Om udgaven *Grundtvigs Værker*. Fremtid, muligheder og udfordringer

Klaus Nielsen

Den igangværende digitale udgave *Grundtvigs Værker* (www.grundtvigsværker.dk eller www.grundtvigsvaerker.dk) skal indtil 2030 udgive N.F.S. Grundtvigs trykte forfatterskab i tekstkritisk etableret og digital form. Materialet – omtrent 37.000 tryksider fordelt på 1.000 værker – forsynes med kommentarer af forskellig slags samt faksimiler af de tilgrundliggende førsteudgaver. Udgivelsesprojektet har i skrivende stund udgivet 32 % af det samlede korpus svarende til ca. 12.000 sider. I sidste nummer af *Grundtvig-Studier* bragtes en præsentation af udgaven: dens opbygning, sigte og retningslinjer belyst ud fra eksempler samt en status for arbejdets fremdrift. I denne artikel rettes blikket fremad for at undersøge, hvilke potentialer for læsere og brugere en digital udgave medfører: avancerede søgemuligheder, korpusanalyser, digital kollationering, samt hvilke udfordringer især for udgavens bevaring det digitale medie indebærer.[1]

Indledning

Denne artikel er den anden af to, som giver en præsentation af udgivelsesprojektet *Grundtvigs Værker* (*GV*). I sidste nummer af *Grundtvig-Studier* blev udgavens opbygning, sigte og retningslinjer belyst ud fra en række eksempler, og jeg gav en status for arbejdets fremdrift. I denne artikel skal

[1] Klaus Nielsen har siden 2013 været udgaveleder for *Grundtvigs Værker* og har været med til at udarbejde retningslinjerne. Første artikel "Om udgaven *Grundtvigs Værker*. Opbygning, principper og status" er trykt i *Grundtvig-Studier 2017* (147-170).

udgavens umiddelbare fremtid behandles med særlig fokus på de digitale potentialer og udfordringer.[2]

Hvor første artikel kunne holdes på et relativt konkret niveau, bliver det anderledes abstrakt for denne. Her skal det dreje sig om fremtidsplaner, hvoraf nogle er søsat, men stadig befinder sig på et tidligt stadie, og andre blot er på ideplan. Endelig er en række blot nævnt her som potentielle muligheder, der kunne være spændende at realisere, men som vi ikke har prioriteret. Flere af disse planer og tiltag stammer fra brugeres ønsker, som vi har haft stor glæde og udbytte af at lytte til. Udfordringen for nærværende artikel er desuden, at den omhandler emner og opgaver, som er under udarbejdelse i skrivende stund. Detaljer, overvejelser og oplysninger om de enkelte tiltag vil potentielt være anderledes eller have forskubbet sig, når artiklen er trykt og i læsernes hænder. I skrivende stund er nærmere bestemt juni måned 2018.

Anvendelse af data

Fordelene ved en digital udgave er mange. Den optager ikke hyldeplads, alle har adgang til den fra hvor som helst på kloden, og man kan søge igennem materialet på kryds og tværs ud fra forskellige parametre. Der er også uhensigtsmæssigheder forbundet med mediet. Skærmlyset anstrenger øjnene og gør det vanskeligt at læse længere tekster. Udgavens webbaserede platform kræver stabil internetforbindelse – et problem, som dog med udviklingen af den mobile datadækning bliver mindre fremover. Og endelig foretrækker mange stadig papirbogen som medie, når der skal læses. Det er den form, hvori vi er vant til at møde litteraturen. Vi kender dens funktionalitet. Den er en solid og stabil medieform – og god at have med ud på havebænken.

Særligt et forfatterskab som Grundtvigs egner sig til at søge igennem. Omfanget gør det vanskeligt for én forsker at læse alt. Og skulle man komme igennem de mange tusinde siders førsteudgaver for ikke at tale om senere optryk og manuskriptmateriale, da vil man have vanskeligt ved at

[2] For kyndig gennemgang af manuskriptet takkes Kirsten Vad, filolog ved Grundtvig Centeret, Ditte Laursen, leder af Afdeling for Digital Kulturarv, Det Kongelige Bibliotek, og Lene Offersgaard, ingeniør ved Center for Sprogteknologi, Københavns Universitet.

håndtere stoffet, huske de indre forbindelser, holde styr på forfatterskabets mange røde tråde, overskue tendenser, stilistiske karakteristika, genrer og emner. Kort sagt kræver et forfatterskab af dette omfang en metodisk afgrænsning for at kunne behandles på en meningsgivende måde. Med adgang til materialet i digital form bliver det imidlertid muligt at gennemse forfatterskabet på en anden måde og måske trække nogle nye røde tråde op, som ikke tidligere har været opdaget.

Brugernes søgebehov er meget forskellige.[3] Nogle søger efter et bestemt citat – de ved så at sige, hvad de leder efter, men ikke hvor det kan findes. Andre søger efter bestemte ord eller begreber i et forsøg på at afdække, hvad Grundtvig har skrevet og evt. ment om dette eller hint. Hvor man i første tilfælde ønsker at få et begrænset antal søgeresultater – og gerne blot dét ene rigtige, man eftersøgte – kan der ligge et potentiale i et bredt søgeresultat, hvis man vil undersøge f.eks. Grundtvigs kvindesyn. Der findes også mange andre tænkelige muligheder. Fælles for disse er imidlertid, at søgeresultatet afhænger af datagrundlagets kvalitet. Dette vil være de fleste bekendt, som har prøvet at lave en søgning i books.google.com. Bøgerne er her blot skannet og OCR-behandlet, dvs. udsat for en maskinel tekstgenkendelse (Optical Character Recognition). Googles OCR er temmelig god, men der findes en del eksempler på fejllæsninger især af gamle bøger trykt med fraktur og på andre sprog end engelsk. På books.google.com vises den elektroniske og søgbare OCR-tekst som et usynligt lag under det skannede billede af bogsiden. Til tider kan man se, hvordan OCR-behandlingen har fejllæst et ord, idet den skannede bogside har en anden ordlyd end den underliggende OCR-tekst, som ligger til grund for det markerede søgeresultat. Andre gange kan man observere det eftersøgte ord på den skannede bogside, uden at søgemaskinen har kunnet finde det. Igen skyldes det, at OCR-behandlingen har fejllæst den trykte kilde.

[3] Jeg vil i nærværende artikel anvende betegnelsen *bruger* om personer, som udnytter den digitale udgaves potentiale til andre formål end læsning, og betegnelsen *læser* om dem, som primært anvender udgaven til at læse i. Der er i denne skelnen ingen værdiladning, og jeg tør som udgaveleder kun håbe på, at udgaven har begge typer modtagere. Distinktionen har imidlertid etableret sig inden for forskningen i digital tekstteori og har sine fordele, når man skal drøfte forskellige anvendelsesmuligheder for digitale udgivelsesprojekter, se f.eks. Rasmussen (2014, særligt kapitel 4.4: 81-87) om den digitale udgaves tre læserroller: læseren, brugeren og bidragsyderen.

Med digitale tekstkritiske udgaver som *GV* er kvaliteten anderledes. Ingen publikation er fejlfri, men med de tekstkritisk etablerede udgaver kommer vi så tæt på et en-til-en-forhold mellem original og digital tekst som muligt. Det betyder, at resultaterne af søgninger i et tekstkritisk etableret korpus er milevidt bedre end søgninger i rå OCR-tekster. Udfordringen er imidlertid omfanget, hvor de færreste udgivelsesprojekter kommer i nærheden af Googles mange millioner titler. Selv for den, som blot vil søge inden for et forfatterskab, kan der være udfordringer: Indeholder udgaven alt, hvad forfatteren har skrevet, eller er en række tekstgrupper udgrænset, f.eks. manuskripter og brevmateriale? *GV* udgiver kun det trykte forfatterskab, hvilket afskærer brugeren fra at kunne søge i det omfattende manuskriptmateriale, herunder bl.a. de utrykte, men nok så betydningsfulde prædikener.

Med *GV* udgiver vi løbende materiale, hvilket betyder, at korpussøgninger på hjemmesiden p.t. kun omfatter 32 % af det samlede materiale, som skal udgives. Det er med andre ord begrænset, hvad disse søgeresultater kan bruges til. Har man held til at finde det citat, man eftersøgte, er der ingen problemer. Søger man efter udtalelser fra Grundtvigs side om et bestemt emne, er 32 % af korpus ikke tilstrækkeligt. På Grundtvig Centeret har vi det samlede materiale i digital form, dvs. alle de trykte (og enkelte utrykte) værker, vi skal udgive. Det stammer fra den indledende skanningsproces, hvormed udgivelsesarbejdet blev igangsat i 2010. Alt materialet blev skannet og OCR-behandlet med en udmærket kvalitet, dvs. egentlige korpussøgninger kan lade sig gøre, og resultaterne er nogenlunde.[4] I forbindelse med hver offentliggørelse af nyt materiale to gange årligt laver vi en opdateret korpusmappe bestående af de udgivne og dermed kontrollerede tekster plus de resterende OCR-tekster. P.t. indeholder mappen altså 32 % kvalitetssikret tekst og 68 % ikke-kontrolleret tekst af svingende kvalitet. Således bliver korpus bedre for hver offentliggørelse, idet andelen af kontrolleret tekst støt fortrænger andelen af OCR-tekst.

Dette materiale kan til hver en tid rekvireres ved kontakt til Grundtvig Centeret, hvilket flere forskere allerede har benyttet sig af. For imidlertid at gøre adgangen til korpusmaterialet enklere har vi indledt et samarbejde med Center for Sprogteknologi ved Københavns Universitet, som tilby-

[4] Minimum 97 % korrekt, hvilket ligger i den gode ende af spektret, jf. Holley (2009). For et anderledes positivt syn på anvendeligheden af OCR-tekster af ringe kvalitet – den såkaldte *dirty OCR* – se Cordell (2017).

der opbevaring i et repositorium for danske tekstkorpora med en platform for data mining og visualiseringsværktøjer. Tiltaget foregår i regi af CLARIN-DK, som er en del af den nationale infrastruktur for digitale forskningsprojekter DIGHUMLAB, et netværk med base på Aarhus Universitet. CLARIN-DK er et konsortium bestående af de fire danske universiteter og KB, hvis mål er at sikre en digital infrastruktur, hvor forskere kan deponere og hente sprogligt materiale af forskellige medietyper.[5] Som en del af tiltaget er der etableret et særligt repositorium integreret med Voyant Tools, et webbaseret analyse- og visualiseringsværktøj, som er open source, dvs. frit tilgængeligt, og som gør det nemmere for forskere uden særlige IT-kundskaber at foretage data mining og korpusanalyser.[6] Mulighederne med Voyant Tools er mange og inkluderer bl.a. analyser af ordfrekvenser, frekvensfordelinger og KWIC-konkordanser (Key Word in Context) – alt sammen med visualiseringer af resultaterne i en lettilgængelig brugerflade. Blandt de allerede etablerede korpora er en række delmængder af Arkiv for Dansk Litteratur (ADL), Georg Brandes' *Hovedstrømninger* og *GV* ordnet i en række større pakker af hensyn til programmets håndtering af tekstmængden. Værkerne er således samlet for hvert årstal, dvs. at alle værker udgivet i f.eks. 1811 findes i én samlet fil. Det gør det enklere for applikationen og for brugeren at anvende den store mængde data. Viser det sig med tiden, at en anden praksis er mere hensigtsmæssig, kan vi foretage ændringer i forbindelse med en af de to årlige opdateringer. Feedback hilses med andre ord velkommen.

I tillæg til disse visualiseringsværktøjer kan man via CLARIN-DK downloade den seneste version af den samlede korpusmappe, hvis man ønsker at bruge andre værktøjer eller blot lave standard fritekstsøgninger i materialet. Korpusmapperne er versionsnummereret i henhold til udgaven, dvs. at når *Grundtvig-Studier 2018* udkommer, vil version 1.13 være offentliggjort, og korpusmappen vil bære samme nummer.[7] Et eksempel på anvendelsen af dette samlede materiale er allerede udgivet i artikelform. Katrine Frøkjær Baunvig og Kristoffer Laigaard Nielbo har analyseret informationsniveauet i det samlede tekstkorpus og beskrevet resultaterne i artiklen "Kan man validere et selvopgør? En fjernlæsning

[5] Se mere på info.clarin.dk/ samt http://dighumlab.com/.
[6] Se mere om analyseværktøjet på repository.clarin.dk/VoyantTools/.
[7] Se hdl.handle.net/20.500.12115/31.

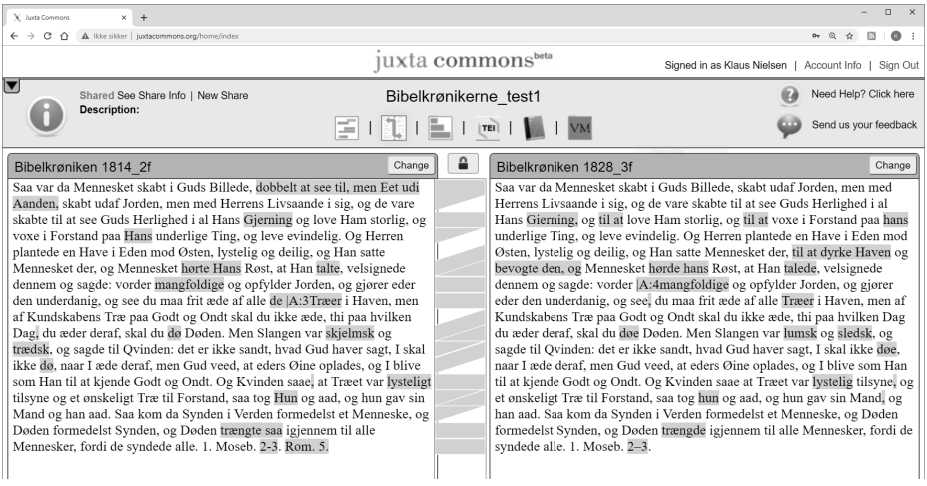
af Grundtvigs forfatterskab og en vurdering af Kaj Thanings reception" (2017). Baunvig og Nielbo benytter sig af *text mining,* også kaldet fjernlæsning eller *distant reading,* dvs. den proces hvormed software så at sige trækker informationer ud af tekster. Specifikt analyseres udviklingen over tid af bl.a. ordfrekvenser, publikationshyppighed og -omfang, titellængde samt teksternes entropi, som måler informationsniveauet per enhed og dermed giver et udtryk for leksikalsk forudsigelighed.[8] Ved hjælp af disse digitale værktøjer er Baunvig og Nielbo i stand til at opstille en række perioderinger baseret på de ekstraherede data, som dels støtter, dels problematiserer tidligere periodiseringer i Grundtvigforskningen baseret på mere manuelle eller analoge nærlæsninger. I takt med at flere forskere bliver fortrolige med digitale analyseværktøjer, skal det blive spændende at følge udviklingen i Grundtvigforskningen og se, hvilke nye billeder af forfatterskabet, personen og receptionen det vil medføre.

En anden måde, hvorpå udgavens data kan udnyttes til andet end almindelig læsning, er i forbindelse med værktøjer til automatisk kollationering og identifikation af varians. Af de forskellige applikationer, som findes, er Juxta formentlig den enkleste at bruge.[9] Den findes både som et program til download og installering på egen maskine og som webapplikation, dvs. som en hjemmeside, hvor man kan uploade sine tekster og få dem behandlet og vist på forskellig vis. Applikationen kan håndtere de fleste gængse tekstformater (txt, xml, html, docx, rtf, pdf), og efter nogle få indledende skridt kan man få vist to tekstkilder ved siden af hinanden med udpegning af forskellene mellem dem. I illustrationen nedenfor er et afsnit fra begyndelsen af *En liden Bibelkrønike for Børn og Menigmand* sat op, således at førsteudgavens ordlyd (1814, 2 f.) vises til venstre, og andenudgavens (1828, 3 f.) til højre. Hvor de to versioner afviger fra hinanden, har applikationen markeret tekstsegmentet, f.eks. er førsteudgavens frase "dobbelt at see til, men Eet udi Aanden" markeret, idet den mangler i andenudgaven. Applikationen skelner ikke mellem større og mindre varianter – eller interessante og uinteressante – hvorfor rent ortografiske varianter, f.eks. *Gjerning* > *Gierning* og *dø* > *døe*, fremhæves på lige fod med sletninger og tilføjelser. Sorteringen mellem varianter må man med andre

[8] Se Baunvig & Nielbo (2017, 59-62). For en mere teknisk gennemgang se desuden Baunvig, Nielbo, Liu & Gao (2018).

[9] Se juxtacommons.org/.

ord selv foretage. Applikationen fungerer bedst, når omfanget af variansen ikke er stort, ligesom den heller ikke er god til at håndtere passager, som har skiftet plads i teksten.



Parallelvisning af *En liden Bibelkrønike* i hhv. første og anden udgave.

En forudsætning for at kunne bruge Juxta er, at man har adgang til tekstkilderne i elekronisk form. Og som tilfældet er i forbindelse med søgning, vil resultatets kvalitet altid være afhængig af teksternes. På *GV* udgiver vi som udgangspunkt kun førsteudgavernes tekst, men en række af Grundtvigs digt- og salmesamlinger indeholder genoptryk, hvorfor disse senere versioner også vil være at finde på *GV*. F.eks. er digtet "Høstgildet" fra 1815 repræsenteret i to versioner: førsteudgaven i *Nyeste Skilderie af Kjøbenhavn,* nr. 86, 28. oktober (sp. 1361-1366) og anden udgave fra samme år i Grundtvigs *Heimdall. Dansk Nyaars-Gave for 1816* (188-195).

Med adgangen til udgavens tekstkorpus via CLARIN-DK kan brugeren altså selv uploade og etablere variantvisninger til brug for genetiske analyser af Grundtvigs revisionspraksis. Heri ligger et spændende forskningsfelt, som den digitale udgave kan være med til at opdyrke.[10]

[10] Inklusionen af senere genoptryk i *GV* er beskrevet i den første artikel, se Nielsen (2017, 156 f.). Her er også en reference med link til den forbilledlige norske tekstportal boskelskap.no, som med deres udgivelse af Camilla Colletts roman *Amtmandens Døttre* (1854/55 og 1879) allerede har etableret et eksempel på, hvordan Juxta kan kombineres med udgivelsen.

Mulighed for udvidelser

Et digitalt udgivelsesprojekt har naturligvis en økonomisk ramme og dermed også en endelig slutdato, men der er intet i vejen for, at nye projekter kan opstå og bygge videre på det allerede etablerede materiale. Man kunne bl.a. forestille sig, at Grundtvig-Arkivets manuskriptmateriale kunne gøres til en del af udgaven, hvis en fremtidig bevilling kan skaffes. På et mere overkommeligt niveau ville udgavens anvendelighed styrkes betragteligt ved inklusionen af en række bibliografisk ekstramateriale i digitalt og søgbart format. I det omfang, *GV* kan udrede evt. forfatterrettigheder, ville det være en oplagt og relativ enkel udvidelse at etablere søgbare versioner af Steen Johansens *Bibliografi over N.F.S. Grundtvigs skrifter* (1948-1954), *Registrant over N.F.S. Grundtvigs Papirer* (1957-1964) samt fortegnelserne over Grundtvigs bogsamlinger fra 1839 og 1873. Fortegnelsen fra 1805 er allerede tilgængelig på *GV*.

En anden mulighed er at supplere udgavens tekstmateriale med opgaveformuleringer rettet mod forskellige uddannelsestrin lige fra folkeskolen til universitetet. På den digitale portal for *Henrik Ibsens Skrifter* findes en række undervisningsresurser, som er tilvejebragt i samarbejde med undervisere fra de forskellige trin.[11] Et lignende samarbejde for *GV* kunne etableres evt. med en forankring på DPU. Spændevidden i forfatterskabet gør, at Grundtvig nemt kan gøres gældende og relevant for alle uddannelsestrin, og elever, studerende og undervisere vil kunne få glæde af de hjælpemidler, som udgaven tilbyder.

Fælles for disse ideer er, at de udspringer af brugeres ønsker, og vi modtager jævnligt forslag til lignende gode ideer. Vi lytter og samler ideerne sammen, men må desværre erkende, at midlerne og tidsplanen kun tillader, at de færreste sættes i værk. På dette punkt er vi stærkt afhængige af eksterne samarbejdspartnere som f.eks. CLARIN-DK nævnt ovenfor. Dette tiltag er et godt eksempel på, hvordan et brugerønske er blevet modtaget, mulighederne undersøgt og overvejet nøje og endelig løst i et samarbejde med dygtige folk inden for et tilgrænsende fagområde.

[11] Se http://www.ibsen.uio.no/undervisningsressurser.xhtml.

Fremtiden hinsides finansiering

I slutningen af 2016 tog Grundtvig Centeret initiativ til et samarbejde mellem Det Kongelige Bibliotek (KB), Det Danske Sprog- og Litteraturselskab (DSL) og Center for Sprogteknologi ved Københavns Universitet (CST) med det formål at opbygge en fælles kodningspraksis samt en fælles visningsapplikation, som vil kunne vedligeholdes og drives centralt og dermed bidrage til at sikre en langsigtet bevaring af og adgang til digitale materialer. Samarbejdet blev i opstartsperioden støttet af DIGHUMLAB og tilknyttet deres portefølje som en såkaldt SIG (Special Interest Group).

Grundtanken bag samarbejdet er enkel. Digitale udgaver såsom *GV*, *Søren Kierkegaards Skrifter, Ludvig Holbergs Skrifter, H.C. Andersens Manuskripter* og lignende har alle samme udfordringer. Parallelt med udarbejdelsen af udgavens indholdsmæssige struktur (tekstkritiske retningslinjer, afgrænsning af materialet og kommenteringsniveau med mere) skal projekterne fastlægge en praksis for kodning af teksterne samt etablere en visningsapplikation. Denne skal omsætte de kodede tekster til et læsbart format med de funktioner, man har valgt at inkorporere i kodningen, f.eks. links til kommentarer, faksimiler, databaser eller varianter. Både applikation og datafiler skal hostes på en server. *GV* hostes p.t. af KB i Aarhus (det tidligere Statsbiblioteket), men har sin egen visningsapplikation, som blev kodet til formålet af en ekstern leverandør i forbindelse med udgavens opstart i 2010.[12] *GV* står selv for vedligeholdelsen heraf, suppleret med en versionskontrol i forbindelse med hver offentliggørelse samt et minimum af teknisk assistance fra KB. Endelig skal der udarbejdes en fornuftig og langsigtet bevaringsplan for projektet, som sikrer, dels at det store arbejde bevares for eftertiden, dels at det ikke forældes i takt med den tekniske udvikling efter projektets afslutning. Inden for IT-verdenen regner man normalt med fem-syv års levetid for software, og vores digitale udgivelsesprojekter skulle meget gerne holde længere end som så. Imidlertid har der ikke været rettet tilstrækkelig opmærksomhed mod dette aspekt af digitale projekter, hverken fra politisk hold eller fra forskningsverdenen.

[12] Faktisk bygger den nuværende visningsapplikation i store træk på visningsapplikationen for den finske udgave *Zacharias Topelius Skrifter* (topelius.fi), som Svenska litteratursällskapet i Finland lod os benytte.

Situationen er ikke ny, idet vi har haft digitale projekter inden for dansk humanistisk forskning siden 1970'erne. Et tidligt eksempel, hvis levetid desværre blev alt for kort, er folkeviseprojektet *Svøbt i mår. Dansk folkevisekultur 1550-1700,* 1-4, red. Flemming Lundgreen-Nielsen og Hanne Ruus (1999-2002). Med til bind 3 fra 2001 fulgte en CD-rom med ca. 1.000 folkeviseversioner i elektronisk format, hvoraf flere af teksterne var beriget med et ortografisk neutralt lag, således at søgning på tværs af viserne blev nemmere for brugeren. Derudover indeholder CD-rommen visekataloger, et udvalg af faksimiler, en komplet ordbog over sprogformen i viserne, supplerende dokumentation til artiklerne i de fire bind samt ikke mindst en visningsapplikation eller grænseflade, som giver brugeren adgang til det omfattende materiale og den dertil udviklede søgefunktion.[13] Til sammen repræsenterer CD-rommens data og applikation et værdifuldt værktøj til folkeviseforskningen. Problemet er blot, at tiden meget hurtigt løb fra både applikationens kodning og det valgte medie. I dag har de færreste computere CD-drev, men de kan dog stadig tilkøbes. Lykkes det at få CD'en i en maskine, vil man desværre opdage, at applikationen ikke længere kan køre på moderne styresystemer. På hjemmesiden for projektet Digitale Undersøgelser af Dansk Sprog (DUDS) kan man læse viseteksterne og søge i dem enkeltvis ved hjælp af browserens fritekstsøgning.[14] På siden linkes desuden til en version af den oprindelige søgemaskine på repositoriet CLARIN-DK, men linket virker imidlertid ikke. Man kan som bruger få adgang, men p.t. kun ved kontakt til CLARIN-DK. Der arbejdes på en løsning, som skal sikre adgang og bedre anvendelsesmuligheder. Således er det store arbejde altså ikke endegyldigt tabt, men eksemplet demonstrerer meget tydeligt, hvilke udfordringer digitale projekter og deres brugere står over for. Skal adgangen sikres for eftertiden, kræver det vedligeholdelse og ikke mindst midlerne til at foretage denne som en vedvarende indsats.

Et tilgrænsende eksempel fra forlagsbranchen er *Den Store Danske Encyklopædi* (1994-2006), som i 2004-2006 udkom på hhv. CD-rom og DVD. Ingen af disse kan i dag køre på moderne styresystemer. Tabet er ikke så stort, idet materialet siden 2009 har været frit tilgængeligt på websitet denstoredanske.dk. Men købere af de to digitale medieformer kan med rette føle sig frustreret over deres korte levetid og ønske sig tilbage til en

[13] Jf. Ruus (2001, 489-505).

[14] duds.nordisk.ku.dk/tekstresurser/aeldste_danske_viseoverlevering/.

tid, hvor encyklopædien i digitalt format var reklamefri og dets indhold stabilt, dvs. ikke udsat for brugeres velmente, men til tider fejlbehæftede revisioner. Desuden annoncerede Gyldendal i 2017, at websitet vil blive lukket pga. manglende rentabilitet. Vi stod med andre ord i fare for at miste den digitale adgang til nationalencyklopædien og dermed muligheden for at søge på tværs af artiklerne. Først i november 2018 dukkede en redningsplan op, idet midler blev afsat i finansloven til opdatering og ajourføring, og encyklopædien er nu sikret levetid frem til og med 2022.[15]

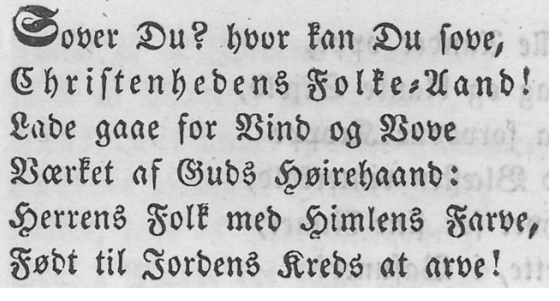
For projekter som *GV*, hvis eneste output er en digital udgave, er holdbare bevaringsplaner af afgørende vigtighed. Den amerikanske litteraturforsker Jerome J. McGann har i sin seneste bog *A New Republic of Letters* (2014) berørt denne problematik. McGann har været med i litteraturforskningens og editionsfilologiens digitale 'vending' siden 1990'erne og har nogle interessante synspunkter på dens fremtid. Han kommer til den foruroligende konklusion, at for at sikre sig adgang i fremtiden til den elektroniske udgave *The Rosetti Archive*, som han selv har skabt og været hovedredaktør af, må han printe samtlige 70.000 digitale filer ud på papir (McGann 2014, 136-138). Udgaven er nemlig havnet mellem to stole, hvad angår den institutionelle forankring, hvorfor ingen føler ansvar for det tekniske vedligehold af hjemmesiden. Selv om *GV* endnu ikke står på en lignende brændende platform, idet projektet er berammet til at løbe indtil 2030, må løsningerne udtænkes og planerne lægges nu.

At langsigtede bevaringsplaner ikke er en del af den nationale forskningspolitik – og af det politiske system som et hele – vidner om en blind vinkel, det er på tide at få oplyst. William Kilbride, leder af det engelske Digital Preservation Coalition, har i en artikel fra 2016 (414) fremhævet en række af de udfordringer, Digital Humanities står over for, hvad angår bevaring. Blandt andet opfordrer han til, at det politiske og øvrige bevillingsgivende system inddrages ved f.eks. at stille krav om, at bevaringsplaner fremover må være obligatoriske i projektbeskrivelser. Da andre projekter og institutioner står med samme udfordring, er et tæt samarbejde netop den rigtige løsning. Inden jeg beskriver de foreløbige tiltag og resultater, vil det dog være passende med en kort introduktion til digitale projekters tekniske side.

[15] Jf. *Aftaler om finansloven for 2019* (2018, 71).

Digitale udgaver er grundliggende opbygget ens. De kan deles op i to hovedbestanddele: data og applikation. Data er de tekster, som udgaven består af, enten i rå, ubehandlet form eller beriget med forskellige oplysninger, links eller lag. Applikationen er det stykke software, som skal omdanne data til en læsbar tekst på skærmen. Andre betegnelser for denne del kan også bruges: hjemmeside og grænse- eller brugerflade. De bruges tit synonymt, selv om der er nuanceforskelle, som jeg dog ikke vil komme ind på her. Det vigtige er, at data *i sig selv* ikke har nogen særlig værdi for den almindelige læser af en digital udgave, og applikationen er derfor en essentiel del af det digitale projekts publikation. Den er så at sige vores eneste mulighed for at præsentere resultatet af arbejdet.

Fremtidssikringen af disse to bestanddele er forskellig. Hvad angår data, er der en høj grad af sikkerhed i anvendelsen af standardformater. De fleste udgivelsesprojekter vælger at benytte opmærkningssproget XML (Extensible Markup Language), som er en fast etableret standard til beskrivelse af data. Der findes tilmed et særligt sæt retningslinjer for opmærkning af tekster i XML, defineret af det internationale netværk eller konsortium TEI (Text Encoding Initiative). TEI blev oprettet i 1980'erne og består af forskere og udgivere med baggrund i Digital Humanities og videnskabelig tekstudgivelse. Konsortiet har opstillet et meget omfattende sæt retningslinjer for, hvordan tekster kan beskrives vha. XML-koder, de såkaldte *TEI P5 Guidelines* (*TEI P5*). I disse specificeres en række *tags* (markører), som sættes rundt om elementer i teksten, og en række *attributter* (beskrivende tilføjelser til tags), som anvendes til nærmere beskrivelse af disse elementer f.eks. i forhold til typografisk opsætning, semantisk struktur eller indhold – alt efter hvilke formål man har defineret for opmærkningen. Et eksempel kan bedst demonstrere dette. Første strofe af Grundtvigs digt "De Christnes Aand" (1842) vises nedenfor i et udklip fra førstetrykket i *Nordisk Tidskrift for christelig Theologi* og efterfølgende i den XML-opmærkede version fra *GV*:


Sover Du? hvor kan Du sove,
Christenhedens Folke-Aand!
Lade gaae for Vind og Vove
Værket af Guds Høirehaand:
Herrens Folk med Himlens Farve,
Født til Jordens Kreds at arve!

```
<lg n="1">
  <l><hi rend="initial">S</hi>over Du? hvor kan Du sove,</l>
  <l><hi rend="spaced">Christenhedens Folke-Aand!</hi></l>
  <l>Lade gaae for Vind og Vove</l>
  <l>Værket af Guds Høirehaand:</l>
  <l>Herrens Folk med Himlens Farve,</l>
  <l>Født til <seg type="com" n="com10">Jordens Kreds</seg> at arve!</l>
</lg>
```

Hver markering består af et start-tag og et slut-tag, som omgiver et element. Således indledes hvert vers ovenfor af start-tagget <l>, hvor l står for *line*, og afsluttes af slut-tagget </l>. Slut-tags angives altid med skråsteg. Verslinjen identificeres således som et element i teksten. Markeringen af verslinjer har ingen attributter i *GV's* kodning. Det har imidlertid markeringen af hele strofen <lg>, som står for *line group*. Start-tagget har fået en n-attribut med værdien 1, hvilket angiver, at de seks vers, som befinder sig inden for hhv. <lg> og </lg>, udgør første strofe.

Også markørerne for fremhævelser har attributter. Det forstørrede begyndelsesbogstav, som indleder digtet, er omgivet af tagget for fremhævelser <hi> (*highlight*) og har fået attributten rend, som står for *rendition* (gengivelse), med attributværdien initial. Hele andet vers er fremhævet med datidens standard fremhævelsesmiddel *spatiering*, hvormed små mellemrum indsættes mellem hvert bogstav. Derfor er hele verslinjen omgivet af start- og slut-tagget for <hi> med attributværdien *spaced*.

Disse eksempler er primært af typografisk karakter, og langt de fleste indføres ved hjælp af automatiske processer. I sidste verslinje findes et eksempel på et link, som peger ud af selve teksten. Frasen "Jordens Kreds" er omgivet af tagget <seg> (*segment*), som identificerer et element til nærmere bestemmelse. Attributten type angiver med sin *com*, at der er tale om et kommentarkrævende element. Det element, som omgives af seg-tagget, er kommentarens lemma, og n-attributtens værdi, com10, peger på den tilhørende forklaring i en separat kommentarfil, som hentes frem, når læseren klikker på lemma i visningsapplikationen. Her får man at vide, at "Jordens Kreds" skal forstås som jorden – slet og ret.

Opmærkningen her er baseret på udgavens behov. Man kan også bruge TEI's XML-opmærkning til mange andre formål, f.eks. en semantisk opmærkning, som registrerer alle substantiver, verber og præpositioner i en

tekst, eller en syntaktisk, som registrerer sætningsenheder, nominalfraser og lignende.

GV's opmærkning er TEI-konform, hvilket betyder, at den overholder de forskrifter for XML-kodning af tekster, som TEI har defineret i deres retningslinjer. TEI-konsortiet tør nok spås en lang levetid, men selv om det skulle hænde, at det om 100 år ikke længere findes, vil XML-opmærkningen ud fra *TEI P5* være dels tilstrækkelig selvforklarende, dels tilstrækkelig dokumenteret til, at opmærkningen vil kunne afkodes, oversættes og vises. Der er således en indbygget fremtidssikring i anvendelsen af disse XML-koder, som sikrer, at brugere ikke kun vil kunne læse de opmærkede tekster, men også kunne gøre brug af det tekstkritiske arbejde og andre forskningsresultater, som ligger i selve koderne.

Anderledes forholder det sig, når vi flytter blikket fra udgavens data til dens applikation. Som eksemplet ovenfor har vist, kan den XML-opmærkede strofe nok læses, men det er ikke nogen rar oplevelse. (Og eksemplet var endda mildt og valgt med pædagogisk omhu; andre værker i *GV* har en meget mere omfattende opmærkning). De XML-opmærkede tekster må oversættes og præsenteres i en læsbar visning, som bl.a. omsætter den typografiske opmærkning til konkret typografi, og som aktiverer den funktionelle del af opmærkningen til operationelle funktioner, f.eks. koblingen mellem lemmaet "Jordens Kreds" med forklaringen com10 i den tilhørende kommentarfil. Dette foregår via en række transformationsprocesser, hvor *stylesheets* giver instrukser til behandling og visning af de forskellige XML-koder. Og her træder vi dels uden for udgavelederens ekspertise, dels uden for den fremtidssikring, som XML-opmærkningen repræsenterer. En vigtig pointe er, at visningsapplikationer ikke er lige så godt sikrede som datafilerne, idet de kræver vedvarende teknisk vedligeholdelse. Den tekniske del er for så vidt ikke en udfordring – det er imidlertid, at den skal være vedvarende.

En succesfuld bevaringsplan, som ikke kun gælder opbevaring af datafiler, men også sørger for en vedligeholdelse af applikationen – en slags kuratering af data – må nødvendigvis have en institutionel forankring. Og som fremhævet af Kilbride (2016, 416) må denne forankring nødvendigvis udgå fra et tæt samarbejde og en åben dialog mellem forskere og institutioner.

Samarbejdet mellem *GV*, KB, DSL og CST blev indledt i slutningen af 2016 og begyndte at tage form med to seminarer eller workshops, som

blev afholdt i hhv. februar og marts 2017. Begge havde de som samlende overskrift: "Digitale udgaver og resurser" – i det første arrangement fokuserede vi på brugerens perspektiv og i det andet på udgiverens. Formålet med begge var at samle erfaringer og synspunkter fra disse to sider og tage den viden med over i det videre arbejde frem mod en fælles løsning.

Blandt oplægsholderne på første seminar var forskellige brugertyper repræsenteret: undervisere, forskere og studerende, som hver især fortalte om deres anvendelse og oplevelse af digitale udgaver og tekstresurser. Fokus var holdt på et praktisk niveau: Hvordan understøtter funktionaliteten forskellige former for anvendelse? Hvordan deles stoffet med andre (f.eks. mellem underviser og studerende)? Hvilke fordele og ulemper har den ene eller anden udgave eller resurse? Disse erfaringer og synspunkter blev efterfølgende drøftet i grupper og sammenfattet til et sæt anbefalinger eller ønsker til, hvordan en digital udgave eller tekstresurse kan fungere bedst muligt. Faktisk var der relativ stor enighed blandt workshoppens deltagere om, hvad der udgjorde de vigtigste funktioner og løsninger. Dette var en opløftende opdagelse. Trods store forskelle i materialet, som omfattede tekstsamlingen ADL, *Ludvig Holbergs Skrifter*, *Diplomatarium Danicum*s middelalderhåndskrifter og den nationale avissamling (Mediestream.dk), var brugernes ønsker og behov nogenlunde de samme. I korte træk var disse følgende:

- Tydelig deklaration af indholdet (type, kvalitet, omfang)
- Ensartet og lettilgængelig henvisningspraksis (unikke links, sidetal til evt. trykt original)
- Velfungerende og stabil søgefunktion (nem navigation mellem søgeresultater, afgrænsning ift. værk, forfatter, periode)
- Mulighed for at downloade udgavens kildemateriale i forskellige formater (pdf, xml, samlede datasæt)
- Øvrige analyseværktøjer til f.eks. data mining.

Andet seminar fokuserede på udgiverens perspektiv og havde bl.a. oplæg fra flere institutioner, som enten arbejdede på eller allerede havde etableret et sæt retningslinjer for kodning af tekster til anvendelse ved flere forskelligartede projekter. Første dags oplæg dannede udgangspunkt for den følgende dags gruppediskussioner og sammenfatning. Ud af drøftelserne var det endnu en gang muligt at opsummere de forskellige synspunkter

og erfaringer, og resultatet heraf blev første skridt på vejen mod den fælles kodningspraksis, som efterfølgende er udarbejdet.

En arbejdsgruppe blev nedsat bestående af medlemmer fra hver af de fire involverede institutioner med den opgave at formulere en kodemanual baseret på *TEI P5*. Resultatet, *TEI Fælles Praksis* (*TEI FP*), er grundlæggende en delmængde af *TEI P5,* som specificerer, hvilke af dennes mange muligheder for kodning af samme tekstelementer der bør bruges. *TEI P5* skal dække over et meget bredt spektrum af udgivelsestyper, og derfor er det gjort muligt at kode samme element på mange forskellige måder. Dette byder på vanskeligheder i forhold til fælles vedligeholdelse, bevaring og visning af udgivelser, som nok kan være kodet ud fra *TEI P5,* men reelt kan se meget forskellige ud med anvendelse af vidt forskellige koder til markering af samme datatyper. Med *TEI FP* er hensigten dels at gøre det nemmere for nye udgivere at udvælge sig et sæt koder til projektet, dels at gøre det enklere for eftertidens pligtafleveringsinstitutioner at sikre bevaring og fortsat adgang til materialet. Rationalet bag er, at udgivelser, som er kodet ens, dvs. som anvender samme anbefalede delmængde af *TEI P5*, kan vedligeholdes og vises med samme sæt værktøjer.

Med *TEI FP* har vi forsøgt at tilgodese så mange hensyn som muligt, og derfor har dokumentationen været sendt rundt til en lang række fagfæller ved både nationale projekter (Digitale Undersøgelser af Dansk Sprog (DUDS), Den Arnamagnæanske Samling, H.C. Andersen Centeret) og internationale (bokselskap.no, Litteraturbanken.se, *Henrik Ibsens Skrifter,* Deutsches Textarchiv og Svenska litteratursällskapet i Finland). Tilbagemeldinger fra disse sparringspartnere har været af stor værdi for det videre arbejde.

Nok kan det være vanskeligt at blive enige om en fælles best practice, når manges synspunkter skal tilgodeses, men med den ovenfor beskrevne proces er vi kommet langt i retningen af en holdbar fælles løsning på de udfordringer, som digitale udgivelsesprojekter står over for. Resultaterne fra de to første seminarer har efterfølgende dannet udgangspunkt for KB's videre arbejde med at udvikle en fælles visningsapplikation med tilhørende forretningsmodel og partnergruppe. Arbejdet hermed er stadig i gang.

I juni 2018 afholdtes det tredje og foreløbigt sidste seminar i rækken. Temaet denne gang var slet og ret fælles visningsapplikationer, og blandt oplægsholderne var repræsentanter for udenlandske projekter, som arbejder med lignende udfordringer: bokselskap.no, Svenska litteratursällska-

pet i Finland og Forum für Editionen und Erschliessung i Schweiz. Som afslutning på seminaret kunne KB vise en meget lovende testversion af deres nye fælles visningsapplikation. Arbejdet hermed vil fortsætte i 2019 med inklusion af *Søren Kierkegaards Skrifter, Trykkefrihedsskrifterne* og materiale fra *GV* som de første delmål. På længere sigt er det tanken, at *GV* i sin helhed skal overføres til KB's applikation. Som nævnt tidligere er *GV* allerede en del af KB's digitale portefølje i kraft af hosting-aftalen, men med dette nye tiltag bliver samarbejdet tættere, idet KB fremover også vil varetage visning og bevaring.

Den fælles visningsapplikation bliver ikke kun en fordel for de enkelte udgivelsesprojekter, men også for brugerne, som opnår en ensartet tilgang til tekstkorpora af forskellig slags. Lige nu skal brugerne selv undersøge, hvilke forfatterskaber der er tilgængelige i pålidelige digitale udgaver. Med en fælles visningsapplikation kan disse potentielt samles, således at brugere nemt kan søge på tværs af forfatterskaber. Desuden skal brugeren kun lære én applikation at kende. En brugerundersøgelse foretaget af bokselskap.no viste, at adgang til større tekstsamlinger af forskellige forfattere øger anvendelsen.[16] Brugere af en fælles visningsapplikation vil ganske enkelt komme til at læse mere end brugere af en forfatterskabsudgave – og det kan vi kun glæde os til.

Afslutning

Det editionsfilologiske arbejde med at genudgive Grundtvigs forfatterskab i en kritisk etableret og kommenteret udgave kan næsten virke som en simpel opgave, når det ses i lyset af de udfordringer, som den hastigt udviklende IT-verden medfører. Forandringerne kan ikke forhindres, og udgivelsesarbejdet må derfor tilpasse sig og indstille sig på de vilkår, det digitale medie opstiller. Dog er det vigtigt at understrege, at digitale projekter ikke bør stå alene over for disse udfordringer, hvorfor koalitioner, sparring og fælles forståelse er afgørende for en succesfuld løsning. Ej heller bør den digitale forskningsverden stå alene, men må samarbejde med

[16] Bokselskaps hovedredaktør, Ellen Nessheim Wiger, præsenterede detaljerne i foredraget "Livet etter lansering – om bruk og brukere av bokselskap.no" ved konferencen *Utgåvor i använding*, 6.-8. oktober 2017 i Helsinki. En artikel følger med de øvrige bidrag i 2019.

bevaringsinstitutioner, bevillingsgivere og det politiske system for at opnå holdbare løsninger. Kun på den måde kan vi sikre os, at fremtidige brugere, både forskere og almindelige læsere, kan få adgang til det værdifulde materiale – værdifuldt i både indholdsmæssig og økonomisk henseende.

Forkortelser

CST	Center for Sprogteknologi, Københavns Universitet
DSL	Det Danske Sprog- og Litteraturselskab
GV	*Grundtvigs Værker* (2010-), Grundtvig Centeret, Aarhus Universitet (www.grundtvigsværker.dk)
KB	Det Kongelige Bibliotek
TEI	Text Encoding Initiative (www.tei.org)
TEI FP	*TEI Fælles Praksis,* version 5, december 2017
TEI P5	*TEI P5 Guidelines,* version 3.3.0, januar 2018 (www.tei-c.org/Guidelines/P5/)
XML	Extensible Markup Language

Litteratur

Værker af Grundtvig (alle udgivet i *GV*)

(1805) *Optegnelse paa de Bøger Jeg ejer og ejende vorder begyndt d. XXde Decbr MDCCCV, Nikolaj Frederik Severin Grundtvig, Ludimagister Egeløkkis,* manuskript i Grundtvig-arkivet på Det Kongelige Bibliotek, København, fascikel nr. 501.1.

(1814) *En liden Bibelkrønike for Børn og Menigmand,* København.

(1815) *Heimdall. Dansk Nyaars-Gave for 1816,* København.

(1815) "Høstgildet" i *Nyeste Skilderie af Kjøbenhavn,* nr. 86, 28. oktober, sp. 1361-1366, København.

(1828) *En liden Bibel-Krønike for Børn og Menig-Mand,* 2. udg., København.

(1842) "De Christnes Aand" i *Nordisk Tidskrift for christelig Theologi,* 4:2, 149-156, København.

Værker af andre forfattere

Aftaler om finansloven for 2019 (2018), København, Finansministeriet.

Baunvig, Katrine Frøkjær og Nielbo, Kristoffer Laigaard (2017), "Kan man validere et selvopgør? En fjernlæsning af Grundtvigs forfatterskab og en vurdering af Kaj Thanings reception" i Henrikson, Paula, Malm, Mats og Söderlund, Petra (red.) (2017), *Textkritik som anaslysmetod,* Nordisk Netværk for Editionsfilologer. Skrifter 12, Stockholm, Svenska Vitterhetssamfundet, 45-67.

Baunvig, Katrine Frøkjær, Nielbo, Kristoffer Laigaard, Liu, Bin og Gao, Jianbo (2018), "A Curious Case of Entropic Decay: Persistent Complexity in Textual Cultural Heritage" i *Digital Scholarship in the Humanities,* fqy054. https://doi.org/10.1093/llc/fqy054

Cordell, Ryan (2017), "'Q i-jtb the Raven': Taking Dirty OCR Seriously" i *Book History* 20, 188-225. http://ryancordell.org/research/qijtb-the-raven/

Fortegnelse over Nik. Fred. Sev. Grundtvigs Bogsamling ved Johan Grundtvig. Christianshavn 1839, manuskript i Grundtvig-arkivet på Det Kongelige Bibliotek, København, fascikel nr. 520.

Fortegnelse over den af N. F. S. Grundtvig efterladte Bogsamling, som bortsælges ved offentlig Auction i Klædeboderne Nr. 38 Mandagen den 29. Sept. 1873 Kl. 10 (1873), København, E.C. Løsers Bog- og Nodetrykkeri.

Holley, Rose (2009), "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs" i *D-Lib Magazine,* 15:3/4. http://www.dlib.org/dlib/march09/holley/03holley.html

Johansen, Steen (1948-1954), *Bibliografi over N.F.S. Grundtvigs skrifter,* København, Gyldendal.

Kilbride, William (2016), "Saving the Bits: Digital Humanities Forever?" i Schreibman, Susan, Siemens, Ray og Unsworth, John (red.) (2016), *A New Companion to Digital Humanities,* West Sussex, Wiley Blackwell, 408-419.

McGann, Jerome J. (2014), *A New Republic of Letters. Memory and Scholarship in the Age of Digital Reproduction*, Cambridge, US, Harvard University Press.

Nielsen, Klaus (2017), "Om udgaven *Grundtvigs Værker.* Opbygning, principper og status" i *Grundtvig Studier 2017,* 147-170.

Rasmussen, Krista Stinne Greve (2014), *Bytes, bøger og læser. En editionshistorisk analyse af medieskiftet fra trykte til digitale videnskabelige udgaver med udgangspunkt i Søren Kierkegaards Skrifter,* ph.d.-afhandling, Københavns Universitet.

Registrant over N.F.S. Grundtvigs Papirer (1957-1964), udg. af Grundtvig-Selskabet af 8. September 1947 og Det Danske Sprog- og Litteraturselskab, 1-30, København.

Ruus, Hanne (2001), "*Svøbt i mår*-cd-rommen" i Flemming Lundgreen-Nielsen og Hanne Ruus (red.) (1999-2002), *Svøbt i mår. Dansk folkevisekultur 1550-1700,* 1-4, København, C.A. Reitzel, bind 3, 489-505.