1002/meet.14504301178

Kirkegaard, B. (2008). *Metadata elements preferred in searching and assessing relevance of archived television broadcast by scholars and students in media studies: Towards the design of surrogate records* (Doktorgradsavhandling). Hentet fra http://nordicom.statsbiblioteket.dk/ncom/files/44339/BKI_Thesis.pdf

Kultur- og kirkedepartementet. (2012). *Vedtekter for Norsk rikskringkasting AS*. Oslo: Kultur- og kirkedepartementet 2012. Hentet fra http://www.regjeringen.no/upload/KUD/Medier/NRK/NRKs_vedtekter_per_juni_2012.pdf

Lakoff, G. (1987). *Women, fire and dangerous things*. Chicago: The university of Chicago.

Lancaster, F.W. (2003). *Indexing and abstracting in theory and practice*. Campaign, Ill: University of Illinois.

Lancaster; F.W, Elliker, C. & Colonell, T.H. (1989). Subject analysis. *Annual Review of Information Science and Technology (ARIST), 24*(1989), 34-84. Hentet fra http://kb.osu.edu/dspace/handle/1811/45109

Li, A. (2013). *How taxonomies help news organizations understand and categorize their content*. Hentet fra: http://www.poynter.org/how-tos/digital-strategies/222187/how-taxonomies-help-news-organizations-understand-and-categorize-their-content/

Liu, D.C. (2013). *Taggeregler i NRK*. Internt dokument fra 21. mai 2013

Marlow, C., Naaman, M., Boyd, D. & Davis, M. (2006). *Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead*. Hentet fra http://www.danah.org/papers/WWW2006.pdf

Mathes, A. (2004). *Folksonomies: Cooperative classification and communication through share metadata*. Hentet fra http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Milstead, J. & Feldman, S. (1999). *Metadata: Cataloging by Any Other Name*. Hentet fra http://www.iicm.tugraz.at/thesis/cguetl_diss/literatur/Kapitel06/References/Milstead_et_al._1999/metadata.html

NISO. (2004). *Understanding metadata*. Hentet fra http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Quintarelli (2005). *Folksonomies power to the people*. Hentet fra http://www.iskoi.org/doc/folksonomies.htm

Raieli, R. (2013). *Multimedia information retrieval: Theory and techniques*. Oxford: Chandos

Rosenfeld, L. (2005). Folksonomies? How about Metadata Ecologies? [Blogginnlegg]. Hentet fra http://www.louisrosenfeld.com/home/bloug_archive/000330.html

Røed, T, Günter, G.S., Johannesen, I., Bakken, T., Ording, E.E., Bakke, A.K. … Bjørnsrud, S. (2011a). *Systematikk og arbeidsflyt for innholdsmetadata i NRK: Gruppe Arbeidsflyt*. Internt dokument fra 1. mars 2011.

Røed, T, Günter, G.S., Johannesen, I., Bakken, T., Ording, E.E., Bakke, A.K. … Bjørnsrud, S. (2011b). *Morgendagens arkiv*. Internt dokument fra 1. mars 2011.

Simonsen, H.G. (2012a, 3. august). Semantikk: språkvitenskap. I *Store norske leksikon*. Hentet fra http://snl.no/semantikk%2Fspr%C3%A5kvitenskap .

Simonsen, H.G. (2012b, 11. desember). Kognitiv Lingvistikk. I *Store norske leksikon*. Hentet fra http://snl.no/kognitiv_lingvistikk.

Salton, G. & McGill, M.J. (1983). *Introduction to modern information retrieval*. New York: McGraw Hill.

Smith, Gene. (2008). *Tagging: People-powered metadata for the social web*. Berkeley, Calif: New Riders.

St.meld. nr. 30 (2006–2007). (2007). *Kringkasting i en digital fremtid*. Oslo: Kultur- og kirkedepartementet.

St. meld. nr. 24 (2008-2009). (2009). *Nasjonal strategi for digital bevaring og formidling av kulturarven*. Oslo: Kultur- og kirkedepartementet.

Svenonius, E. (2000). *The intelectual foundation of information organization*. Cambridge, Mass.: MIT Press.

Søbak, V. (2013). *Desentralisert indekseringspraksis: En studie av det semi-kontrollerte vokabularet i NRK* (Masteroppgave). Hentet fra https://oda.hio.no/jspui/handle/10642/1588

Tennis, J. T. (2006). Social tagging and the next steps for indexing. *17th Annual ASIS&T SIG/CR Classification Research Workshop, November 3-8, 2006 Hilton Austin, Austin, Texas, USA.* doi: 10.7152/acro.v17i1.12493

Tonkin, E. (2006). Searching the long tail: Hidden structure in social tagging. *17th Annual ASIS&T SIG/CR Classification Research Workshop, November 3-8, 2006 Hilton Austin, Austin, Texas, USA*. doi: 10.7152/acro.v17i1.12494

Tremoen, H.N., Günther, G.S., Engan, T., Johnsen, J.A. & Howlid, M.H. (2013). *Evaluering av metadataregistrering i NRK: Er metadataføringen god nok for intern gjenfinning?* Internt dokument fra 18. januar 2013.

Vanderwal, T. (2005). Explaining and Showing Broad and Narrow Folksonomies [Blogginnlegg]. Hentet fra http://www.vanderwal.net/random/entrysel.php?blog=1635

Voorbij, H.J. (1998). Title keywords and subject descriptors a comparison of subject search entries of books in the humanities and social sciences. *Journal of Documentation 54*(4), 466-476. doi: 10.1108/EUM0000000007178

Voss, J. (2007). Tagging, Folksonomy & Co. Renaissance of Manual Indexing. Hentet fra arXiv:cs/0701072

Weinberger, D. (2006).*Taxonomies and Tags: From Trees to Piles of Leaves*. http://www.hyperorg.com/blogger/misc/taxonomies_and_tags.html

Yi, K. & Chan, L.M. (2009). Linking folksonomy to Library of Congress subject headings: an exploratory study. *Journal of Documentation, 65*(6), 872-900. doi: 10.1108/00220410910998906

7.1 Personlig kommunikasjon

Therese Engan, gruppeleder Metadatahjelp ved Arkiv og research, NRK.

Oddbjørn Holgersen, rådgiver og tidligere seksjonsleder Film- og bearbeiding, NRK.

John Arne Johnsen, mediearkivar, Metadataseksjonen i Arkiv og research, NRK.

Danielle Chiosso Liu, mediearkivar, Metadataseksjonen i Arkiv og research, NRK.

Maja Wettmark, seksjonsleder Metadataseksjonen i Arkiv og research, NRK.