SLÆGTSFORSKEREN

MEDLEMSBLAD FOR DANSKE SLÆGTSFORSKERE · **ÅRGANG 36 · NR. 2 · 2021**

I Link-Lives laver vi simple, computergenererede livsforløb for alle danskere 1787 til 1968

Af Anna Kristiane Mortensen, Barbara Revuelta-Eugercios og Anne Løkke

Kirsten Sanders har bedt os skrive lidt om, hvad meningen er med forskningsprojektet Link-Lives. Hvad har vi fundet ud af indtil nu? Vil slægtsforskere kunne få glæde af det, vi laver eller kommer Link-Lives til at spolere fornøjelsen ved at lave slægtsforskning?

Hvad er Link-Lives

Link-Lives er et forskningsprojekt, hvor vi finder og "linker" oplysninger om samme person i alle de transskriberede folketællinger, kirkebøger og københavnske begravelsesprotokoller foreløbig 1787 til 1901, senere helt til 1968. Resultatet bliver rygraden i en ny forskningsinfrastruktur. Partnere i projektet er Rigsarkivet, Københavns Stadsarkiv og Københavns Universitet. Se mere på Link-Lives.dk

Formålet er at gøre det muligt for historikere, epidemiologer, økonomer og andre forskere at forske i de rige, danske kilder til enkelte menneskers liv i stor skala. Når Link-Lives forskningsinfrastrukturen kommer op at køre, vil forskere kunne zoome ind og ud mellem statistiske oversigter over hele befolkningen og de enkelte mennesker, som udgør en gruppe. For eksempel søskendeflokke med mange døde i en mæslingeepidemi, kvindelige håndværksmestre eller familier, hvor alle børn blev voksne. På den måde kan vi få et helt nyt perspektiv på livet i Danmark de seneste århundreder.

Forskningsinfrastrukturen kan blive ved at vokse, fordi stadig flere transskriberede kilder kan inddrages og andre forskningsprojekter, som Rigsarkivets Multigenerationsregister-projekt kan give synergi.
Læs mere om Multigenerationsregistret her: www.sa.dk/da/om-rigsarkivet/samarbejder/etablering-af-nyt-multigenerationsregister/.

Men det er fremtiden. Lige nu laver vi den forskning, der gør det muligt at lave computergenererede links. Teamet består af historikere og computerscience-eksperter.

Hvad er en algoritme?

- En algoritme er en opskrift, der består i præcist beskrevne trin, som fortæller en computer, hvad den skal gøre med input-data, så det fører til et resultat.
- Et computerprogram og en app består af mange algoritmer.
- En algoritme skrives i et programmeringssprog for eksempel Phyton, C++, Basic.
- Eksemplet nedenfor er en forsimplet udgave af vores mest simple algoritme. Den sammenligner to folketællinger og laver regelbaserede links. Den er "oversat" til en slags dansk. De ord, der står med STORE bogstaver, er *operatorer*, som er defineret i forvejen i et programmeringssprog.

LÆS (Navn_1860); LÆS(Navn_1850);
LÆS (Alder_1860); LÆS (Alder_1850);
LÆS (Fødested_1860); LÆS (Fødested_1850);

HVIS Navn_1860 er næsten lig Navn_1850
OG Alder_1860 = Alder_1850 +/- 2
OG Fødested_1860 er næsten lig Fødested_1850
og der ikke er mere end 1 match → ✓ Link
ELLERS → ✗ Ikke Link

Regelbaserede links og Machine Learning baserede links

Et "link" er i Link-Lives én linje i en database, der giver argumentet for at en personregistrering i én kilde vedrører samme person som en personregistrering i én anden kilde. For eksempel at en regelbaseret sammenligning af folketællingen 1860 med folketællingen 1850 viser at én Bodil Madsen i 1860 er den samme person som én bestemt Bodil Madsen i 1850, fordi navn, fødested og alder er ens, og der ikke er andre kandidater i hele Danmark, hvor dette er tilfældet.

Regelbaserede algoritmer, der sammenligner alle personer i to databasefiler, er standardmetoden til at generere sådanne links. Vi har sammenlignet alle de transskriberede folketællinger fra 1787 til 1901 for hele Danmark. To folketællinger ad gangen i alle relevante kombinationer. Først kun med attributterne køn, navn, fødested og alder. Derefter med en algoritme, der også inddrager oplysninger om husstanden. På den måde kommer det i linkets attributter til at fremgå, om husstand er brugt eller ikke brugt som argument for linket. Vi afstår fra at lave et link, når der er to eller flere lige gode kandidater til et link. Det gør vi, fordi vi prioriterer troværdige links over mange links, når vi er tvunget til at vælge.

Indtil nu har vi slet ikke brugt attributterne erhverv og bopæl, fordi det så vil være meget nemmere at linke folk, der hele livet har samme erhverv eller bor det samme sted end at linke folk, der flytter meget eller skifter erhverv. Det har forskere internationalt ønsket at undgå for at få et retvisende billede af, hvor ofte folk flytter og hvor tit de skifter erhverv, når data bruges til at lave statistik for at undersøge disse variable.

Lige nu danner alle vores regelbaserede links mellem folketællingerne stumper af livsforløb (det kalder vi *life-course frames*) for cirka halvdelen af befolkningen 1787-1901. Vi arbejder videre med at programmere andre algoritmer, der kan generere regelbaserede links fra begravelser til folketællinger. Stadig kun to kilder ad gangen: altså begravelsesregistrene i kirkebøgerne er én kilde, der skal sammenlignes med én folketælling ad gangen. Én begravet person, der er gammel nok, vil optræde i flere folketællinger. Så når vi har fået først begravelserne og derefter resten af kirkebøgerne med, vil vi have mange links per person og dermed bedre beskrevne livsforløb for en større del af befolkningen end folketællingerne kan give alene. Vi tror på, at vi kan komme op på at have regelbaserede livsforløb bestående af troværdige links på 60 - 70 % af befolkningen 1787-1901 i løbet af et års tid. Det er fint i forhold til regelbaseret linkning internationalt, men de danske kilder er så gode, at vi ikke synes det er en tilfredsstillende succesrate på længere sigt.

Derfor arbejder vi samtidig med at udvikle Machine Learning baserede links. Her bruger man links lavet af mennesker for en lille del af kildedata som træningsdata til at oplære et computerprogram til at kende forskel på links og ikke-links. Derefter lader man maskinen lave links af hele det store kildedatasæt. For at kunne bedømme kvaliteten af det maskinen laver, deler man sin pulje af menneskeskabte links i to: træningslinks og "facitliste"-links, så man kan tjekke om maskinen er blevet lige så troværdig som mennesker til dette arbejde. De foreløbige testresultater tyder på at denne metode faktisk kan blive lige så dygtig som de mennesker, der har lavet træningslinkene.

Menneskeskabte links

Erfarne historikere og slægtsforskere, der arbejder længe, hårdt og går detektivisk til værks, kan finde næsten hele deres slægt fra fødsel til død eller udvandring i folketællinger og kirkebøger. Så succesraten på danske kilder kan blive meget tæt på 100 % for mange 1800-tals danskere, når man inddrager oplysninger fra mange kilder til at bekræfte eller afkræfte fundene. Men er det muligt for historikere og slægtsforskere at finde flere troværdige links end de regelbaserede algoritmer, når de også kun må sammenligne to folketællinger ad gangen? Det spørgsmål besluttede vi at undersøge nærmere. Fem erfarne historikere og slægtsforskere linkede uafhængigt af hinanden alle beboere i det samme sogn i folketællingen 1850 til folketællingen i hele landet i 1845 efter reglerne for regelbaseret linkning. Et andet hold af fem gjorde det samme med et andet sogn 1860 til 1850. Resultatet var, at for 1850 til 1845 var alle fem slægtsforskere og historikere enige om, hvad der var det rigtige link for lidt mere end 80 % af de sognets beboere i 1850, der ville kunne linkes (potentielt linkbare = over fem år gamle i 1850). Resten havde de enten ikke fundet, eller de var uenige om, hvilket link der var det bedste. For 1860 til 1850 var de enige om lidt mindre end 80 %. Et lovende resultat, fordi 80 % er en meget høj succesrate i international sammenhæng. Men også et resultat, der viste os, at når vi laver træningslinks til Machine Learning, skal vi håndtere, at selv trænede linkere ikke når til det samme resultat for alle links, selv om de er enige om langt de fleste.

Derfor bruger vi to trænede linkere, når vi laver træningsdata til Machine Learning. De to skal linke samme data uafhængigt af hinanden. Deres links blive herefter sammenlignet, hvorefter de begge på ny tager stilling til de links, hvor

de er kommet til et forskelligt resultat. De links som de to fortsat ikke er enige om, bliver derefter gennemgået af en tredje erfaren linker, der beslutter om et af de to bud er mere troværdigt end det andet. Hvis det ikke er tilfældet, linkes ikke.

Ved linkningen af træningsdata for 23 landsogne har vi tjekket hvor ofte de to linkere havde forskellige bud på et link, og hvor høj en succesrate de tilsammen nåede efter hele proceduren med tjek og tredjepart. Ved linkningen af et sogn ad gangen fra FT 1860 til hele landet 1850 og tilsvarende for FT 1850 til 1845 var der forskellig opfattelse af, hvad der var det bedste link for mellem 4 % (Sejerø 1850-45) og 15 % (blandt andre Krønge 1860-50) af de potentielt linkbare i startfolketællingen. Resten af sognene lå derimellem. Der var, som forventet, størst uenighed om sogne, hvor mange beboere flyttede ind og ud. Linkningsprocenten efter sammenligning og tredjeparts gennemgang var mellem 70 % i det mest besværlige sogn og 98 % i det nemmeste. Resten lå derimellem med et gennemsnit på 87 %.

Hele denne omstændelige dokumentation af, hvad mennesker gør, når de linker to folketællinger ad gangen, har givet os ny viden om forskellen på regelbaserede, computergenererede links og de links, mennesker kan lave med de samme oplysninger. Nu er vi gået i gang med at undersøge forskellen på links lavet af mennesker, regelbaserede algoritmer og Machine Learning algoritmer, når vi inddrager kirkebøger og de københavnske begravelsesprotokoller. Vi forventer, at både mennesker og Machine Learning algoritmer vil kunne se mere komplicerede mønstre i livsforløb, så flere af de svære links vil kunne finde et match, og at links, vi allerede har lavet, vil kunne be- eller afkræftes, ligesom vi må være forberedt på, at nogle af de lette links vil ende med et nyt spørgsmålstegn.

Kan slægtsforskere bruge Link-Lives?

I løbet af efteråret 2021 vil vi på Link-Lives.dk være klar til at vise en test-udgave af de foreløbige livsforløbsstumper, som vores regelbaserede og Machine Learning baserede algoritmer på det tidspunkt kan præstere. Man vil kunne se og søge i både livsforløbsstumperne og de transskriberede kilder, vi har brugt, uanset om de indgår i en livsforløbsstump eller ej. Vores gæt er, at hvad angår livsforløb, vil der ikke være så meget at hente for slægtsforskere, fordi de fleste vil have bedre styr på deres egen slægt end vores algoritmer har. Algoritmerne kan ikke opnå den dybde i viden om et menneske og en slægt, som slægtsforskere opbygger. Det, algoritmerne giver, er en anden slags viden, der er simpel og rå, men som dækker hele befolkningen to kildefiler ad gangen, så vi kommer til at vide, præcis hvem, der er matchet i disse to kilder og hvem, der mangler et match.

Det betyder at de to typer viden kan komplettere hinanden og at Link-Lives kan få stor glæde af slægtsforskeres erfaring og viden, hvis de har lyst til at bidrage. Derfor er vi i gang med at udvikle en feedback funktion, så man kan logge ind og markere om algoritmerne leverer troværdige links. Hvis vi kan få det til at virke, og tilstrækkeligt mange har lyst til at bidrage, vil vi senere forsøge at finde en god måde at udvide den interaktive del på, så slægtsforskere kan lave nye links baseret på de kilder, der allerede er i systemet.

For forskerne i Link-Lives vil det være meget værdifulde bidrag, hvis mange vil deltage, fordi det vil give mulighed for at give Machine Learning algoritmerne flere typer træningsdata at arbejde med: både dem, der er lavet sogn for sogn af Link-Lives linkere og dem, der er lavet slægt for slægt. Endnu er det for tidligt at sige, hvilken og hvor stor en forskel det vil gøre, men Machine Learning eksperter ved fra erfaringer med at linke andre oplysninger, at det kan gøre en stor forskel i hvilken rækkefølge match laves. Vi vil gerne undersøge, om det også gør en forskel for livsforløb. For slægtsforskere vil der måske blive noget at hente ved at kunne sammenligne egne links med algoritmernes og med andre slægtsforskeres, der kommer til slægten fra en anden gren.

Vi tror altså ikke, at Link-Lives vil ødelægge fornøjelsen ved at lave slægtsforskning. Tværtimod kan det på længere sigt ende med, at vi sammen, historikere, slægtsforskere, universiteter og arkiver, kan opbygge en Dansk Historisk Befolkningsdatabase med links mellem mange forskellige slags kilder. Det vil være en fantastisk ressource for forskere fra mange discipliner. Men det vil også være en måde, slægtsforskere kan publicere deres forskning på: En ny type Danmarkshistorie: alle danskeres Danmarks historie, hvor mange slægters billeder, breve, dagbøger, avisudklip med videre kan linkes på relevante andre kilder til en person (for eksempel personens optræden i en folketælling eller en kirkebog) med oplysninger i linket om hvem, der har uploadet. Så vil de mange små historier tilsammen fortælle en Danmarkshistorie, der griber langt flere menneskers erfaringer end vi har kunnet skrive frem i bogform.