Statistisk sentralbyrå
Statistics Norway

Datatilstander SSB

2. utgave

TALL SOM FORTELLER

NOTATER / DOCUMENTS

2024/44

Standardutvalget

I serien Notater publiseres dokumentasjon, metodebeskrivelser, modellbeskrivelser og standarder.

Publisert: 28. oktober 2024

ISBN 978-82-587-1053-7 (elektronisk)
ISSN 2535-7271 (elektronisk)

Standardtegn i tabeller	Symbol
Ikke mulig å oppgi tall Tall finnes ikke på dette tidspunktet fordi kategorien ikke var i bruk da tallene ble samlet inn.	.
Tallgrunnlag mangler Tall er ikke kommet inn i våre databaser eller er for usikre til å publiseres.	..
Vises ikke av konfidensialitetshensyn Tall publiseres ikke for å unngå å identifisere personer eller virksomheter.	:
Desimaltegn	,

Forord

Det er behov for å ha en felles beskrivelse av de ulike datatilstandene som inngår i statistikk-produksjon i SSB. En felles forståelse og terminologi vil være en hjelp i det kontinuerlige arbeidet med å forbedre statistikkproduksjonen, bidra til etterprøvbarhet og tilrettelegge for økt gjenbruk av data.

Statistisk sentralbyrå, 1. oktober 2023

Arvid Olav Lysø

Sammendrag

Notatets formål er å beskrive de fem datatilstandene i SSB, og kravene som stilles til dokumentasjonen av dem.

Innhold

1. Innledning

Målet med dette notatet er å bidra til en tydeligere beskrivelse av datatilstander i statistikk-produksjonen i SSB. Beskrivelsene er konkrete slik at de kan være til hjelp i det kontinuerlige arbeidet med å forbedre statistikkproduksjonen, bidra til etterprøvbarhet og tilrettelegge for økt gjenbruk av data.

FNs prosessmodell, Generic Statistical Business Process Model (GSBPM)[1] beskriver produksjonsprosessen for offisiell statistikk. Den beskriver og definerer alle prosesser og del-prosesser, fra arbeidet som gjøres før data samles inn, og til det endelige statistikkproduktet er formidlet.

Fra et datasett foreligger i SSB i sin minst bearbeidede form, vil det, som et resultat av de ulike stegene i produksjonsprosessen, endres. En datatilstand er et resultat av at et datasett har gått gjennom gitte operasjoner og prosesser. Veldefinerte datatilstander er viktige for å oppnå:

- standardiserte og gjenkjennbare produksjonsløp
- etterprøvbare statistikker
- intern og ekstern gjenbruk av data
- Identifisering av data som er arkivverdige etter Lov om arkiv (arkivlova).

Det er naturlig at hovedfokuset i SSBs kvalitetsarbeid[2] er rettet mot statistikkene. Samtidig har både forventninger og krav til SSBs evne til å dele data, økt betydelig de senere år. Det betyr at vi i tillegg til å produsere hovedproduktet «statistikk», vil ha økt fokus på å produsere gjenbrukbare datasett med høy kvalitet. En viktig forutsetning for gjenbruk er at de som vil bruke dataene, kan forstå dataene og vite hvilke endringer de har gjennomgått. Bruk av data i SSB, og gjenbruk i og utenfor SSB, krever gode metadata. Definisjoner av datatilstander og andre statistikkbegreper, må derfor i størst mulig grad være avstemt med internasjonale statistiske rammeverk og definisjoner.

Begrepet «etterprøvbarhet» brukes flere steder i notatet. Ideelt sett bør vi produsere statistikk slik at ettertiden eller en uavhengig instans med tilgang til dataene og vår dokumentasjon, vil komme til samme statistiske resultater som oss selv.

Datatilstandene som beskrives er *kildedata, inndata, klargjorte data, statistikk og utdata*. De tre første tilstandene, er i hovedsak mikrodata som gir informasjon om enkeltenheter, mens statistikk og utdata i hovedsak er aggregerte data. For å beskrive datatilstandene på en god måte, trenger vi også begrepet «startdata». Dette er ikke en datatilstand, men er mer knyttet til selve produksjonsprosessen. Startdata er de dataene en starter klargjøringsprosessen med. Det vil si at startdata kan bestå av egen datatilstand inndata, samt ulike data som er delt fra andre i SSB, f.eks. andres klargjorte data. Noen statistikker vil ha startdata som kun består av egne inndata, andre vil ha startdata som kun består av delte data fra andre i SSB, mens noen vil ha startdata som både består av egne inndata og delte SSB-data. For å gjøre det enklere å dele og gjenbruke data, foreslår dette notatet at datatilstandene er stabile og følger av at et **datasett** tilfredsstiller gitte kriterier, dvs. at en datatilstand ikke er avhengig av hvilke prosesser datasettet brukes i. Et datasett som har fått status som (tilfredsstiller kravene til) «klargjorte data»[3], vil fortsette å være klargjorte data selv om datasettet brukes som startdata for en annen statistikk.

[1] Generic Statistical Business Process Model GSBPM. (Version 5.1, January 2019). Norsk oversettelse (ssb.no)
[2] Statistikkloven §5 Krav til offisiell statistikk: statistikken skal være relevant, nøyaktig, aktuell, punktlig, tilgjengelig og klar, sammenlignbar og sammenhengende.
[3] Se avsnitt 2.3, variabler er beregnet og nøyaktigheten forbedret.

I tillegg til hovedtilstandene, defineres noen viktige begreper som statistikkprodukt og kvalitetsindikator som er nært knyttet til datatilstander og statistikkproduksjon. Notatet beskriver hvilke metadata som er de mest relevante for hver tilstand, og gir anvisning for når dataeiers variabelnavn skal benyttes og når SSB-definerte variabelnavn skal benyttes.

1.1. Obligatoriske datatilstander

Dette notatet gir overordnede beskrivelser av alle datatilstandene som inngår i statistikk-produksjonen i SSB. Samtidig fastsettes det at tilstandene kildedata, klargjorte data, statistikk og utdata er obligatoriske. Dette innebærer at de som er ansvarlig for en gitt statistikk, også er ansvarlig for at disse tilhørende og obligatoriske datasettene produseres, dokumenteres[4], langtidslagres og tilgjengeliggjøres i henhold til tilstandsbeskrivelsene i dette notatet. I den forbindelse er det viktig å understreke at «dokumenteres og langtidslagres», spesielt for utdata, ikke betyr at all dokumentasjon nødvendigvis skal lagres på Dapla[5]. For eksempel kan utdata for (tabell)oppdrag, være dokumentert i Websak. Disse utdataene behøver da ikke «dobbelt-dokumenteres» på Dapla.

Det anbefales at de som er ansvarlige for statistikkene, jevnlig vurderer om det er etterspørsel og behov også for inndata. Særlig for datasett som inneholder persondata, kan det være aktuelt å dokumentere og langtidslagre inndata siden dette er den første datatilstanden som inneholder pseudonymiserte data.

[4] Datasettene vil dokumenteres i en egen løsning. I skrivende stund (2023) arbeides det med å implementere en slik løsning på Dapla.
[5] Dapla er SSBs skybaserte dataplattform.

2. Datatilstander

2.1. Kildedata

Kildedata er data lagret slik de ble levert til SSB fra dataeier, det vil si på dataeiers dataformat og datamodell, samt med informasjon om tidspunkt og rekkefølge for avlevering. Kildedata kan inneholde personlig identifiserbar informasjon og andre sensitive opplysninger. De kan også være ustrukturerte eller strukturerte. Kildedata er en del av statistikkenes dokumentasjon, og kan være en nødvendig kilde til forskning og for å lage nye statistikker. Uten at kildedataene er tilgjengelige, vil det ikke være mulig å etterprøve SSBs statistikker. De originale kildedataene vil ofte komprimeres og krypteres for lagring etter at de delene av datasettet som skal brukes videre, er transformert til inndata.

Kildedata, særlig de som hentes inn via API, kontrolleres for å avdekke eventuelle datafeil slik at dataeier kan kontaktes. Aktuelle kontroller kan være frekvens (data kommer ikke/kommer sjeldnere enn forutsatt), omfang (færre variabler enn forutsatt), volum (betydelig færre observasjoner enn ventet) og innhold (f.eks. andel null-verdier mye større enn forventet eller verdiene åpenbart ulogiske).

Eksempler på kildedata er:

- **Grunndata** (ofte kalt masterdata): Data som endrer seg relativt sjelden og som ofte anvendes i mange ulike sammenhenger. Basisdata om personer (Folkeregisteret), virksomheter (Enhetsregisteret) og steder (Matrikkelen), er eksempler på grunndata.
- **Transaksjonsdata**: Data som reflekterer at utveksling av informasjon eller hendelser har funnet sted. Banktransaksjoner, kvitteringsdata og lønningsdata, er eksempler på transaksjonsdata.
- **Administrative data**: Data som er sammenstilt, primært av offentlige myndigheter, for bred anvendelse, som rapporterings- eller skatteformål. Administrative data kan, i likhet med transaksjonsdata, reflektere at en utveksling eller hendelse har funnet sted. Tolldeklarasjoner og a-ordningen, er eksempler på administrative data.
- **Statistiske data:** Data som er samlet inn med produksjon av statistikk som hovedformål. Spørreundersøkelser/skjemadata brukes der vi ikke har etablert maskinell innsamling eller der det ikke finnes data med nødvendig frekvens, aktualitet, detaljeringsnivå eller avgrensing som er nødvendig for statistikkproduksjonen. De brukes også for å følge europeiske forordninger, som i noen tilfeller krever datainnsamling med en gitt skjemautforming.
- **Aggregerte data og rapporter**: Data som dataeier har bearbeidet før SSB mottar dem. Dette innebærer at deler av produksjonsprosessen er utført utenfor SSB, og at SSB ofte ikke har innsyn i alle algoritmene eller prosessene som er benyttet i fremstillingen av dataene.

Data som oppstår hos dataeier, vil ofte utgjøre en form for dokument (slik som for eksempel en kassakvittering). For å minimere mulighet for feil i databehandlingen hos dataeier, bør vi tilstrebe å fange disse dataene

- så nær som mulig kilden der de oppstår
- mest mulig ubearbeidet[6] (kopi av originaldokument)
- så nær som mulig tidspunktet hvor dataene oppstår.

[6] Med «ubearbeidet» menes at dataene ikke har vært utsatt for omfattende koblinger, transformasjoner eller aggregeringer. Data hvor dataeier har korrigert feil og gjør enkle uttrekk for SSB, vil fortsatt kunne ansees som ubearbeidet i denne sammenheng.

I tilfeller hvor det er behov for data utover originaldokumentet, kan det være nødvendig å hente data fra flere kilder eller å hente data som er bearbeidet av dataeier.

I tilfeller hvor SSB mottar bearbeidede data, som for eksempel administrative data fra registre, er det viktig med et samarbeid med dataeierne slik at dataeierne selv tar et ansvar for kvalitetssikring og feilretting før SSB tar imot data.

Aktuelle metadata
Informasjon på datasettnivå slik som dataeier, hvilket område dataene omhandler og tidsinformasjon, er relevante metadata for kildedata. Metadata om enkeltvariabler følger av, og er begrenset til, den informasjonen dataeier selv avleverer.

2.2. Inndata

Inndata er hovedsakelig kildedata som er transformert til SSBs standard lagringsformat[7] og ev. omstrukturert[8]. Inndata er altså data som er hentet inn fra eksterne og tilpasset statistikkens formål. Disse eksterne dataene vil, ev. sammen med delte data fra andre i SSB, utgjøre en statistikks startdata. Noen makrostatistikker vil imidlertid ikke ha inndata, kun startdata som består av delte data fra andre i SSB. Andre makrostatistikker vil bygge på en blanding av inndata (egne innsamlede data) og delte data fra andre.

Inndata har blitt behandlet gjennom ulike stadier av dataminimering, pseudonymisering, omkoding, omstrukturering og omformatering. Inndataene skal ikke inneholde personlig identifiserbar informasjon, og skal være strukturerte og lagret i standardiserte formater og systemer. For mange statistikker vil det også være validering[9] knyttet til inndata. Dette betyr at variabelnavn og -innhold i inndata er uendret fra slik de er i kildedata, bortsett fra at

- dataene er minimert slik at kun variabler som er nødvendige i den videre produksjonsprosessen, inngår.
- direkte identifiserende variabler (f.eks. fødselsnummer) er pseudonymisert
- det benyttes standard kodeverk (Klass[10]) der det er mulig (f.eks. kjønn)
- dataene kan være omstrukturert og tilpasset statistikkformålet
- tegnsett, datoformat, adresse mm er endret til SSBs standardformat

Selv om dataene minimeres, vil det ofte inngå flere variabler i inndata enn de som framgår i den endelige statistikken. Årsaker til dette er bl.a. at

- vi ønsker å kunne foreta kontroll, feilkorrigering, imputering, vekting, analyse og lignende operasjoner for å forbedre kvaliteten, lage statistikken og andre estimater (f.eks. av usikkerhet)
- det kan være behov for deling til andre statistikker og til forskning
- det er behov for å kunne videreutvikle og forvalte kode effektivt i takt med økt kunnskap og endringer i datakilder.

Statistikkloven og personvernfordningen (GDPR) bestemmer at alle data som samles inn og behandles skal «… være adekvate, relevante og begrenset til det som er nødvendig for formålene de behandler (dataminimering)»[11]. For å tilfredsstille kravet til adekvate og relevante data, må det

[7] Eksempelvis tegnsett UTF8, datoformat YYYY-MM-DD. S
[8] Ofte kan det være hensiktsmessig å omstrukturere dataene ihht prinsippene for «Tidy Data» (hver observasjon er en rad, hver variabel er en kolonne, hver celle er en verdi), se https://vita.had.co.nz/papers/tidy-data.pdf
[9] Maskinell validering der valideringsrapporter sendes tilbake til oppgavegiver som deretter sender inn dataene på ny.
[10] https://www.ssb.no/klass/
[11] Kapittel II, Artikkel 5, pkt 1c

samles inn nok variabler til at vi kan gjennomføre nødvendige kontroller og analyser, slik at vi får god kvalitet i våre produkter. Dette vil, som for andre typer data, oftest innebære behov for flere variable enn de som brukes direkte i beregninger for å lage statistikken. Samtidig må det sikres at vi ikke samler inn data som ikke er nødvendige for formålet dataene skal brukes til.

Aktuelle metadata
Det er i utgangspunktet de samme metadataene som vil være aktuelle for inndata, som for kildedata, men dersom inndata skal deles, kan det være behov for flere eller andre metadata. I tillegg kan det være aktuelt å supplere med metadata som gir informasjon om bearbeidingen som er gjort i SSB.

2.3. Klargjorte data

Klargjorte data er inndata hvor

- variablene er beregnet gjennom utregninger og koblinger mellom datasett
- nøyaktigheten er forbedret gjennom kontroll av dataenes gyldighet og korrigerende tiltak i form av for eksempel filtrering, editering eller imputering
- metadata med variabeldefinisjoner er lagt til.

I klargjorte data er det som regel *ikke* foretatt aggregeringer. Dette innebærer at klargjorte data i hovedsak reflekterer enkeltobservasjoner, på lik linje med kildedataene. Dersom kildedataene SSB mottar er aggregater, vil de klargjorte dataene ofte være aggregert på samme nivå som kildedataene. I klargjorte datasett er det SSB som står for variabelnavn og variabeldefinisjoner.

Det er et mål at klargjorte data i størst mulig grad skapes automatisk og ved bruk av algoritmer, uten manuelle operasjoner. Endringene som er gjort i data, skal uansett være sporbare og dokumentert på en slik måte at statistikkene blir etterprøvbare.

I klargjorte datasett er det gjort konkrete valg for hvordan feil og mangler i inndataene skal håndteres, og hvordan klargjorte variabler etableres. Valgene vil kunne være forskjellige for forskjellige statistikker. Når klargjorte data skal gjenbrukes til andre formål enn det dataene opprinnelig ble klargjort for, må det derfor gjøres en konkret (ny) vurdering om hvorvidt det er nødvendig å supplere med nye eller gjøre andre endringer i dataene.

Merk at det skal ikke gjøres endringer på inndata som følge av klargjøringsprosessen. Dette innebærer at inndata som er korrigert for feil og mangler i klargjøringen, lagres som klargjorte data, og ikke som inndata.

Data kan også være klargjort for delingsformål[12]. Slike data skal tilfredsstille kravene i dette dokumentet og er ellers gjenstand for samme kvalitetskrav som klargjorte data for statistikkformål[13].

Klargjorte data fra andre statistikker
I tillegg til data som samles inn direkte for en gitt statistikk, vil det ofte være behov for ulike typer klargjorte data fra andre statistikker[14] og støttedata i produksjonsprosessen slik som for eksempel

- populasjonsinformasjon fra et av populasjonsregistrene
- utvalgsinformasjon

[12] Eksempler er forløpsdata som NUDB (Norsk utdanningsdatabase) og FD-Trygd (Forløpsdatabasen for Trygd).
[13] Retningslinjer for Europeisk statistikk
[14] Klargjorte data fra andre statistikker kan både komme fra andre statistikker i SSB og fra eksterne kilder som Eurostat.

- klargjorte data fra andre statistikker
- vekter beregnet fra andre statistikker.

Selv om disse dataene allerede er klargjort gjennom populasjonsforvaltning eller andre statistikkers klargjøringsprosesser, vil det ofte være behov for tilpasning og transformasjon når de skal integreres med øvrige data i en gitt statistikkproduksjon, typisk i klargjøringsfasen for den gitte statistikken. Makrostatistikken er et eksempel på at flere kilder og /eller klargjorte data settes sammen til en statistikk.

Statistikker som benytter klargjorte data fra andre, er ansvarlige for å ta vare på informasjon om hvilke datasett som er brukt, slik at produksjonen kan etterprøves.

Versjoner
Avhengig av om klargjøringsprosessen involverer manuelle operasjoner, eller om klargjorte data kan gjenskapes i sin helhet fra kildedata ved å kjøre programmer, vil det være behov for å lage og lagre ulike versjoner av klargjorte data. Disse skal navngis og versjoneres i henhold til «Navnestandard og versjonering av datasett i DAPLA»[15].

Aktuelle metadata
For klargjorte data er dette variabeldefinisjoner, det vil si beskrivelse av hver enkelt variabel og hvordan den er beregnet. I tillegg trengs dokumentasjon av hvilke nøyaktighetsforbedrende tiltak som er utført og hvorfor de er gjort.

2.4. Statistikk

Statistikk er data som faller inn under definisjonen av statistikk i statistikkloven: «Tallfestede opplysninger om en gruppe eller et fenomen, og som kommer frem ved en sammenstilling og bearbeidelse av opplysninger om de enkelte enhetene i gruppen eller et utvalg av disse enhetene, eller ved systematisk observasjon av fenomenet.» (Statistikkloven § 3a).

Statistikk vil som oftest være aggregerte data eller estimerte størrelser. De enkelte variablene i statistikk er definert av SSB og er gitt navn av SSB i tråd med gjeldende praksis. Disse er ikke nødvendigvis sammenfallende med variabelnavnene og -definisjonene i klargjorte data.

Noen ganger følger statistikken av de klargjorte dataene ved enkle aggregeringer (ujustert statistikk). Andre ganger vil det benyttes statistiske metoder for å estimere tall basert på utvalgsdata og beregnes indekser, kalenderjusterte tall, sesongjusterte tall og trendserier (justert statistikk) i tillegg.

Statistikk og klargjorte data deles internt ved at de inngår i andre statistikkers startdata, særlig makrostatistikkenes, og kan da inneholde konfidensielle og detaljerte data som ikke publiseres.

Aktuelle metadata
Som for klargjorte data, er variabeldefinisjoner, det vil si beskrivelse av hver enkelt variabel og hvordan den er beregnet, aktuelle metadata for statistikk. I tillegg skal det dokumenteres hvilke metoder og programmer/kode som er benyttet for å produsere datatilstanden.

[15] Internt dokument i SSB

2.5. Utdata

Utdata er statistikk der kravene til konfidensialitet er ivaretatt. Det er denne datatilstanden som publiseres av SSB selv eller utleveres til andre. Regler og mekanismer for utlevering av for eksempel kildedata eller klargjorte data, faller utenfor rammene for dette notatet.

Transformasjonen fra statistikk til utdata kjennetegnes ved at

- konfidensielle data er undertrykt ved prikking av celler, aggregering eller andre metoder som ivaretar krav til konfidensialitet
- kvaliteten er vurdert å være god nok for publisering i henhold til krav om kvalitet i offisiell statistikk[16]
- eventuelle krav til sampublisering er ivaretatt. Dette kan for eksempel være samtidig publisering av ujusterte, kalender- og sesongjusterte tall.

Eksempler på utdata er:

- statistikkbanktabeller
- tabelloppdrag
- Internasjonal rapportering.

Versjoner
Dersom publiserte tall revideres etter publisering, skal alle versjoner tas vare på.

Aktuelle metadata
Aktuelle metadata for utdata er som for statistikkdata. I tillegg skal det dokumenteres hvilke programmer eller kode som er brukt for å lage produktet. Som nevnt i kap. 1.1, vil ikke dokumentasjonen av alle utdata nødvendigvis lagres på Dapla, men arkiveres i systemer som for eksempel Websak.

2.6. Konfidensialitet

Vær oppmerksom på at mange konfidensialitetsmetoder tar klargjorte data som utgangspunkt, og ikke statistikk. Ofte produseres da mange statistiske tabeller samtidig, slik at samordnet konfidensialitet er sikret. Resultatet av en slik metode er konfidensielle utdata, og også de ubeskyttede statistikkdataene. Det har da ingen hensikt å produsere utdata i et eget trinn.

2.7. Arkivering

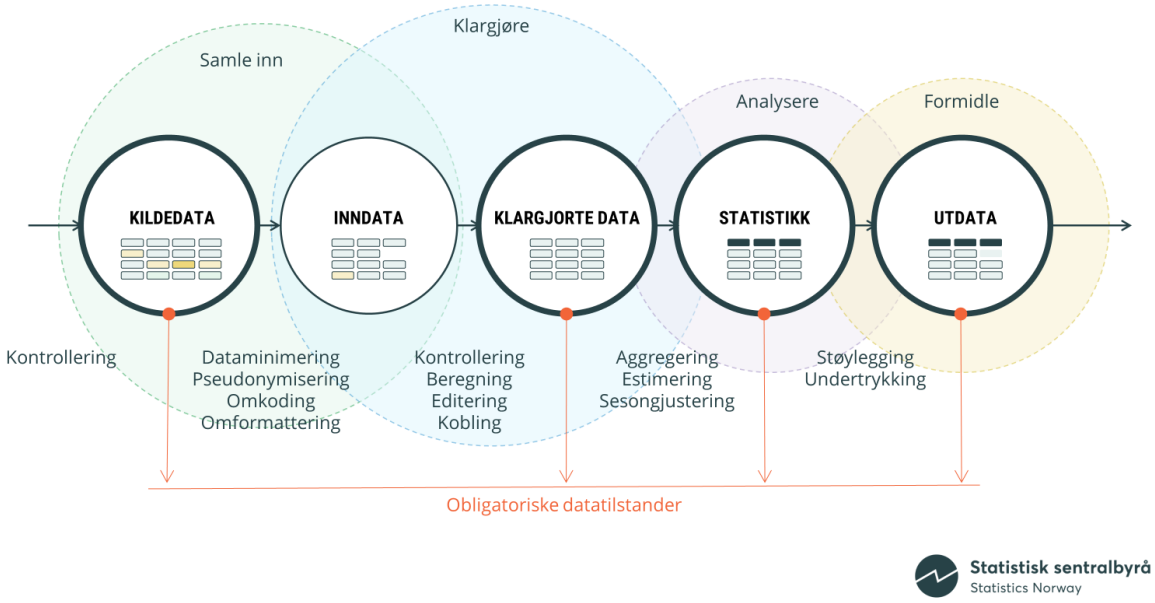
Når det gjelder arkivverdige filer/datasett og avlevering eller deponering til Arkivverket, er dette et område som er under diskusjon i SSB. Både avlevering/deponering[17], og muligheten for at SSB selv kan oppbevare arkivmaterialet[18], er en del av diskusjonen. Det er klargjorte data som anses som arkivverdige.

[16] https://www.ssb.no/omssb/kvalitet-i-offisiell-statistikk
[17] Paragraf 16 i Arkivlova: Ved avlevering går eigedomsretten til arkivet over til mottakarinstitusjonen. Ved deponering har deponenten og seinare ervingane hans eigedomsretten til arkivet. Eigedomsretten går likevel over til mottakarinstitusjonen når erveføLgda etter deponenten vert broten, eller når det er gått hundre år sidan deponeringa.
[18] Paragraf 10 i Arkivlova: Statlege arkiv skal avleverast til Arkivverket i samsvar med dei føresegnene som vert fastsette i medhald av § 12 i denne lova. Riksarkivaren kan likevel gje samtykke til at statlege arkiv vert avleverte til institusjonar utanfor Arkivverket, eller at dei framleis skal oppbevarast av det arkivskapande organet

2.8. Datatilstander oppsummert

Figur 2.1 oppsummerer de ulike datatilstandene og hvilke hovedtransformasjoner som utføres for å bringe dataene fra en tilstand til en annen.

Figur 2.1. Datatilstander og hovedtransformasjoner



Tilstandene kildedata, klargjorte data, statistikk og utdata er obligatoriske, noe som betyr at tilhørende datasett produseres, dokumenteres, langtidslagres og tilgjengeliggjøres i henhold til tilstandsbeskrivelsene.

Det trengs imidlertid noen overordnede kommentarer for å nyansere hva som er obligatoriske datatilstander. I «Navnestandard og versjonering av datasett på Dapla»[19] opererer vi med begrepene «dataprodukter»[20] og «statistikkprodukter»[21]. Dokumentasjonskravene til disse vil variere. For det vi omtaler som «dataprodukter», eksempelvis FREG, BOF, a-ordningen, FD-Trygd og NUDB[22], vil kun kildedata og klargjorte data være obligatoriske datatilstander. Dataproduktene er i hovedsak mikrodata i form av registerinnsamlinger/populasjonsregistre som klargjøres for «sekundærbruk», dvs. at de inngår i andre statistikkprodukter og/eller forskning.

For statistikkprodukter som samler inn egne data (fra for eksempel registre eller via spørreundersøkelser), er kildedata, klargjorte data, statistikk og utdata obligatoriske.

[19] En intern standard for navngiving og versjonering av datasett.
[20] Ikke alle data i SSB kan knyttes direkte til en statistikk i Statistikkregisteret. Data bearbeides til andre bruksområder og formål, f.eks. klargjøring av data til forskning og utlån, bearbeiding av data som skal inngå andres statistikker, og data som inngår i populasjonsregistre. Disse kalles dataprodukter.
[21] Alle SSBs tidligere og nåværende statistikkprodukter inngår i Statistikkregisteret. Før publisering på ssb.no må alle statistikkprodukter være registrert i Statistikkregisteret med informasjon om bl.a. statistikkens navn, emne-område, eier og publiseringstidspunkt. I tillegg får statistikkene tildelt et kortnavn.
[22] hhv Folkeregisteret, Bedrifts- og foretaksregisteret, a-ordningen, Forløpsdatabase for trygd og Norsk utdanningsdatabase

For makrostatistikker som nasjonalregnskap, miljøregnskap, etc. er i mange tilfeller kun klargjorte data, statistikk og utdata obligatoriske, i og med at de kun bruker andres klargjorte data (eventuelt statistikk/utdata) og ikke henter inn kildedata (inndata).

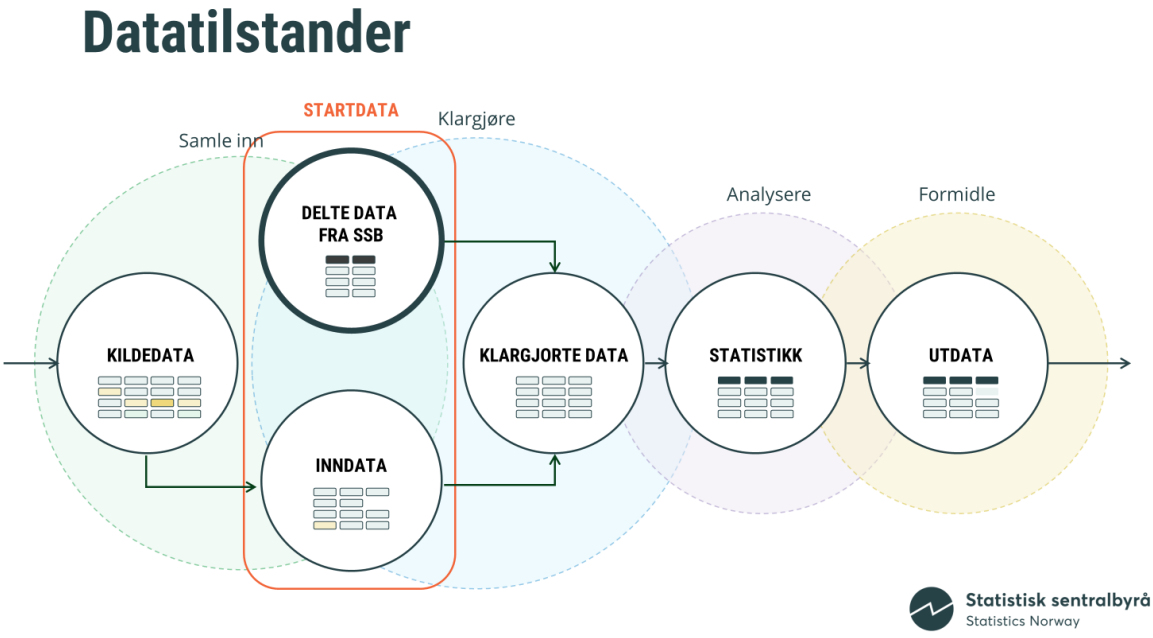
Dette vil si at kun klargjorte data er obligatoriske for alle produksjonsløp, mens inndata alltid er valgfritt. Alle statistikkansvarlige bør imidlertid vurdere om ikke også inndata skal lagres. Inndata kan være en viktig datatilstand når det gjelder deling, fordi det er den første datatilstanden med pseudonymiserte data. Det er også den tilstanden som viser hvilken dataminimering som er gjort på kildedataene.

Kvalitetsindikatorer skal etableres og følges opp for alle datatilstander. En må derfor kunne hente data som kan brukes til å beregne kvalitetsindikatorer, fra alle lagrede datatilstander. Eksempler på kvalitetsindikatorer er «andel ugyldige data», «enhetsfrafall» og «editeringsandel».

I figuren under vises forholdet mellom startdata og datatilstandene. Startdata er som tidligere nevnt, knyttet til prosess, og er de dataene en starter klargjøringsprosessen med. "Delte data fra SSB" er data som er delt av andre i SSB, det være seg deres inndata, klargjorte data, statistikk eller utdata.

For noen statistikker vil startdata kun består av egne inndata, andre vil ha startdata som kun består av delte data fra andre i SSB, mens noen vil ha startdata som både består av egne inndata og delte SSB-data.

Figur 2.2. Startdata og datatilstander



3. Andre definisjoner

3.1. Statistikkprodukt

Et statistikkprodukt er statistikk som alene eller sammen med andre statistikker beskriver en gruppe eller et fenomén og er formidlet slik at det gir mening for brukerne. Ulike målgrupper har ulike behov. For noen brukere kan dette bety en tabell for raskt å slå opp de ferskeste tallene, for andre brukere kan en graf som viser både siste tall og utviklingen over tid, gi mer innsikt. For utforsking kan en interaktiv modell være det som er mest nyttig for brukerne. Ofte vil ulike former for tekstlig beskrivelse og analyse være del av et statistikkprodukt.

I «Navnestandard og versjonering av datasett på Dapla», beskrives statistikkprodukt slik: Alle SSBs tidligere og nåværende statistikkprodukter inngår Statistikkregisteret. Før publisering på ssb.no må alle statistikkprodukter være registrert i Statistikkregisteret med informasjon om bl.a. statistikkens navn, emne-område, eier og publiseringstidspunkt. I tillegg får statistikkene tildelt et kortnavn.

3.2. Prosessdata

Prosessdata er informasjon om hendelser på dataene og oppstår underveis i produksjonsprosessen i alle datatilstander. Alle hendelser på og endringer i data logges, med informasjon om: Når (tidspunkt for hendelsen / endringen), hvem endret data eller trigget hendelsen, hva (ny, endret, slettet), hvordan (manuelt, automatisk) og gjerne også hvorfor (rekontakt, faglig skjønn etc). Prosessdata inngår ofte, sammen med metadata og data, som grunnlag for å skape kvalitetsindikatorer. Et eksempler på dette er kvalitetsindikatoren for «andel data editert», fordelt på manuell og automatisk editering.

3.3. Kvalitetsindikator

Kvalitetsindikatorer[23] er numeriske størrelser eller statistikk, som bidrar til å belyse kvaliteten i data. Kvalitetsindikatorer kan være enkle opptellinger, eller resultat av beregninger eller analyser. Ulike deler av produksjonsprosessen kan ha ulike kvalitetsindikatorer. Slike indikatorer bør etableres, beregnes og følges opp for alle datatilstandene. Det er utarbeidet en oversikt over anbefalte kvalitetsindikatorer i SSB. Det er behov for å arbeide videre med kvalitetsindikatorer i SSB. Det er få statistikker som har implementert slike indikatorer i produksjonen, og det er behov for å samle erfaringer med bruk av de anbefalte indikatorene.

Eksempler på kvalitetsindikatorer er:

- Enhetsfrafall: Andel enheter ikke rapportert i forhold til antall enheter i utvalget som skal rapportere. Rapporteres til finansdepartementet på aggregert nivå.
- Editeringsandel: Andel verdier editert i forhold til antall mulige verdier.
- Potensiell aktualitet: Antall dager fra statistikken er klar til publisering til den faktisk blir publisert. Formålet med indikatoren er å kunne gi en pekepinn på om det er rom for å forbedre aktualiteten på statistikken.

3.4. Pseudonymisering

Datatilsynet definerer pseudonymisering som avidentifisering av personopplysninger slik at de ikke kan knyttes til en bestemt person uten bruk av tilleggsopplysninger (for eksempel en koblingsnøkkel) som lagres adskilt og tilstrekkelig sikkert. Pseudonymiserte personopplysninger er ikke anonyme.

[23] Begrepet «produksjonsindikatorer» har også vært brukt tidligere om enkelte kvalitetsindikatorer.