

Notat

Juridisk fremgangsmåde for byrådsdata

Digitaliseringsstyrelsen har i samarbejde med Open Data DK og fem kommuner udført et pilotprojekt, der har afdækket, hvordan sprogdata fra kommunale byråds- og udvalgs møder kan tilgængeliggøres til brug for udvikling af sprogmodeller. Som led i projektet har Digitaliseringsstyrelsen fået udarbejdet en analyse af de ophavsretlige og databeskyttelsesretlige forhold, som skal iagttages i forbindelse hermed. Dette notat opsamler den juridiske fremgangsmåde, som Digitaliseringsstyrelsen har benyttet i projektet. Fremgangsmåden beror desuden på en række konkrete vurderinger, som er foretaget af Digitaliseringsstyrelsen, navnlig: hjemmelsanalyse, risikovurdering, tærskelanalyse og konsekvensanalyse. Fremgangsmåden skal ikke ses som en vejledning men skal udelukkende demonstrere, hvordan Digitaliseringsstyrelsen er gået til opgaven. Fremgangsmåden kan derfor kun bruges som inspiration. Eventuel genbrug af metoden i lignende projekter bør kun ske efter en konkret vurdering af om forudsætningerne er sammenlignelige. Digitaliseringsstyrelsen opfordrer desuden anvendere til selv at sikre, at anvendelsen af datasættet sker i overensstemmelse med lovgivningen, herunder databeskyttelsesreglerne. Dertil anbefaler Digitaliseringsstyrelsen, at det overvejes, om man i anvendelsen af datasættet tilføjer nye risici for de registrerede. I dokumentationen på datasættet findes der en henvisning til den tekniske fremgangsmåde, hvor selve koden og dokumentation for løsningen kan findes.

1. Ophavsret

Et værk er beskyttet af ophavsretten, hvis det lever op til originalitetskravet, der tilsiger, at værket skal være udarbejdet under "frie og kreative valg". Hvis dette er tilfældet, og der er tale om et værk udarbejdet i et ansættelsesforhold, tilkommer ophavsretten i udgangspunktet den medarbejder, der har udarbejdet værket, medmindre andet er aftalt. Det betyder, at en kommune som organisation ikke uden videre kan dele eller videreanvende ophavsretligt materiale, som den ligger inde med, med henblik på udvikling af sprogmodeller. Derfor har pilotprojektet udelukkende fokuseret på tekster, der vurderes at være fri for ophavsret.

Det beror som udgangspunkt på en konkret vurdering, om en given tekst er fri for ophavsret. I en stikprøveanalyse foretaget før projektet, er samtlige bilag undersøgt til et tilfældigt byrådsmøde i Aarhus Kommune og et tilfældigt miljø-

og teknikudvalgsmøde i Kalundborg Kommune (samlet over 200 dokumenter). De eneste bilag, som i stikprøven er blevet anført ophavsretlige bemærkninger til er: illustrationer, fotografier, logoer, forslag til plakattekst samt rapporter udarbejdet af eksterne private organisationer. Projektet har imidlertid kun haft fokus på at arbejde med tekstdata, hvorfor grafiske elementer ikke er gjort til genstand for tilgængeliggørelse. Fra stikprøven er det derfor udelukkende rapporter udarbejdet af private organisationer, der har været anledning til at sortere fra. Dette med en forventning om, at sådanne rapporter i langt de fleste tilfælde vil være helt eller delvist beskyttet af ophavsretten og derfor bør udelades.

Der er imidlertid mange øvrige typer af bilag, som kan findes på kommunale byråds- og udvalgsmøder, og som i stikprøven umiddelbart vurderes at være fri for ophavsret. Det kan dog ikke konkluderes, at alle andre tekster end eksternt udarbejdede rapporter *altid* vil være fri for ophavsret. Den juridiske vurdering peger dog på, at tekster, som bærer præg af rutinemæssige eller rent faktuelle forhold, herunder sagsfremstillinger og referater, *i langt de fleste tilfælde* må vurderes at være fri for ophavsret. Digitaliseringsstyrelsen antager, at netop disse kvaliteter gør sig bredt gældende for den type af dokumenter, der fremføres på åbne byrådsmøder og åbne miljø- og teknikudvalgsmøder, hvilket stikprøveanalysen underbygger. Dertil kommer, at der på hhv. byrådsmøder og miljø- og teknikudvalgsmøder drøftes emner, som er afgrænsede til kommunal forvaltning eller underområder hertil. Derfor forventes materialet på disse møder at være af forholdsvis ensartet karakter på tværs af kommunerne, hvorfor der ikke forventes at være væsentligt større spændvidde i typen af bilag, som fremføres på sådanne møder, end hvad stikprøverne har vist.

I første dataudgivelse arbejder Digitaliseringsstyrelsen kun med referaterne fra byråds- og miljø- og teknikudvalgsmøderne. For disse dokumenttyper har projektet taget udgangspunkt i en generel vurdering af, at der kun indgår materiale, som i langt de fleste tilfælde er fri for ophavsret, og som dermed vurderes egnet til tilgængeliggørelse med henblik på brug til udvikling af sprogmodeller. Digitaliseringsstyrelsen arbejder på en proces for at finde en løsning til at frasortere ophavsretligt materiale, således at bilagene fra møderne også kan blive efterbehandlet og udstillet.

2. Databeskyttelsesret

De tekstdata, der indgår i projektet, indeholder personoplysninger. Selvom personoplysningerne allerede ligger offentligt tilgængelige, gælder databeskyttelsesreglerne stadig. Nedenfor listes de relevante databeskyttelsesretlige forhold, som analysen og Digitaliseringsstyrelsen har peget på, samt hvordan disse er blevet håndteret i projektet.

- **Behandlingsgrundlag for almindelige personoplysninger:** Digitaliseringsstyrelsen har identificeret et behandlingsgrundlag for tilgængeliggørelsen af kommunale byråds- og udvalgsdata i databeskyttelsesforordningens artikel 6, stk. 1, litra e om behandling af personoplysninger når det er nødvendigt af hensyn til udførelse af en opgave i samfundets interesse. Digitaliseringsstyrelsen har i den forbindelse lagt vægt på, at projektet udspringer af Den Fællesoffentlige Digitaliseringsstrategi 2022-2025 og Strategisk indsats for kunstig intelligens. Ydermere har projektet budgetmæssig hjemmel via finansloven¹. Slutteligt er der lagt vægt på, at projektet ikke vurderes at være væsentlig indgribende for de registrerede, særligt henset til, at de anvendte data allerede er offentliggjorte.
- **Behandlingsgrundlag for følsomme personoplysninger:** I en stikprøveanalyse foretaget forud for projektet, er der identificeret følsomme personoplysninger i form af politisk overbevisning, hovedsageligt i form af kommunalpolitikeres partipolitiske tilhørsforhold. Ud fra stikprøven har Digitaliseringsstyrelsen vurderet, at denne type oplysninger tydeligvis er offentliggjort af de registrerede selv, fx i forbindelse med opstilling på valglisten. Følsomme personoplysninger, som tydeligvis er offentliggjort af den registrerede selv, må gerne lovligt behandles ifølge databeskyttelsesforordningens artikel 9, stk. 2, litra e.

Der er derudover blevet identificeret ét tilfælde af fagforeningsmæssigt tilhørsforhold, hvilket Digitaliseringsstyrelsen vurderer ikke *tydeligvis* er offentliggjort af de registrerede selv, og som der derudover ikke vurderes at være hjemmel til at behandle med henblik på udstilling til brug for udvikling af sprogmodeller. Det kan dermed ikke udelukkes, at lignende eller øvrige typer af følsomme personoplysninger kan findes i tekstmaterialet fra andre byråds- og udvalgmøder.

Det er derfor besluttet, at projektet kun må behandle følsomme personoplysninger i det omfang disse tydeligvis er offentliggjort af de registrerede selv. Dette har indebåret, at der er brugt ressourcer på at screene for andre følsomme personoplysninger, og at tekster, hvori sådanne indgår, er blevet frasorteret. Til det formål er der blevet udviklet en automatiseret løsning for screening og frasortering.

Frasorteringen udgør i juridisk forstand også en behandling, og Digitaliseringsstyrelsen vurderer, at databeskyttelsesforordningens artikel 9, stk. 2, litra g i den forbindelse kan benyttes som behandlingsgrundlag. Databeskyttelsesforordningens artikel 9, stk. 2,

¹ Tekstanmærkning nr. 104 og 105 ad § 25.12.02 til § 25 på finansloven for finansåret 2024.

litra g indebærer, at følsomme personoplysninger kan behandles, hvis det er nødvendigt af hensyn til væsentlige samfundsinteresser på grundlag af EU-retten eller medlemsstaternes nationale ret. Det vurderes, at behandlingen vil være nødvendig af hensyn til væsentlige samfundsinteresser, jf. databeskyttelseslovens § 7, stk. 4.

Hvis nogen derudover skulle blive opmærksomme på, at der alligevel forekommer følsomme personoplysninger, som ikke tydeligvis er offentliggjort af den registrerede selv, vil Digitaliseringsstyrelsen håndtere det som en utilsigtet hændelse, og det skal herefter sikres, at tekster hvori disse oplysninger indgår, ikke bliver gjort til genstand for tilgængeliggørelse til brug for udvikling af sprogmodeller. Såfremt dataene allerede er blevet gjort til genstand herfor, vil de pågældende tekster fjernes fra det udstillede datasæt og anvendere af datasættet vil blive underrettet om hændelsen via metadata for datasættet.

- **Formålsforenelighed:** Digitaliseringsstyrelsen vurderer med afsæt i Datatilsynets vejledning om Offentlige myndigheders brug af kunstig intelligens, at formålet med behandlingen i projektet er foreneligt med det formål, som dataene oprindeligt blev indsamlet til. I vurderingen lægges der særligt vægt på, at personoplysningerne allerede er offentliggjorte, hvormed det må være forventeligt for de registrerede, at personoplysningerne efterfølgende kan viderebehandles til nye formål, idet oplysningerne ved offentliggørelsen er tilgængelige for enhver. Ydermere vurderes behandlingen ikke at indebære en væsentlig højere risiko for de registrerede end den offentliggørelse, kommunerne allerede har foretaget. Slutteligt er personoplysningerne blevet pseudonymiseret, hvormed sandsynligheden for identificering er mitigeret.
- **Risikovurdering, tærskel- og konsekvensanalyse:** Digitaliseringsstyrelsen har udarbejdet en risikovurdering og en tærskelanalyse for behandlingen af de pågældende personoplysninger. Digitaliseringsstyrelsen har desuden udarbejdet en konsekvensanalyse, idet tærskelanalysen har givet anledning hertil, da der er anvendt nye teknologier i behandlingen og formålet med behandlingen er at tilvejebringe træningsdata til udvikling af yderligere nye teknologier. Analyserne peger på, at risici og konsekvenser forbundet med behandlingen er lave, efter at de mitigerende foranstaltninger er indført, særligt taget i betragtning at personoplysningerne pseudonymiseres, at følsomme personoplysninger frasorteres, samt at oplysningerne allerede findes offentligt tilgængelige på internettet.
- **Oplysningspligt:** Forud for behandlingen har Digitaliseringsstyrelsen oplyst de registrerede om den nye behandling af deres

personoplysninger. Oplysningsteksten indeholder identiteten og kontaktoplysninger på den dataansvarlige, kontaktoplysninger på databeskyttelsesrådgiveren, formålet med behandlingen, behandlingsgrundlaget, de berørte kategorier af personoplysninger, den kilde, oplysningerne er hentet fra, modtagere af oplysningerne, kriterier, der anvendes til at fastlægge tidsrummet for opbevaring, retten til at klage til Datatilsynet samt retten til at udøve sine rettigheder.

I analysen er det noteret, at hvis der kan argumenteres for, at det vil være uforholdsmæssigt vanskeligt at finde frem til de enkelte registrerede, kan oplysningspligten opfyldes ved at massekommunikere via en hjemmeside. Digitaliseringsstyrelsen har vurderet, at der i projektet behandles data af en karakter og i et omfang, hvor det dels er en uforholdsmæssig vanskelig opgave at afdække, hvilke personer, der er registreret, og at det, selv derefter, kan være uforholdsmæssigt vanskeligt, hvis ikke umuligt, at finde kontaktoplysninger på disse personer. Digitaliseringsstyrelsen har derfor håndteret oplysningspligten via massekommunikation på Digitaliseringsstyrelsens hjemmeside, som kan findes [her](#).

- **Håndtering af direkte personidentifikatorer:** For at overholde dataminimeringsprincippet må der ikke behandles personoplysninger, som ikke er nødvendige for formålet med behandlingen. Behandlingen i projektet har været tredelt, idet der 1) indsamles personoplysninger via scraping af de fem kommuners hjemmesider, 2) foretages en teknisk efterbehandling af data og 3) foretages en udstilling af data på ubestemt tid. Udstillingen i sig selv udgør dermed en databehandling, og Digitaliseringsstyrelsen har i denne konkrete sammenhæng vurderet, at direkte personidentifikatorer som navne, e-mailadresser og telefonnumre ikke skal indgå i de data, der udstilles. Dette skyldes, at tilgængeliggørelse med henblik på videreanvendelse til brug for udvikling af sprogmodeller ikke forudsætter, at datasættet indeholder direkte personidentifikatorer.

Dataminimeringsprincippet lægger derfor til grund, at der, *om muligt*, skal foretages en anonymisering af dataene, førend de udstilles. Digitaliseringsstyrelsen er imidlertid nået til den konklusion, at data, der allerede er offentligt tilgængelige, ikke kan anonymiseres effektivt, idet der forholdsvis let vil være mulighed for at re-identificere de registrerede ved brug af den offentlige adgang til de oprindelige data. Der vil alligevel stadig være behov for at tage dataminimerende foranstaltninger i brug, såfremt dette er muligt, hvorfor analysen peger på, at der skal foretages en pseudonymisering forud for udstilling af dataene. Pseudonymisering indebærer, at personoplysninger erstattes med pseudonymer med

henblik på at mindske muligheden for at identificere fysiske personer. Pseudonymiserede oplysninger udgør stadig personoplysninger, fordi det ved brug af supplerende oplysninger fortsat er muligt at re-identificere fysiske personer. Derfor skal de databeskyttelsesretlige regler fortsat iagttages, selvom oplysningerne er blevet pseudonymiseret.

Digitaliseringsstyrelsen har derfor erstattet direkte personidentifikatorer med et pseudonym med henblik på at mindske risikoen for, at sprogmodeller under træning på de pågældende data præsenteres for autentiske direkte personidentifikatorer. Forud for pseudonymiseringen er der lavet en undersøgelse af, hvilke typer af direkte identifikatorer, der indgår i tekstmaterialet.

- **Håndtering af quasi-identifikatorer:** Quasi-identifikatorer, dvs. indirekte identifikatorer så som pronomener, ansættelsessted eller stillingsbetegnelser, er i udgangspunktet ikke blevet pseudonymiseret. Valget om at beholde quasi-identifikatorer beror på et fagligt hensyn om at bevare semantisk sammenhæng og kontekst i et givet dokument, da disse kvaliteter bidrager væsentligt til sprogmodellers præstationer. Dertil kommer, at givet at der er tale om allerede offentligt tilgængelige oplysninger, vil dataminimeringsforanstaltninger ikke hindre mennesker i at skaffe sig adgang til de pågældende personoplysninger. Foranstaltningerne skal derimod blot sikre, at de sprogmodeller, der trænes på dataene, ikke bliver præsenteret for autentiske direkte personidentifikatorer.
- **Tidsopmærkning:** For at imødekomme databeskyttelsesforordningens rigtighedsprincip er der i metadataopmærkningen for de udstillede data blevet indført tidsopmærkning, således at anvendere af datasættet kan forholde sig til, om dataene er aktuelle eller forældede i forhold til de formål, de måtte blive videreanvendt til.
- **Databehandleraftale:** Forud for brug af en ekstern leverandør til behandling af personoplysninger (indsamling, frasortering af følsomme personoplysninger, pseudonymisering), er der udarbejdet en databehandleraftale.