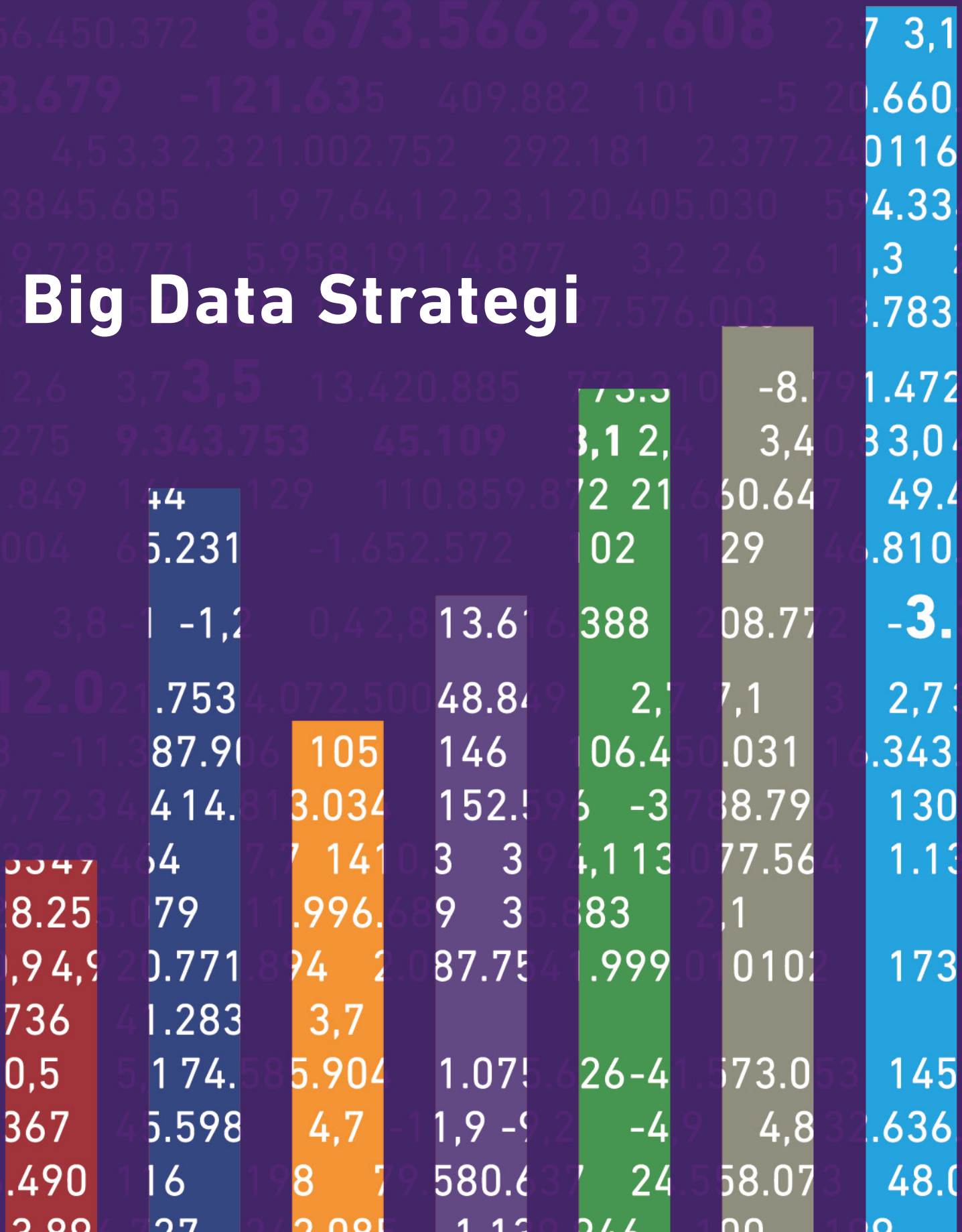
DANMARKS STATISTIK

Big Data Strategi



Big Data Strategi 2018-2020

DANMARKS STATISTIK

Big Data Strategi 2018-2020

Udgivet af Danmarks Statistik
Oktober 2018
Forside: Grafik af Danmarks Statistik

Pdf-udgave:
Kan hentes gratis på www.dst.dk/publ/BigDataStratDK
ISBN 978-87-501-2305-7

Adresser:
Danmarks Statistik
Sejrøgade 11
2100 København Ø

Tlf. 39 17 39 17

e-mail: dst@dst.dk
www.dst.dk

Indledning

Big Data er et resultat af den stadigt stigende digitalisering, der betyder at såvel borgere som virksomheder afsætter elektroniske 'fodaftryk' i forbindelse med en lang række mere eller mindre almindelige og hverdagsagtige handlinger. Det sker via aktiviteter på internettet (indkøb i meget bred forstand, transport, sociale medier, medieforbrug i øvrigt og registreringer af egne forhold fx helbred og sportsaktiviteter) og forretningsmæssige transaktioner (køb, salg af varer og tjenester og transport af varer). Anvendelsen af digitale sensorer i målere (fx elmålere) og maskiner i bred forstand (fx transportmidler og landbrugsredskaber) (det såkaldte 'internet og things') er en yderligere bølge i skabelsen af Big Data.

Big Data adskiller sig fra andre kendte datakilder (administrative data og survey data) på en række områder, i forhold til mængde (stor), kilder (flere forskellige), hastighed (høj – der skabes data hele tiden), forskellighed (i kilder og deres struktur), og troværdighed (af kilder ifht. et givet formål).

Big Data kan være (en del af) svaret på en række udfordringer som officiel statistik står overfor som faldende svarprocenter, reducerede økonomiske rammer og ønsket om hurtigere statistik

Danmarks Statistisk Strategi 2022 fastslår, at der i strategiperioden skal udarbejdes en handlingsplan for udnyttelse af Big Data, og at der skal samarbejdes med producenter af Big Data om anvendelse af data i statistikproduktionen (Danmarks Statistisk, Strategi 2022:11).

Udmøntningen af Danmarks Statistiks Big Data Strategi vil især have fokus på anvendelsen af data i relation til eksisterende statistikker, og på at indgå i data-partnerskaber med andre med henblik på at berige de eksisterende kilder til den officielle statistik.

Denne strategien er et bidrag til det, og beskriver de strategiske satsninger om Big Data på følgende områder:

- Erfaring med Big Data i Danmarks Statistik
- Kompetencer og kompetenceudvikling
- De juridiske aspekter
- Partnerskaber
- Internationalt samarbejde
- Samt hvordan arbejdet med dette organiseres

Erfaringer med Big Data i Danmarks Statistik

Danmarks Statistik har allerede en vis erfaringen med brugen af Big Data.

Brugen af *stregkodedata* som input til forbrugerprisindekset er det eneste eksempel på anvendelsen af Big Data i statistikproduktionen. Og et eksempel der viser de udfordringer og behov for grundige undersøgelser, der er ved anvendelsen af Big Data i statistikproduktionen. Arbejdet blev igangsat i 2010 og førte til, at stregkodedata indgik i produktionen af forbrugerprisindekset fra 1. januar 2016. I den periode blev der gennemført grundige undersøgelser og trukket på erfaringer fra andre lande før den endelige model for anvendelsen af de nye data var på plads.

AIS-data(Automatic Identification System) er digitale meddelelser om positioner på alle skibe i danske farvande[1]. Danmarks Statistik har deltaget i arbejdsgruppe under ESSnet Big Data, hvor anvendelsen af AIS-data blev undersøgt. AIS-data kan fx være supplerende datakilde til Passager- og Færgefart, og være input til Grønt Nationalregnskab (emissioner af CO2 og NOx) og Turismestati-stik.

Elmålerdata fra energinet.dk. I 2020 skal alle elmålere i Danmark (og resten af EU) være af smartmeter-typen, der sender forbruget til leverandøren hver 15. minutter. Dette giver mulighed for at følge elforbruget meget detaljeret. Danmarks Statistik deltager i arbejdsgruppe under ESSnet Big Data, hvor anvendelsen af SmartMeterdata er undersøgt. Elmålerdata kan fx være datakilde til mere detaljeret energistatistik samt til boligstatistikken.

Webscraping – indsamling af data direkte fra nettet. I 2016 undersøgte Danmarks Statistik muligheden for at integrere webscraping i dataindsamlingen for statistikken for ledige stillinger. Den gang viste der sig en række problemer, som muligvis siden er løst. Bl.a. har Eurostats ESSnet Big Data gennem de seneste år, arbejdet målrettet med at hente data fra nettet og forsøgt at løse de problemer, de har mødt undervejs. Webscraping anvendes primært som kilde til kvalitetssikring af data, hvor de primære data kommer fra en anden kilde.

Betalingskortdata - indeholder betalingskorttransaktioner enten foretaget i fysiske terminaler eller over internettet. Et vigtigt formål med at få adgang til betalingskortdata er, for Danmarks Statistik, at forbedre udgiftssiden i betalingsbalances rejsepost da dette er en notorisk svær post at opgøre. Der er dog en bred vifte at potentiale anvendelsesmuligheder for betalingskortdata. Indtil videre er det lykkedes at få adgang til et testdatasæt, der dækker en periode på 1½ år.

Og desuden er der en række initiativer i relation til indberetning til især erhvervsstatistik om anvendelse af *Digital revisor* til regnskabsstatistikken, *Sensor-data* til transport og landbrugsstatistik, *data fra platformsøkonomien* som en yderligere automatisering af virksomhedernes indberetninger af data til Danmarks Statistik. Desuden findes der yderligere Big Data datakilder, som kan undersøges nærmere.

[1] Alle skibe over en vis størrelse (>300 bruttotons, alle passagerskibe samt alle fiskeskibe over 15 meters længde) skal være udstyret med en AIS-transponder.

Indsats

- Foretag en analyse af aktuelle og fremtidig brug af Big Data datakilder til eksisterende og ny officiel statistik herunder en konceptualisering af forskellige former af brug af Big Data kilder – som fx erstatning for eksisterende datakilder, som supplement til eksisterende datakilder som basis for hurtigere eller ny statistik

- Udbygge brugen af *stregkodedata* til en fuld dækning af supermarkedskæder og til at dække andre områder fx tankstationer samt foretage analyse af andre mulige anvendelsesområder

- Afprøvning af brugen af *AIS-data* på et eller flere konkrete statistiksområder fx grønt nationalregnskab og havnestatistikken

- Udbygge erfaringen med brugen af *elmålerdata* med henblik på at kunne anvende dem i en konkret statistikproduktion med data fra 2020 og frem fx boligstatistikken

- Genbesøge muligheden for at anvende *webscrabing* til kvalitetssikring af oplysninger om ledige stillinger fra den offentlige sektor med inddragelse af erfaringer fra andre statistikbureauer bl.a. det hollandske og det slovenske.

- Afdække mulighederne for at anvende *betalingskortdata* på en række eksisterende og nye statistikområder som fx betalingsbalancens rejsepost og internethandel ved brug af de eksisterende testdata

- Afdække mulighederne for øget automatisering at private virksomheders inkl. landbrugets indberetninger til Danmarks Statistik ved at indgå frivillige datapartnerskaber med leverandørerne af systemløsninger til den private sektor

- Undersøge mulighederne for anvendelse af yderligere datakilder fx mobiltelefondata og data (fx rejsekort) om offentlig transport til 'trængselsstatistik', data fra sociale medier til statistik om befolkningens 'stemning' og markdatabasen til landbrugsstatistik

Kompetencer og kompetenceudvikling

Danmarks Statistiks procesmodel[2] finder mest direkte anvendelse på traditionelle survey-baserede statistikker, men kan også anvendes til at beskrive statistikker baseret på Big Data.

Kompetencebehovet vedrørende Big Data handler om, hvordan man kommer fra rådata i ofte ukendte eller besværligt håndterbare formater til brugbare fakta og viden, der kan uddrages og formidles. Udover de rent tekniske udfordringer kræves der også stor viden om hvordan data kan anvendes, idet den datagenererende proces sjældent er velbeskrevet. Fx er det ikke trivielt at beskrive en relevant undersøgelsespopulation.

For det praktiske kompetenceudviklingsbehov i Danmarks Statistik arbejder vi med roller og kompetenceniveauer for at adressere forskellige medarbejdertyper.

Der er to hovedroller: Statistikansvarlige/statistikmedarbejdere samt it-medarbejdere og ansatte i IT. En sekundær rolle er ledelsen, som skal have et grundlæggende niveau af forståelse for at kunne forholde sig forretningsmæssigt til Big Data-mulighederne.

Kompetenceudviklingen vedrørende Big Data gennemføres primært internt. Der er kompetente medarbejdere såvel i IT som i Metode og Analyse, som vil levere relevant undervisning og oplæring suppleret med eksterne ressourcer til få, udvalgte områder (eksempelvis machine learning).

Det tværgående samarbejde er en nødvendig kompetence og disciplin, hvis et Big Dataprojekt skal lykkes. Specielt inden et konkret dataområde baseret på Big Data er driftsmodent og publicerings- eller formidlingsklart vil det kræve samarbejde på tværs af statistikkontorerne, Metode og IT med fokus både på IT kompetencer og kompetencer i statistikkontorerne til fortolkning af data fra Big Data kilder

Desuden er networking både internt og eksternt relevant og værdifuldt. Det er meget givtigt at følge med i andre statistikbureauers fremdrift på Big Dataområdet og dialogen med nordiske og internationale kolleger giver basis for værdifuld sparring og viden.

Det vurderes, at behovet for værktøjer og teknologisk kapacitet indtil videre er dækket med R-miljøet i samspil med de eksisterende Oracle og SAS platforme.
Nye værktøjer og kapacitet skal primært drives af konkrete behov i samspil med de generelle overvejelser der er i Metode og IT om fremtidens værktøjsportefølje.

[2] En tilpasning af den generiske GSBPM – Generic Statistical Business Process Model.

Indsats

- Skabe overblik over de konkrete værktøjer og udviklingen af brugen af disse værktøjer, der anvendes i de statistiskbureauer, der er længst fremme i brugen af Big Data til officiel statistik

- Etablere et overblik over kompetencer i statistikkontorer og IT og en plan for kompetenceudvikling for anvendelsen af Big Data til officiel statistik

- Konkretisere kompetenceudviklingsplanerne i Danmarks Statistiks reviderede IT-strategi med fokus på kompetencebehov til at løfte udfordringen ved anvendelse af Big Data i produktionen af officiel statistik

- Afdække behovet for kompetencer i relation til udviklingen og anvendelsen af nye former for lagring af meget store datamængder

- Organisere kompetence-partnerskaber på tværs af huset og ud af huset (digital task forces) til hurtigt (sprint) at udvikle og afprøve ideer på strategisk udvalgte indsatsområder

De juridiske aspekter

Dette afsnit beskæftiger sig med mulighederne for ad lovgivningens vej at pålægge private virksomheder (dataejere) at afgive Big Data til Danmarks Statistik til statistiske formål.

Der er i den forbindelse to spor. Der kan enten arbejdes på at få en national lovgivning, der sikrer dette nationalt, eller en EU-forordning, der så vil gælde alle medlemsstater – en kombination af disse vil også være en mulighed.

Hvis der arbejdes for en løsning via national lovgivning, kan man arbejde videre med det grundlag, der søgtes skabt i forbindelse med forarbejderne til revisionen af lov om Danmarks Statistik.

Mulighederne via den nationale lovgivning er imidlertid sat på stand-by efter, at loven er vedtaget uden et hjemmelsafsnit, der giver mulighed for at stille krav ved indsamlingen af Big Data.

Det spor, der derfor ligger for, er hjemmel via en EU-retsakt, forordning (som har direkte retskraft i alle medlemsstater), direktiv (som skal implementeres i medlemsstaternes egen lovgivning) eller via EU-softlaw, som ikke er umiddelbart bindende (fx forskellige former for aftaler).

I forhold til at bruge lovgivningsinstrumentet til at gøre indberetning af Big Data til statistiske formål obligatorisk skal man være opmærksom på, at der siden 2017 er sket et kraftigt skift i offentlighedens bekymring for sikkerheden af borgernes digitale data – både hos offentlige og private dataejere. Det skyldes bl.a. diskussionerne i forhold til implementeringen af EU-databeskyttelsesforordningen og eksempler på usikker omgang hos det offentlige med borgernes data.

Anbefalingen kan derfor være i forbindelse med et eventuelt lovgivningsinitiativ at skabe sikkerhed for, at de Big Data, som indsamles af de nationale statistikinstitutioner til statistiske formål enten har en karakter, så de ikke kan anvendes til andre formål – heller ikke, hvis de videregives ulovligt, samt at forskningsanvendelsen begrænses, så situationer med misbrug ikke kan ske. Det vil sige, at der skal arbejdes med den ret brede formulering i den danske databeskyttelseslovs § 10, stk. 2, hvor det blot hedder, at oplysninger, der er behandlet til statistiske eller videnskabelige formål, ikke siden må behandles til andre formål. Det vil sige, at det skal fremgå af den pågældende lovgivning, hvordan den snævre anvendelse sikres.

Indsats

- Udarbejde et notat baseret på internationale erfaringer om forhold omkring de juridiske aspekter af officiel statistiks adgang til anvendelse af Big Data i produktionen af officiel statistik

- Deltage i udviklingen omkring adgang til Big Data til statistiske formål i det internationale statistiske system (Eurostat og FN)

- Deltage aktivt i debatten med indlæg og arrangementer, der viser konkrete eksempler på det samfundsmæssige potentiale, der ligger i at sikre statistik- og forskningsinstitutioner adgang til Big Data på måder, så borgernes og virksomhedernes data er sikret.

Partnerskaber

Danmarks Statistik har, som det fremgår, indgået partnerskab med en række supermarkedskæder om modtagelse og anvendelse af scanner data som input til prisstatistikken. Der ud over har der været sporadiske kontakter til andre potentielle leverandører af Big Data bl.a. på mobiltelefoniområdet.

Der har også været kontakt til erhvervsorganisationer (DI og Dansk Erhverv) om Big Data og til den akademiske verden (KU, DTU og ITU) bl.a. om etableringen af en master-uddannelse i data science som overbygning på de samfundsvidenskabelige bacheloruddannelser på Københavns Universitet.

Endelig har der været kontakt til Microsoft Danmark om brugen af Big Data, ligesom der har været kontakt til og besøg fra Microsoft i Seattle med en drøftelse af potentielle samarbejdsprojekter.

FN's arbejdsgruppe om Big Data har som en del af sin strategi udviklet en god kontakt til større tech-virksomheder på det globale niveau (Microsoft, Google, Amazon og Nielsen) med henblik på at indgå et samarbejde om Big Data projekter, der er nyttige og relevante for udarbejdelsen af officiel statistik ved brug af Big Data.

Der er næppe tvivl om, at der uanset om der via lovgivning opnås adgang til brugen af Big Data til produktionen af officiel statistik eller ej, så skal udviklingen på området ske i form af partnerskaber.

Partnerskaber kan have flere formål ud over leverance af Big Data til officiel statistik. Det kan være partnerskaber om gensidig udveksling af data, hvor data fra den private dataleverandør indgår i Danmarks Statistiks produktion af officiel statistik, og hvor data fra Danmarks Statistik anvendes til at berige en private dataleverandørs datakilder, det kan være partnerskaber om kompetenceudvikling og om lagring af data, og det kan være partnerskaber, hvor Danmarks Statistik får indsigt i og indgår som partner i udviklingen af nye måder at anvende data på hos den private dataleverandør, samt partnerskaber hvor Danmarks Statistisk, som det er tilfældet med data fra Energinet, stiller data til rådighed for forskere og analytikere.

Indgåelse af partnerskaber såvel når det gælder datakilder som teknologi og kompetenceudvikling er en central del af strategien på området. En tilgang der understøttes af bestræbelserne i FN's arbejdsgruppe om at indgå partnerskaber på globalt niveau til støtte af de enkelte statistikbureauers arbejde på området.

I Danmarks er det også en mulighed at prøve at etablere data-partnerskaber med leverandører af 'dataindberetningssystemer' til den private sektor, og med erhvervsorganisationer om anvendelsen af nye måder at indberette data på.

Indsats

- Der udarbejdes en præsentation, der viser potentielle partnere, Danmarks Statistiks potentiale som Big Data partner med fokus på Danmarks Statistiks særlige styrker omkring dataanvendelse, datadokumentation og datadeling

- Kontakten til potentielle Big Data leverandører (herunder Erhvervsorganisationerne) systematiseres med henblik på en dialog om potentialet i anvendelsen af Big Data til officiel statistik

- Kontakten til Akademia og tech-virksomhederne systematiseres med en undersøgelse af potentialet i at etablere et 'advisory board' for brugen af Big Data til officiel statistik med deltagelse af bl.a. KU, DTU, ITU og danske tech-virksomheder.

- Der gennemføres en samtalerunde (tech-lunches) med udvalgte Big Data leverandører, Akademia og tech-virksomheder som led i planlægningen af Danmarks Statistiks arbejdsplan for 2019

- Muligheden for at indgå frivillige digitale partnerskaber med systemleverandørere på udvalgte områder undersøges og konkretiseres

- Muligheden for at gennemføre konkrete projekter med Microsoft undersøges i relation i Danmarks Statistiks arbejdsplan for 2019

Internationalt samarbejde

FN's statistikkomite (UNSC) etablerede på sit 45. møde i 2014 en arbejdsgruppe til fremme af brugen af Big Data i officiel statistik (Global Working Group on Big Data for Official Statistics). Arbejdsgruppen, som Danmarks Statistik har været formand for siden 2016, har etableret en række teams, der har beskæftiget sig med brugen af udvalgte datakilder (satellitdata, mobiltelefondata, scannerdata og social medier data) som input til produktionen af officiel statistik. Desuden har arbejdsgruppen siden 2014 arrangeret en årlig konference om Big Data for Official Statistics med deltagelse både af statistikbureauer og private teknologivirksomheder, samt etableret en base med eksempler på brugen af Big Data i officiel statistik[3].

Eurostat har ligeledes arbejdet med udviklingen af brugen af Big Data i officiel statistik gennem en årrække bl.a. ved etableringen af en styregruppe på området, som Danmark er medlem af. For statistikbureauerne er den mest konkrete del af Eurostats arbejde med Big Data etablering af et såkaldt ESSnet, der som et samarbejde mellem en række lande (herunder Danmark) afprøver mulighederne for anvendelsen af konkrete Big Data kilder som input i produktionen af officiel statistik. Eurostat har desuden gennem styregruppen igangsat et arbejde der skal afprøve mulighederne for på europæisk niveau at få et juridisk grundlag for adgangen til Big Data til produktionen af officiel statistik. Endelig arbejder Eurostat også med kompetenceudvikling på området bl.a. ved afholdelse af kurser i forskellige data-værktøjer, der kan anvendes til organisering og analyse af Big Data.

Både i relation til arbejdet i FN og EU er der på meget nyttig vis tale om at etablere og fremme mulighederne for, at de enkelte landes statistikbureauer i fællesskab kan få erfaringer med og udvikle kompetencer blandt medarbejderne til at anvende Big Data i produktionen af officiel statistik.

Endelig er der nogle få statistikbureauer, der er særligt langt fremme i arbejdet med Big Data. I Europa gælder det Nederland og UK og til dels Estland, mens det udenfor Europa især er Australien og Canada. Muligheden for at lære af dem og muligheden for at arbejde sammen med dem skal inddrages i Danmarks Statistiks arbejde med udviklingen af anvendelsen af Big Data i officiel statistik.

[3] Man kan finde yderligere oplysninger om gruppens arbejde på https://unstats.un.org/bigdata/.

Indsats

- Skabe og vedligeholde overblik over medarbejdernes deltagelse i Big Data relevante aktiviteter internationalt

- Indgå i konkrete projekter i anden runde af Eurostats ESSnet om Big Data til officiel statistik

- Indgå i arbejdet i relevante task teams i FN's arbejdsgruppe om Big Data til officiel statistik

- Få konkret indblik i arbejdet med Big Data i officiel statistik i de lande, der er længst fremme på området bl.a. via kontakter i FN's arbejdsgruppe om Big Data.

- Arrangerer en konference med international deltagelse med fokus på et dansk publikum om Big Data, Data Science og statistik som opfølgning på den konference Danmarks Statistik afholdte i efteråret 2016 i samarbejde med Københavns Universitet og Dansk Industri

Organisering

Arbejdet med Big Data som kilde til officiel statistik er som arbejdet med andre datakilder tværgående, og involverer såvel statistikafdelingerne som IT. Desuden er der behov for ledelsesmæssige initiativer vedrørende arbejdet med adgang til Big Data (de juridiske aspekter og partnerskaber) og finansieringskilder til det samlede Big Data arbejde i Danmarks Statistik samt koordinering mellem de internationale initiativer i Eurostat og FN og det konkrete arbejde i Danmarks Statistik.

Der er med andre ord flere aspekter vedrørende organisering og finansiering, der skal udredes og besluttes

Derfor etableres der i august 2018 et projekt med støtte fra Porteføljesekretariatet, der:

- Foretager en analyse af aktuel og fremtidig brug af Big Data datakilder til eksisterende og ny officiel statistik herunder en konceptualisering af forskellige former for brug af Big Data kilder – som fx erstatning for eksisterende datakilder, som supplement til eksisterende datakilder som basis for hurtigere eller ny statistik, samt skitsere en ramme for beskrivelse af Big Data kilders kvalitet

- Udarbejder en præsentation, der viser potentielle partnere, Danmarks Statistiks potentiale som Big Data partner – og på den baggrund

- Gennemfører en samtalerunde (tech-lunches) med udvalgte Big Data leverandører, akademia og tech-virksomheder som led i planlægningen af Danmarks Statistiks arbejdsplan for 2019

Og kommer med konkrete forslag til opfølgning på følgende indsatser i arbejdsplanen for 2019:

- Organisere kompetence-partnerskaber på tværs af huset og ud af huset (digital task forces) til at udvikle og afprøve ideer på strategisk udvalgte indsatsområder

- Udarbejde et notat baseret på internationale erfaringer om forhold omkring de juridiske aspekter af officiel statistiks adgang til anvendelse af Big Data i produktionen af officiel statistik

- Deltage aktivt i debatten med indlæg og arrangementer, der viser gode eksempler på det samfundsmæssige potentiale, der ligger i at sikre statistik- og forskningsinstitutioner adgang til Big Data

- Arrangerer en konference med international deltagelse med fokus på et dansk publikum om Big Data, Data Science og statistik som opfølgning på konferencen i 2016

DANMARKS STATISTIK

Danmarks Statistik
Sejrøgade 11
2100 København Ø

Tlf. 39 17 39 17
www.dst.dk
dst@dst.dk