# Re-kategorisering af Infomedia-kilder

Kvantitativ indholdsanalyse (‘content analysis’ på engelsk) er en klassisk disciplin i medieforskningen (Krippendorff 1980, Eskjær & Helles 2015). I forskning såvel som i studerendes opgaveskrivning ved de danske universiteter benyttes skrevet journalistisk indhold fra landets største aviser ofte som empirisk materiale i analyser, der sigter mod at redegøre for den danske nyhedsdæknings intensitet, diskursive mønstre, narrative karakteristika mv. I Danmark kan størsteparten af skriftligt journalistisk indhold tilgås via databasen, Infomedia, hvortil de danske forskningsbiblioteker har licens.

Infomedia er imidlertid ikke et forskningsarkiv, der er kurateret med henblik på medieforskning, men en database, der først og fremmest er beregnet til kommerciel medierådgivning og relativt aktuel nyhedsovervågning. Dette har en række uhensigtsmæssigheder i forskningsregi. Særligt i forskningsprojekter med *mediehistoriske interesser* som eksempelvis vores kortlægning af den danske debat om HPV-vaccination gennem 20 år giver Infomedias inddeling af kilder i ‘landsdækkende dagblade’, ‘lokale ugeaviser’, ‘magasiner’, ‘webkilder’ osv. en række metodiske og analytiske udfordringer. Den nuværende indeksering er præget af en tid, hvor printmediet var det dominerende skriftlige medie, og logikken bag tager kun dårligt højde for, hvordan nyhedsmediebilledet har udviklet sig i de to seneste årtier. Dels har den digitale nyhedsdistribution bevirket, at (papir)**publikationsfrekvens** (dagligt, ugentligt) og et medies **materielle karakteristika** (avis, magasin, web) er blevet sekundære parametre i bestemmelsen af en nyhedskildes type. I dag har mange printmedier for eksempel en digital udgave, hvor indhold publiceres løbende - i modsætning til tiden hvor det redaktionelle arbejde kun blev offentliggjort, de få gange om ugen, måneden eller året papirudgaven gik i trykken.

Dels er Infomedias kildetyper **ikke gensidigt udelukkende**. Dette skaber det en række **dubletter** i databasen, fordi digitale udgaver af papirartikler figurerer som selvstændige artikler fra særskilte ‘webkilder’. Da der ikke nødvendigvis er navnesammenfald eller nogen eksplicit sammenkædning i databasen mellem eksempelvis et ‘landsdækkende dagblad’ som *Berlingske* og dennes tilsvarende ’webkilde’ [*b.dk*](http://b.dk), er disse besværlige at korrigere for (maskinelt) i eksempelvis en frekvensoversigt over hvor hyppigt en term som ‘HPV’ er nævnt i de danske medier over tid.

Et lignende problem med dubletter gør sig gældende med de fleste af de **gratisaviser**, der kom frem i begyndelsen og midten af 00’erne. Selvom flere af disse i analytisk øjemed med rette kan betragtes som landsdækkende, er gratisaviserne typisk registreret i Infomedia som flere forskellige ‘regionale og lokale dagblade’ og deres digitale udgaver som særskilte ‘webkilder’. *Urban* figurerer for eksempel som hele 6 selvstændige lokaludgaver og derudover 2 webkilder. Det betyder, at hver enkelt *Urban*-artikel, der nævner ’HPV’ i princippet tæller med op til 8 gange i en umiddelbar frekvensopgørelse af antallet af artikler i Infomedia, der indeholder tekststrengen ‘HPV’, som den henholdvist Suppl et al. (2018) og Hansen & Schmidtblaicher (2019) for eksempel har lavet. Tilsvarende tæller en artikel fra *Metroxpress* med op til 4 gange (3 fysiske lokaludgaver og 1 webkilde, mens den i flere dele af landet husstandsomdelte gratisavis, *Nyhedsavisen*, blot tæller 1.

Holder man sig derimod til kun at analysere ‘landsdækkende dagblade’, er *Urban* og de andre gratisaviser med en enkelt undtagelse fraværende på listen. Det samme gælder andre landsdækkende nyhedsmedier, der ikke har nogen papirudgave, eksempelvis [*DR.dk*](http://DR.dk), [*nyheder.tv2.dk*](http://nyheder.tv2.dk) og [*Zetland.dk*](http://Zetland.dk). Disse er indekseret som ‘webkilder’ sammen med (og dermed i metodisk øjemed på linje med) andre, mere nicheorienterede kilder, blandt andet netmediet [*Kongressen.com*](http://Kongressen.com), der beskæftiger sig med amerikanske samfundsforhold, [*Cancer.dk*](http://Cancer.dk), der er Kræftens Bekæmpelses foreningswebsite, som udgiver artikler fra papirmagasinet *Tæt på Kræft* i digitalt format, og det populærvidenskabelige netmedie [*Videnskab.dk*](http://Videnskab.dk). Kategorien ‘webkilder’ er imidlertid over årene kommet til at omfatte intet mindre end 1.346\* kilder af meget forskellig karakter, rangerende fra [*Ramboll.dk*](http://Ramboll.dk) og [*RockUnderBroen.dk*](http://RockUnderBroen.dk), der måske lidt overraskende figurerer på en liste over nyhedsmedier, over mindre nyhedssites som [*sundhedspolitisktidsskrift.dk*](http://sundhedspolitisktidsskrift.dk) og [*Dagens.dk*](http://Dagens.dk) til store nyhedssites med eller uden komplementerende papirudgivelse såsom [*Jyllands-Posten.dk*](http://Jyllands-Posten.dk) og [*DR.dk*](http://DR.dk).

Dertil kommer at Infomedias egen kategoriseringen af kilder (efter medietype) er både inkonsistent, tilfældig og fejlbehæftet: I kategorien ‘landsdækkende dagblade’ er *Avisen Danmark* eksempelvis registeret som fire separate medier (med navnetilføjelserne *eftermiddag, fyns amts avis, morgen,* og *trekantsområdet*). I kategorien ‘webkilder’ finder vi henholdvis [Jyllands-Posten.dk](http://Jyllands-Posten.dk) og [Jyllands-posten.dk](http://Jyllands-posten.dk) (altså med stort og lille p) som to forskellige medier, mens Børsen optræder som 3 ‘landsdækkende dagblade’ med *Børsen, Børsen Lørdag / Søndag* og *Børsen Tillæg* og 5 ‘webkilder’ med [*Borsen.dk/nyheder/finans*](http://Borsen.dk/nyheder/finans), [*Borsen.dk/nyheder/penge*](http://Borsen.dk/nyheder/penge), [*Borsen.dk/play*](http://Borsen.dk/play), [*Borsen.dk/prointernational*](http://Borsen.dk/prointernational) og [*Borsenpro.dk/finans*](http://Borsenpro.dk/finans) *(abonnementsområde)*. Samtidig figurerer *Weekendavisen* som et ‘landsdækkende dagblad’ til trods for at avisen aldrig har kørt med daglige (print)udgivelser. Der kan dog anes et analytisk ræsonnement bag at placere ***Weekendavisen*** i samme kategori som medier som *Information* og *Berlingske.* Dette ræsonnement hører imidlertid op, når også medier som ***Licitationen - Byggeriets Dagblad*** og ***Effektivt Landbrug*** forefindes i samme kategori.

I kategorien ‘regionale og lokale dagblade’ optræder *JydskeVestkysten* som hele 10 forskellige papirmedier (til trods for at det redaktionelle indhold om HPV er identisk), mens *Fyens Stiftstidende*, der dækker et lignende geografisk område, blot er registreret som 1 medie. Forklaringen er sandsynligvis at måderne avisernes annoncering er organiseret på, er forskellig. Det gør dog ikke opdelingerne hensigtsmæssige i forskningsregi. Kigger vi endvidere længere ned i kategorien med ‘lokale ugeaviser’ bliver udsvingene i registreringslogikken blot endnu større.

Alt I alt bevirker disse forhold - og flere andre, som vi ikke har plads til at komme ind på her - at Infomedias ad hoc-inddeling af nyhedskilder efter en blanding af eventuelt ikke-eksisterende papirudgaver, (annoncerings)geografi og materielle karakteristika er analytisk utidssvarende og ofte decideret problematisk i forhold til aktuelle forskningsinteresser. I denne artikel præsenterer vi derfor en **re-kategorisering** af Infomedias kilder, som vi har udarbejdet i forbindelse med vores case-studie om den danske debat om HPV-vaccination. Denne re-kategorisering består dels af en ”**clustering**“, der rydder op i Infomedias (fejl)registreringer og samler analytisk nært beslægtede kilder, der af forskellige grunde har forskellige kildekoder i databasen, under navnet på det **mediebrand**, der er afsender. Således kommer mediebrandet *Børsen* til at indeholde alle outlets, der kan tilskrives at være *Børsen* i analytisk forstand, altså både de tre udgaver af papiravisen, der oprindeligt var registreret som ‘landsdækkende dagblade’, og de 5 URL’er, der var registreret som ’webkilder’. Denne clustering er tilgængelig som et script.

Dels består re-kategoriseringen af en omkodning af Infomedias kilder i en række nyudviklede **genrekategorier**, der forsøger at tager højde for aktuelle forskningsinteresser i at arbejde kvantitativt med visuelle repræsentationer af nyhedsstof over tid. I denne omkodning har vi blandt andet fundet det relevant at udvikle en særskilt genrekategori til ‘gratisaviser’ samt at gøre kategoriernes navne fri for markører, der indikerer kanal, udgave, materiale/format eller har ophæng på annoncørforhold. Eksempelvis finder vi det meningsfuldt at erstatte kategorien ‘landsdækkende dagblade’ med ‘landsdækkende nyheder’.

Metodisk tager denne omkodning udgangspunkt i [Infomedias egen kildeliste, der er tilgængelig på databasens forside | En fortrolig leverandørliste | Et udtræk fra CHCAA over det materiale, vi har modtaget]. Denne liste kodes først af 2 medieforskere med henblik på navneclustering. Dernæst udvikles analytiske genrekategorier, hvorefter listen igen kodes af 2 medieforskere med henblik på at tildele hvert mediebrand en genrekategori. I begge kodninger laves der intercoder reliability test på 10% af indholdet. I tilfælde med uenighed diskuteres ræsonnementet bag.

Når re-kategoriseringen er færdig, udarbejder vi en ny figur, der viser de danske mediers intensitet i dækning af HPV gennem 20 år og sammenligner resultatet med en tilsvarende figur udarbejdet af Suppl et al. 2018. Vi tester desuden hvordan rensningen/rekategoriseringen påvirker CHCAA’s analyser af bl.a. Novelty/resonnance og diskuterer resultatet.