VISUALISERING AF DANMARKS KULTURARV

Visualisering som metode til at afsløre mønstre i store mængder data er et stærkt redskab til at bygge bro mellem datalogi og andre forskningsområder. Værdien af visuelt design bliver her demonstreret med et eksempel bygget på Dansk biografisk leksikon.

Forfatteren

Stefan Jänicke er adjunkt ved Institut for Matematik og Datalogi, Syddansk Universitet. stjaenicke@imada.sdu.dk

I sin ph.d.-afhandling fra Leipzig Universitet, Tyskland i 2016, undersøger Stefan anvendelsen af visualiseringsteknikker til at understøtte sammenlignende analyser af data indenfor digital humaniora. Han har siden beskæftiget sig med at udvikle informationsvisualisering og visuelle analytiske teknikker indenfor et bredt spektrum af interdisciplinære forskningsprojekter, der adresserer aktuelle forskningsspørgsmål indenfor (digital) humaniora, samfundsvidenskab, biologi og sport.

Artiklen er oversat fra engelsk og redigeret af Carsten R. Kjaer.

Maleriet *Udslidt* af den socialrealistiske maler Hans Andersen Brendekilde (1857-1942) kan ses som en kunstnerisk implementering af ordsproget "et billede siger mere end tusind ord". Maleriet skildrer det barske liv i landdistrikterne i slutningen af det 19. århundrede i Danmark, og det viser en mand, der er faldet om (måske død?) af sit hårde arbejde og hans skrigende kone på den bare mark, der indrammer scenen. Selvom forrige sætning også giver en detaljeret beskrivelse af problemstillingen, formidler maleriet øjeblikkeligt og meget bedre det socialkritiske budskab til observatøren.

Figur 1. Maleriet *Udslidt* af H.A. Brendekilde (1889)

Når vi forskere snakker om *visualisering* som en metode til analyse af data, der vises på skærmen, udnytter vi her den menneskelige evne til hurtigt at opfatte mønstre, der ligger gemt i disse data. En visuel designproces oversætter fænomener fra den virkelige verden til visuelle repræsentationer, som gør fænomenerne lettere tilgængelige for observatøren. I denne proces transformeres objekter relateret til fænomenet til dataenheder, der er beskrevet ved specifikke attributter og relationer mellem dem. Disse attributter og relationer oversættes derefter til visuelle repræsentationer, som kan afspejle de mønstre, der forekommer.

En sådan proces involverer altid en forvrængning, og derfor skal den udføres med omhu for at sikre, at man på baggrund af en visualisering kan drage brugbare konklusioner. Visualiseringer er specielt nyttige værktøjer for fagfolk, der ikke selv er eksperter i datalogi, til at generere og bekræfte hypoteser ud fra digitale datasæt indenfor deres fagområde.

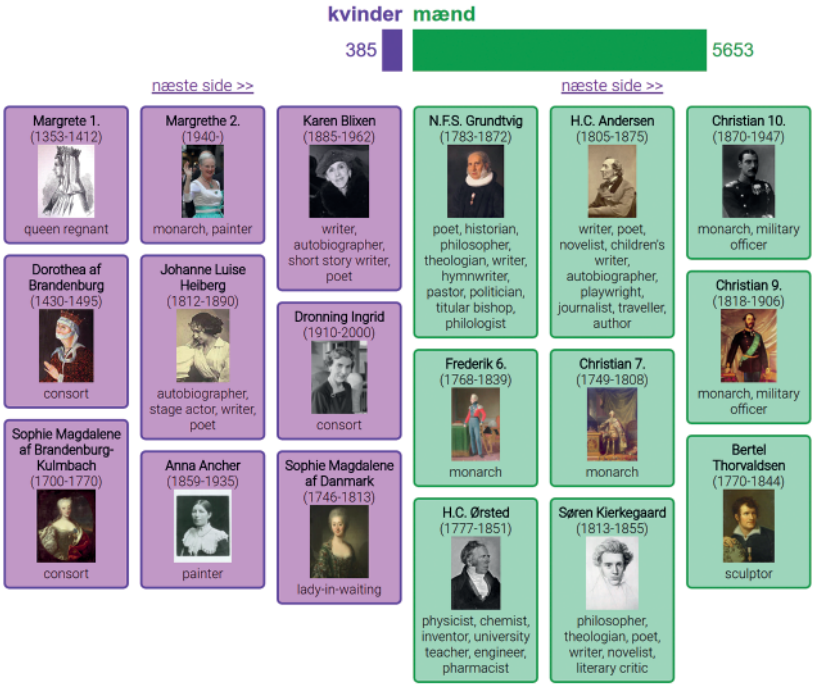
Igennem de seneste ti år har jeg arbejdet sammen med eksperter fra forskellige fagområder, og de fleste af mine projekter har været tværfaglige samarbejder med forskere fra humaniora indenfor rammerne af det, man kalder *digital humaniora*. De seneste tiårs bestræbelser på at digitalisere vores kulturarv har betydet, at humaniora-forskere i dag har adgang til store mængder digitale data. Digital humaniora som forskningsområde går ud på at udvikle løsninger, som kan få mening ud af alle disse data. I denne sammenhæng har visuel udforskning vist sig som et værdifuldt instrument, der får mere og mere betydning for tilvejebringelsen af

ny viden. I denne artikel vil jeg give eksempler på, hvordan visualisering kan afsløre information om forskellige aspekter af Danmarks kulturarv.

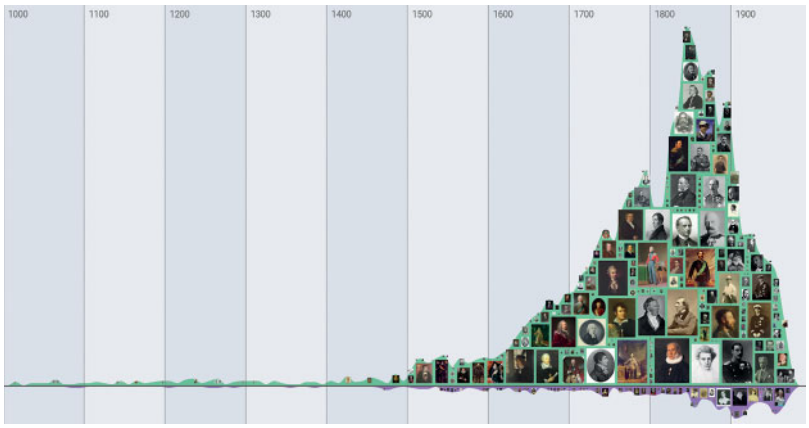
Visuel udforskning af Dansk Biografisk Leksikon

I en artikel om kulturarv understreger den humanvidenskabelige forsker Elena Franchi, at alle mennesker bidrager til verdens kultur. Hvor vigtig en person er for kulturarven kan aflæses af, hvor meget der bliver skrevet om denne person. De mest indflydelsesrige personer for Danmarks kulturarv findes derfor sandsynligvis i Dansk Biografisk Leksikon (DBL). Den tredje og sidste udgave, der blev udgivet mellem 1979 og 1984, indeholder biografisk information om cirka 20.000 personer, både danskere og udlændinge, der i høj grad har bidraget til dansk socialliv og kultur. For omkring 6.000 af disse personer findes der en indgang på den danske udgave af Wikipedia. Det kan man dermed opfatte som en digital udgave af Dansk Biografisk Leksikon (her kaldet dDBL), der omfatter de personer med mest nutidig relevans. Her findes en række strukturerede biografiske oplysninger om køn, levetid, fødested og erhverv for et stort antal personer, hvilket giver et passende grundlag for en visuel udforskning, hvor vi fokuserer på sammenligning af kvinder og mænd i dDBL.

Kønsforholdet i dDBL kan afbilledes som et simpelt søjlediagram (se figur 2). Denne første meget basale form for visuelle repræsentation af data afslører den stærke kønsubalance i DBL nøjagtigt, da det viste forhold svarer til det, man finder i den tredje trykte udgave, hvor cirka 7-8 % er kvinder. Søjlediagrammet er koblet med en listevisning, der ordner alle kvinder og mænd i henhold til deres antydede indflydelse på den danske kulturarv. Denne indflydelse er skønnet ud fra mængden af tekst i de pågældende personers Wikipedia-opslag: jo mere tekst, der er skrevet om en person, jo større skønnes vedkommendes indflydelse at være.



Figur 2. Øverst ses et søjlediagram, der viser antallet af henholdsvis mænd og kvinder, der findes i dDBL (som er de personer, der er omtalt på Wikipedia og som også findes i 3. udg. af Dansk Biografisk Leksikon fra 1979-1984). Nedenunder ses en listevisning, hvor personerne er rangordnet efter deres indflydelse vurderet ud fra mængden af tekst om dem. Samtlige personer er listet, og man kan klikke sig ind på de korresponderende Wikipediaopslag ved at klikke på en person.

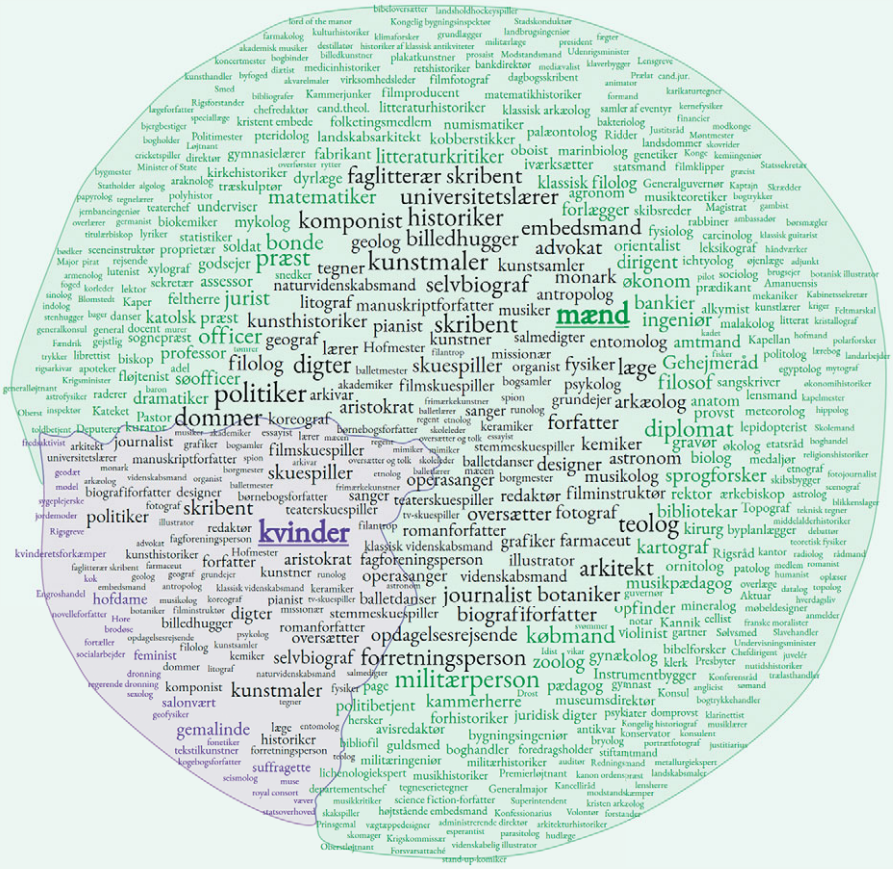


Figur 3. Tidslinje, hvor den grønne kurve viser mænd, og den lilla kurve viser kvinder i dDBL. Personerne er grupperet i årtier ud fra gennemsnittet af leveår (for eksempel er H.C. Ørsted placeret i 1840, da det ligger præcist midtvejs mellem fødsels- og dødsåret (1805-1875). Endnu levende personer er placeret efter fødselsår plus tyve år, og hvis fødselsåret ikke er kendt, placeres personen efter dødsår). Højden på kurverne afspejler antallet af personer. De små billeder angiver de vigtigste personer for den pågældende tid vurderet ud fra mængden af tekst om dem i dDBL.

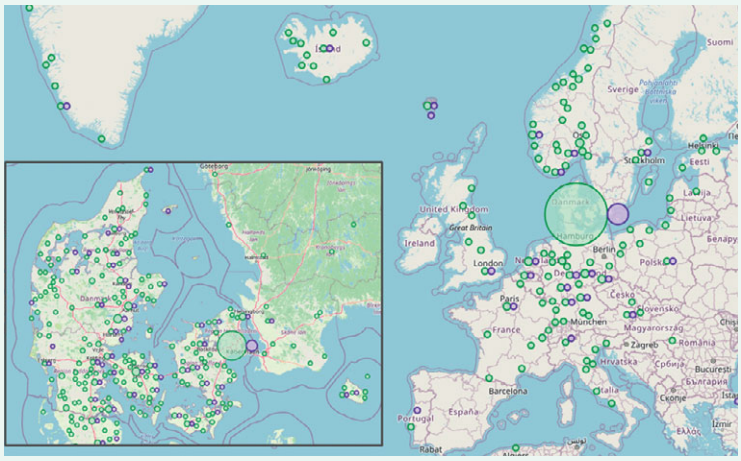
For at understøtte muligheden for at søge information i de biografiske data har jeg designet tre interaktive, visuelle grænseflader, som alle understøtter en sammenlignende analyse af biografiske oplysninger i relation til køn.

Den første af disse visuelle grænseflader viser en tidslinje, hvor kvinder og mænd er adskilt i to separate kurver (se figur 3). De små billeder placeret på kurverne viser de personer, der har haft den største indflydelse i bestemte tidsperioder. Mens det store antal mænd i det 19. århundrede tegner den danske guldalder, en periode med usædvanlig kreativ produktion, fik kvinder tilsyneladende stor opmærksomhed, da den første udgave af DBL blev udgivet mellem 1887 og 1905.

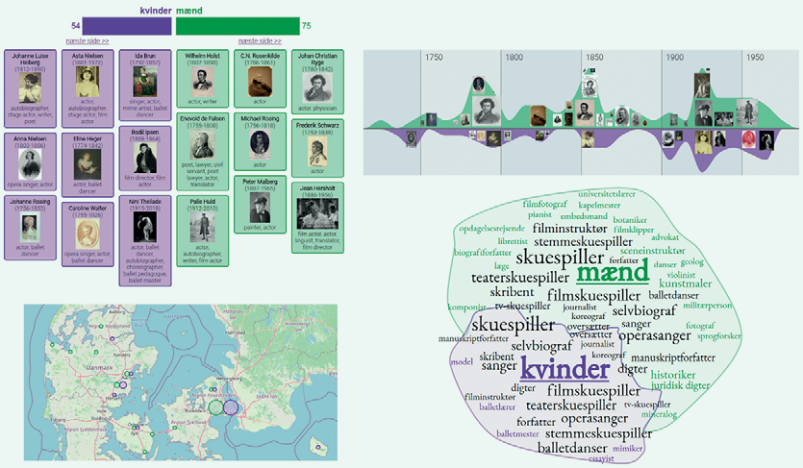
Den anden visuelle grænseflade er et kort, hvor personernes fødeste-



Figur 5: Ordsky, der sammenligner erhverv for de kvinder (lilla) og mænd (grøn) i dDBL, hvor oplysning om erhverv findes (i alt 5.127 personer). Sorte ord er erhverv, som både mænd og kvinder udførte. Hvis lige mange kvinder og mænd udøvede et erhverv, placeres ordet tæt på ordskyens centrum, mens ordet flyttes væk fra centrum, når kønsforholdet er skævt. Skriftstørrelsen afspejler, hvor mange personer, der udøvede det pågældende erhverv.



Figur 4. Kort der viser placeringen af fødesteder for personer i dDBL. Størrelsen på en cirkel afspejler antallet af personer, der er født i en bestemt region. Man kan bemærke, at mange personer, der har spillet en rolle for dansk kulturarv, faktisk er født uden for Danmark.



Figur 6: En visualisering af et søgeresultat i dDBL, som er fremkommet ved at vælge tidsintervallet 1700 til 2000 på tidslinjen og samtidig vælge erhverv relateret til skuespil i ordskyen. Dette giver 129, personer med et rimeligt afbalanceret kønsforhold. Kortet afslører, at skuespil er et ret "urbant erhverv", da personerne typisk blev født i store byer.

der er vist med cirkler (det skal dog nævnes, at oplysninger om fødested kun gives for cirka halvdelen af personerne i dDBL). Denne grænseflade understøtter spørgsmål som: *Hvilke personer af historisk betydning blev født i min by?* Selvom de fleste personer i dDBL blev født i Danmark, dokumenterer den vidt udbredte fordeling af cirkler en international indflydelse på Danmarks kulturarv.

Den tredje visuelle grænseflade er en såkaldt ordsky-visualisering, der illustrerer forskelle og ligheder på erhvervene hos kvinder og mænd i dDBL (figur 4). Ord farvet med lilla eller grønt repræsenterer erhverv, der henholdsvis kun blev udøvet af kvinder eller mænd, mens ord farvet med sort er erhverv, der blev udført af begge køn. Skriftstørrelsen afspejler, hvor mange personer, der har udøvet det pågældende erhverv. De sorte ord er endvidere placeret i forhold til deres kønsbalance.

Ved brug af det samme farveskema giver hver af de tre visuelle grænseflader et oveblik over fordelingen af person-attributter vedrørende køn, tid, fødselssted og profession. Størrelsen på de visuelle attributter afspejler altid mængder, dvs. hvor ofte en specifik attribut optræder hos alle personerne i datasættet. Alle grænseflader er forbundet med hinanden, og muligheden for at vælge specifikke attributter via klik med musen gør det muligt at udforske datasættet visuelt på mange måder. Alle visninger opdateres kun med de personer, der matcher den specifikke forespørgsel, og forespørgslen kan sammensættes af valg, der vedrører flere attributter. Hvis man for eksempel vælger et tidsinterval fra 1700 til 2000 på tidslinjen og erhverv relateret til skuespil i ordskyen, efterlader dette 129 personer med et langt mere afbalanceret kønsforhold (se figur 6).

Dette casestudie om dDBL illustrerer, hvordan visualiseringer kan frembringe letopfattelige mønstre, der kan guide brugeren i en visuel

udforskningsproces. Visualiseringer er i stand til at fremhæve vigtige dataenheder, som i tilfældet med dDBL er personer, der væsentligt har bidraget til Danmarks kulturarv.

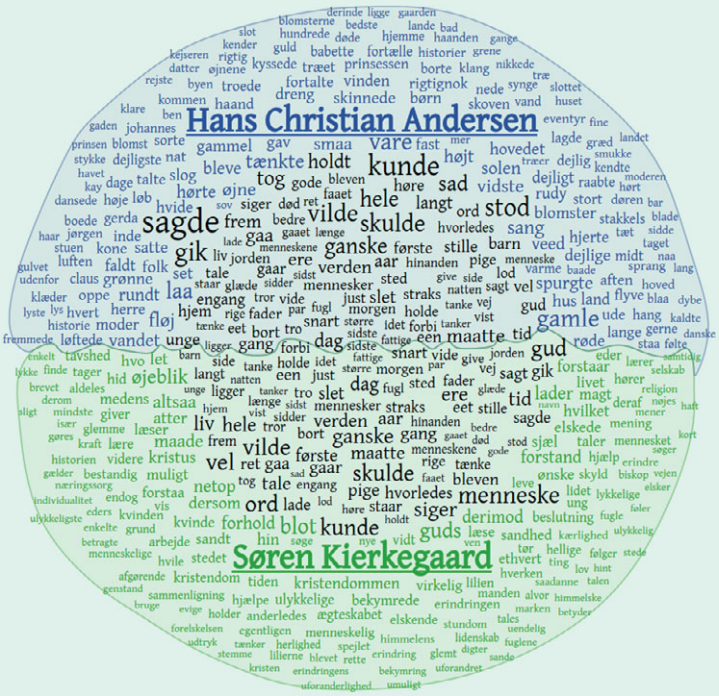
H.C. Andersens vs Søren Kierkegaard

To af de mest indflydelsesrige mandlige personer i dDBL er Hans Christian Andersen og Søren Kierkegaard, som begge anerkendes som de vigtigste danske forfattere i den danske guldalder. Mens H.C. Andersen først og fremmest er kendt for sine eventyr, anses Kierkegaard på baggrund af sit forfatteskab som den første eksistentialistiske filosof. Begge forfattere havde et ambivalent forhold til hinanden, der gik fra foragt til respekt, og de henviste også til hinanden i deres værker. Vi undersøger, om denne ambivalens afspejles i de to forfatteres værker ved en tilgang kaldet *Distant reading*.

Distant reading er et udtryk opfundet af den digitale humanist Franco Moretti, og det beskriver kvantitative tilgange til tekstanalyse, hvor man betragter teksterne i et oversigtsperspektiv, typisk ved at sammenligne hyppigheden af forskellige ord.

Figur 7 giver et sådant kvantitativt syn på et fragmenteret, men repræsentativt tekstkorpus, der inkluderer 89 historier af Andersen og 12 skrifter af Kierkegaard. I figuren er de 250 mest anvendte ord fra begge forfattere visualiseret.

Ved første øjekast afspejles de to forfatteres tematiske områder godt af de blå og grønne sektorer i ordskyen. Mens blåfarvede ord som "prinsessen", "slot / tet" eller "kejseren" eksplicit henviser til Andersens eventyr, viser andre ord, der betegner planter, dyr eller farver, nøjagtigt de scener, som historierne fortælles i. Kierkegaards fokus på religion er på sin side afspejlet af ofte anvendte ord som "gud", "biskop" eller "kristendommen". Den delte del i skyens centrum indeholder primært adjektiver og funktionsord. Begge



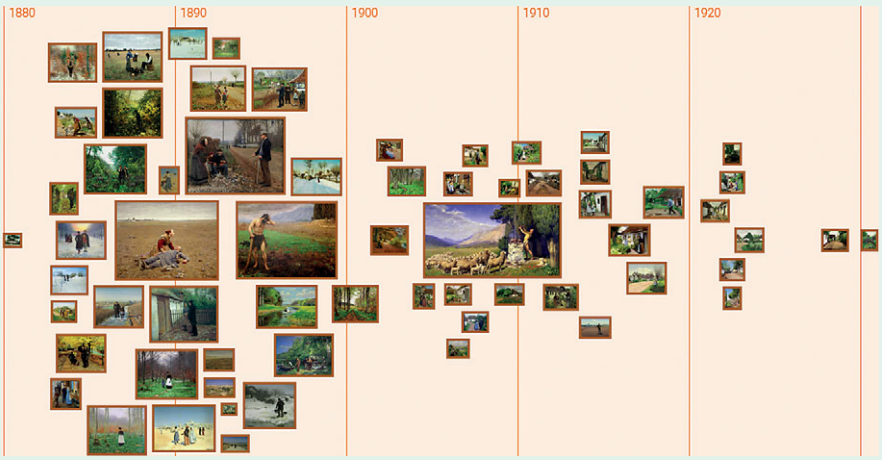
Figur 7. Kvantitativ analyse af 89 historier (med i alt 240.734 ord) af H.C. Andersen (blå) og 12 skrifter (90.770 ord) af Søren Kierkegaard (grøn). I analysen tælles hyppigheden af forskellige ord fra begge forfattere, og de normaliseres i forhold til antallet af ord. Derudover fjernes stopord (dvs. meget almindelige ord som "og", "eller", mv., der normalt udelades fra søgning i elektroniske tekster, da de ikke bidrager væsentligt til meningen). De 250 mest anvendte ord fra begge forfattere er her visualiseret i en ordsky.

Figure 8: Analyse af den kontekst, hvor ordet "verden" optræder i skrifterne af henholdsvis H.C. Andersen og Søren Kirkegaard. Farverne på ordene angiver her den gennemsnitlige afstand til et ord, der optræder sammen med ordet verden. Ord med rødt er typisk deskriptive og optræder meget tæt ved ordet verden i teksten, for eksempel som nærmeste nabo før eller efter. Gule ord definerer typisk den bredere kontekst af "verden", da de optræder længere væk.

forfattere maler imidlertid deres personlige billede af verden, da begge ofte bruger ordet "verden". Figur 8 analyserer den kontekst, hvor ordet verden optræder, mere detaljeret ved hjælp af en anden visuel kortlægning. Nu giver farven information om den gennemsnitlige afstand til et ord, der optræder sammen med "verden", i den skrevne tekst.

Mens Andersens verden er indrammet af temmelig positivt ladede ord som "dejligste", "bedste" eller "lande", er Kierkegaards billede mere kontrastfyldt, da der også optræder mere negative ord som "uddød" eller "hader".

Hvad der yderligere er synligt i begge illustrationerne er, at Andersens



Figur 9. 64 af Hans Andersen Brendekildes malerier arrangeret på en tidslinje. Størrelsen på billederne afspejler deres virkelige størrelse. Det er umiddelbart tydeligt, at Andersens værker fordeler sig i to distinkte perioder – den første med store, ofte socialkritiske malerier, og den anden med mindre og mere idylliske skildringer.

Figur 10. Udateret maleri af Hans Andersen Brendekilde. Størrelsen på billedet antyder, at det mest sandsynligt hører hjemme i hans første kreative periode og dermed snarere kan tolkes som en skildring af det hårde liv på landet frem for et idyllisk sceneri.

karakterer snarere tænker (der er mange varianter af ordet "tænke") og spørger ("spurgte"), mens Kierkegaards karakterer synes at forstå (mange varianter af "forståelse"). Man kan sammenfatte det sådan, at Andersen tegner et fantasifuldt, bevidst naivt billede af verden, der stiller spørgsmål til, hvordan den fungerer, mens Kierkegaard leverer moderne, realistiske svar. Således kan man konkludere, at de to forfattere har været gensidigt befrugtende, komplementære personligheder i deres tid.

Et fjernt syn på Brendekildes malerier

I lighed med begrebet *Distant reading* refererer *Distant viewing* til en kvantitativ analyse af visuelt materiale. Ved at kortlægge billeder på en meningsfuld måde, kan vi afsløre mønstre og derved måske blive i stand til at drage konklusioner. Dette er allerede antydet med eksemplet med dDBL, hvor personer med indflydelse på dansk kulturarv er vist på en tidslinje (figur 3). Vi anvender nu en lignende metode på Hans Andersen Brendekilde og arrangerer 64 af hans malerier på en tidslinje (figur 9). Den relative størrelse på billederne i visualiseringen afspejler den virkelige størrelse på malerierne, hvilket betyder at højden på tidslinjen afspejler, hvor stort et areal Brendekilde malede i et bestemt tidsinterval.

Visualiseringen gør det med det samme tydeligt, at Brendekildes værker placerer sig i to distinkte, kreative perioder. Den første periode fra omkring 1880 til 1900 har det førnævnte socialkritiske maleri *Udslidt* i sit centrum. Det er omgivet af andre malerier som *En Landevej* eller *Fortrykt*, der ligeledes skildrer problemstillinger vedrørende livet i landdistrikterne. Andre malerier illustrerer scener i Middelhavsområdet med udgangspunkt i Brendekildes besøg i Italien, Egypten, Palæstina og Syrien.

I den anden periode fra omkring 1900 til 1930 skiller det religiøse maleri *Abels tilbud* (1908) sig ud. Denne periode er ellers kendetegnet ved små malerier, der viser idylliske scener fra landlivet, hvilket står i kontrast til de socialkritiske temaer i den første periodes malerier.

Samlingen indeholder også udaterede malerier, såsom det, der er vist i figur 10. Den afbillede scene giver plads til fortolkninger. Det kan vise en idyllisk familiescene om vinteren med en gammel mand, der skovler sne med sit barnebarn og en kat, der viser vejen til husets varme indre. Eller det kan vise det hårde, isolerede liv på landet, hvor man var oppe imod naturens vældige kræfter. Da maleriet er ret stort, synes sidstnævnte fortolkning, der placerer maleriet i Brendekildes første kreative periode, mere sandsynlig, fordi han i den anden kreative periode sjældent skildrede det idylliske landsliv på store lærreder.

Bygger bro mellem fagområder

Selvom en del af de mønstre, der afsløres af sådanne visualiseringer, næppe vil komme som en overraskelse for eksperter indenfor de relevante fagområder, er det veldokumenteret i litteraturen, at sådanne nye perspektiver er i stand til at generere nye hypoteser om selv meget velstuderet materiale. En yderligere fordel ved denne intuitive og legende visuelle formidling af information er, at den har stort potentiale for at engagere almene borgere i at interagere med samfundsrelevante data.

Eksemplerne i denne artikel viser, at en og samme visualiseringsteknik kan understøtte vidt forskellige forskningsundersøgelser. Denne generaliserbarhed er uvurderlig, da den rummer potentialet til at bygge bro mellem datalogi og mange andre fagområder. Uden visualiseringer vil algoritmiske metoder ofte være utilgængelige, resultater vil være uforståelige og iboende mønstre i data vil forblive skjult. ■

Videre læsning

Elena Franchi, *What is cultural heritage?*, in Smarthistory, August 8, 2015, accessed May 10, 2020

https://smarthistory.org/what-is-cultural-heritage

https://da.wikipedia.org/wiki/Kategori:*Personer_i_Dansk_Biografisk_Leksikon*, accessed June 11, 2020

B. Shneiderman, *The eyes have it: a task by data type taxonomy for information visualizations*, Proceedings 1996 IEEE Symposium on Visual Languages, Boulder, CO, USA, 1996, pp. 336-343.

Da Kirkegaard røg i flæsket på Andersen: https://videnskab.dk/kultur-samfund/da-kierkegaard-rog-i-flaesket-pa-andersen