Interne notater

STATISTISK SENTRALBYRÅ

86/8 29. januar 1986

BYRÅETS LAGRING AV DATA

Av

Johan-Kristian Tønder

Innhold

1. Innledning

I juli 1983 startet en stor debatt i svenske media om måten folke- og boligtellinger skulle gjennomføres på i framtida. Debatten utvikla seg til å bli en generell debatt om personvern og spesielt om forholdet mellom personvern og bruk av administrative data lagra på EDB-medium, i offentlig statistikk. Norske aviser ble oppmerksom på debatten i nabolandet, og stilte spørsmål om hvordan forholdet mellom administrative data og statistikk var i Norge, særlig med hensyn til folke- og boligtellingene. Debatten i Norge ble kortvarig, og stort sett foregikk den i en nøktern tone. Den viste likevel at det var en del uro hos politikere og publikum. Den tekniske utviklinga har ført til hjelpemidler som delvis øker og delvis reduserer personvernproblemene ved bruk av EDB på persondata.

Statistisk Sentralbyrå ønsket å sette i gang en utredning om disse forhold i forbindelse med spørsmålet om alternative planer for gjennomføringen av en folke- og boligtelling i 1990. Den generelle delen av denne utredninga skulle derfor omfatte følgende emner:

- Byråets adgang til administrative registre
- Byråets lagring av egne og andres data
- Regler for kobling og gjenbruk av data i Byrået

Dette notatet skal ta for seg Byråets lagring av data. Det vil innskrenke seg til lagring av data på EDB-medium. Vi ser altså bort fra lagring av skjema, lister og andre hjelpemidler for formidling og lagring av data. Ikke fordi dette ikke er viktig nok, men fordi bakgrunnen for at lagringsspørsmålet ble tatt inn i utredninga, var Byråets lagring og bruk av data på EDB-medium. Videre vil vi konsentrere oss om persondata og opplysninger som kan knyttes til personer.

2. Dagens regler for lagring og sikring av data i Byrået

Etter at Lov om personregistre m.m. (Personregisterloven) trådte i kraft 1. januar 1980, er reglene for Byråets lagring av persondata regulert av den konsesjonen som Datatilsynet har gitt Byrået 2. april 1981. Reglene er i første rekke gitt ved interne rundskriv som bygger på konsesjonsvilkårene.

Rundskrivet Prl 1 (RDr/VFr, 20.7.81) sier i pkt. 4.1 at Byråets personregistre bør anonymiseres så snart som mulig. Byrået har tolket dette slik at de bør anonymiseres så snart det statistiske formålet med å ha identifikasjonskjennemerker (fødselsnummer) knyttet til opplysningene

er nådd. For å kunne lage statistikk over overganger mellom ulike stadier i et menneskes liv ("livsløpsanalyser"), beholder en fødselsnummeret på en fil for hver årgang av et register. Andre filer (mellomprodukter, koblinger o.l.) blir slettet eller anonymisert etter at arbeidet med statistikkårgangen er avsluttet. Anonymiseringen foregår ved at en sletter navn, adresse og fødselsnummeret, eller erstatter det siste med et serienummer uten identifiserende kjennetegn.

Systemet for datasikring har vært under omarbeiding, og det vil her bli gitt en kort presentasjon av hvordan det nye systemet fungerer. Foreløpig gjelder systemet bare for data som er lagret i Byråets IBM-maskiner. Etter hvert vil dette gjelde alle Byråets data.

Hvert enkelt datasett i Byrået "eies" av det fagkontor som har ansvaret for statistikken på det feltet datasettet primært skal dekke (f.eks. eies utdanningsdataene i Byrået av det kontoret som har ansvaret for utdanningsstatistikken). Lederen av fagkontoret har ansvaret for de datasettene kontoret eier. Datasett som brukes til flere kontors statistikker, eies av det kontor som har kontakten med den som leverer registeret til Byrået.

Alle personer i Byrået som skal ha tilgang til ett eller annet datasett, må ha et passord. Passordet skal ikke være kjent for andre personer, og det må derfor underskrives en taushetserklæring om dette. Passordet endres automatisk hver måned.

Lederen for det fagkontor som eier datasettet bestemmer hvilke personer som skal ha tilgang til det. Det gjøres på et særskilt skjema som sendes SSB's driftskontor (se vedlegg 1). Driftskontoret administrerer sikkerhetssystemet, og gir maskinen beskjed om hvilke personer (passord) som blir godtatt som brukere på det enkelte datasett.

Lederne ved fagkontorene må vurdere hvem som bør få tilgang til de ulike datasettene utifra de arbeidsoppgavene vedkommende er tiltenkt. Det skal altså ikke være noen automatisk tilgang til alle kontorets datasett for alle kontorets ansatte.

Den samme vurderingen gjelder selvsagt også dersom personer ved andre kontor ønsker tilgang til datasettene for selv å kunne kjøre statistikk el.l. fra dem. Generelt kan en si at det skal foreligge et prosjektskriv godkjent av Adm. direktør, der bruken av vedkommende datasett er spesifisert. En dokumentasjon av dette, eventuelt i form av et kjørediagram, skal foreligge før adgang til datasettene blir gitt til de personer som er ansvarlig for driften av prosjektet, enten personene befinner seg på et fagkontor, ei forskningsgruppe eller på Systemkontoret eller Driftskontoret.

Byråets datasett kan deles inn i fire typer: Produksjonsdatasett, testdatasett, brukerdatasett og katalogdatasett.

Produksjonsdatasett er datasett som brukes til produksjon av statistikk og analyse.

Testdatasett kan være utdrag av andre typer datasett eller det kan være spesielt oppbygde datasett med det formål å teste programmer som så skal brukes til bearbeiding av datasett eller produksjon av statistikk og analyse.

Brukerdatasett er datasett som opprettes for en person ved fagkontor/forskningsgruppe. Det kan være programmer som personen har laget, publikasjonsmanus o.l., men det kan også være et utdrag eller en forenklet utgave av et produksjonsdatasett. Et brukerdatasett skal i prinsippet ikke inneholde data fra et produksjonsdatasett. Det er idag ikke lagt inn tekniske hindringer for at dette kan skje. Slike tekniske hindringer vil kunne føre til tidstap i databehandlingen, og Byrået overlater derfor til den kontorleder som eier produksjonsdatasettene å sikre at medarbeiderne ikke lagrer produksjonsdata i et brukerdatasett.

Katalogdatasett er datasett som angir standardiserte inndelinger og kataloger som brukes i Byrået, f.eks. kommunenummerkatalog, postnummerkatalog.

I sammenheng med datasikkerhet, er det bare produksjons-, test- og brukerdatasett som er av interesse, ettersom det bare er disse som inneholder personopplysninger som en trenger å sikre utifra personvernhensyn.

Den enkeltes bruk av et datasett kan deles inn i fire grupper:

1) Personen kan utføre alle slags operasjoner på datasettet (produsere tabeller, lese opplysninger, endre opplysninger eller fjerne opplysninger).
2) Personen kan se på hvilke opplysninger som står på datasettet.
3) Vedkommende kan skrive (endre) datasettet.
4) Vedkommende kan fjerne datasettet.

Det blir også fastsatt for hvor lang tid personen har denne tilgangen til datasettet.

Personer ved andre kontor enn kontaktpersonen ved Driftskontoret og funksjonærene ved fagkontoret som eier datasettet, får bare tillatelse til å lese hva som står på datasettet.

Det er imidlertid vanskelig å hindre en bruker som har fått lesetillatelse å bruke datasettet til andre formål, f.eks. å kopiere et datasett med personidentifiserte opplysninger. Dette kunne i prinsippet gjøres ved å lage bestemte "leseprogram" som avgrenset hva brukere med lesetillatelse kunne gjøre med datasettet. En slik ordning er ikke innført i Byrået, og en må derfor stole på at brukeren med lesetillatelse ikke misbruker tilliten og bruker datasettet til annet enn til å lage statistikk.

Det er bare personer ved eierkontoret og ved Driftskontoret som har full adgang til produkjonssettene, og bare de personene som er godkjent av kontorleder ved fagkontoret. Tillatelse til bruken av brukerdatasett er derimot delegert til den brukeren som har opprettet settet. Det er ikke gjort tekniske begrensninger i hva for opplysninger som kan være med i et brukerdatasett. De kan altså inneholde identifikasjonsopplysninger. Ettersom retten til å tillate andre (ved og utenfor kontoret) å bruke datasettet er tillagt den som har opprettet det, har kontorlederen ingen kontroll med hvilke opplysninger fra kontoret sine datasett som blir spredt til andre kontor i Byrået. I den grad kontorlederen gir en av sine ansatte adgang til å lagre produksjonsdata på et brukerdatasett, må vedkommende innføre de kontroller og restriksjoner som er nødvendig for å hindre at personopplysninger fra produksjonsdatasettene blir spredt til andre kontor.

Personer som slutter i Byrået eller har permisjon, mister adgang til Byråets data ved at passordet blir slettet. Eventuelle brukerdatasett som vedkommende måtte ha til disposisjon på sluttidspunktet, blir ikke slettet, men forelagt kontorlederen ved det kontor der vedkommende arbeider med spørsmål om det/de skal slettes eller overføres til andre personer ved kontoret.

Dersom personen flytter til et annet kontor i Byrået, mister han/-hun adgangen til de filene vedkommende måtte ha hatt adgang til ved sitt forrige kontor etter noen dager, dersom ikke kontorlederen ved dette kontoret ber Driftskontoret om å hindre dette. I løpet av disse dagene kan imidlertid vedkommende kopiere de brukerdatasettene han/hun ønsker å ha med seg til sitt nye kontor, dersom ikke kontorlederen ved det tidligere kontoret regulerer dette.

3. Praksis ved lagring av data i Byrået

For å belyse hvordan Byrået lagrer sine data på EDB i praksis, skal vi bruke lagringen av data til Folke- og boligtelling 1980 (Fob80) som eksempel. Andre kontors praksis kan avvike litt fra dette.

Dataene til Fob80 ble samlet inn ved bruk av skjema som alle personer 16 år eller eldre sendte inn, og ved bruk av opplysninger fra en rekke administrative og statistiske registre. Skjemaopplysningene ble bearbeidet (kodet, registrert og kontrollert) og koblet sammen med registeropplysningene til et datasett som vi her kan kalle beredskapssettet. Ut fra dette settet ble det laget en rekke andre datasett som skulle brukes for ulike formål (spesielle publikasjoner, spesielle oppdrag osv.). I praksis fant en det nemlig nødvendig å ha flere produksjonsdatasett enn beredskapsdatasettet fordi det ble for omfattende for mange av oppgavene den skulle brukes til. I tillegg kommer så brukerdatasett som den enkelte funksjonær ved Folketellingskontoret har opprettet.

Ved Folketellingskontoret har alle funksjonærer full adgang til alle datasett, såvel produksjons- som brukerdatasett. Årsaken er at alle arbeider med det samme materialet, og at arbeidsoppgavene i liten grad er spesialisert slik at enkelte personer bare bruker spesielle datasett. Arbeidsoppgaver fordeles etter hvem som har kapasitet i øyeblikket, og da er det fleksibelt og effektivt å ha samme tilgang til alle datasett. Ved andre fagkontor har en andre former for arbeidsdeling, og dermed andre løsninger for tildeling av adgang til datasett.

Folketellingsdatasettet er mye brukt til statistikk og analyse ved andre kontorer og forskningsgrupper i Byrået, delvis alene og delvis koblet mot andre data. For at saksbehandlere ved de andre kontorene skal få sette seg inn i materialet, søker de om tillatelse til å se på datasettet. Dersom de kan legge fram et godkjent prosjektskriv der behovet for folketellingsstatistikk er dokumentert, får de slik tillatelse av kontoret. Melding om godkjent søknad sendes Driftskontoret som forteller maskinen hvilke folketellingsdatasett personen ved det andre fagkontoret får tilgang til. Tillatelsen er begrenset i tid og gjelder bare det prosjektet det er søkt om lesetillatelse for. Det kan også stilles andre betingelser fra Folketellingskontoret si side.

Før det nye sikringssystemet kom, var det ikke god kontroll med bruken av data som tilhørte andre kontor i Byrået. I de fleste tilfeller blir det kontoret som eier dataene orientert ved at de blir kontaktet med spørsmål o.l. Men det er også tilfeller der datasett er kopiert, helt eller delvis, uten at eierkontoret vet om det. En systematisk oversikt

over dette, har ikke vært laget og de nye reglene er ikke gitt "tilbakevirkende" kraft for å dekke disse datasettene. Dermed står de som lagde disse filene som eiere av datasettene, selv om de inneholder individdata. Datasettene er imidlertid dokumentert på Driftskontoret slik at de opprinnelige eierkontorer kan skaffe seg en oversikt.

På grunnlag av lesinga av datasettet kan det settes opp spesifikasjoner for produksjonskjøringer (tabeller o.l.). Dersom det andre fagkontoret skal bruke den filen de har fått lesetillatelse for, vil Driftskontoret gjennom lesetillatelsen vite at Folketellingskontoret har godtatt at produksjonsdatasettet eller brukerdatasettet tas i bruk i vedkommende prosjekt. De kan derfor bruke datasettet i produksjonskjøringa. Ofte vil det være fagkontoret selv som styrer produksjonskjøringa, og det er da ikke mulig å kontrollere om vedkommende tar ut opplysninger som f.eks. identifikasjonsnummer.

Alle EDB-lesbare data i Byrået er lagret ved Driftskontoret. Det aller meste lagres på magnetband, men noe blir lagret på magnetplater (disker). De ulike typene av datasett blir lagret slik:

Produksjonsdatasett blir oftest lagret på magnetband. Når datasettet skal brukes, blir det kopiert til platelager. Når jobben er utført blir platelagerkopien slettet. Dersom kjøreplanen viser at produksjonsdatasettet skal brukes flere ganger i løpet av en periode, kan platelagerkopien beholdes over flere dager.

Testdatasett blir lagret på platelager for å være til disposisjon til enhver tid for de som ønsker å teste programmene sine.

Brukerdatasett blir lagret på platelager.

Katalogdatasett blir lagret på platelager.

Folketellingskontoret setter opp ei prioritetsliste for i hvilken rekkefølge de ønsker å få utført kjøringene sine.

Både Folketellingskontoret og Driftskontoret kan til enhver tid få prioriteringslista opp på dataskjermen, og Folketellingskontoret kan føre på nye oppdrag eller endre prioritering. Hovedtrekkene blir imidlertid avtalt mellom den ved Folketellingskontoret som har ansvaret for kontakten til Driftskontoret og Systemkontoret, og kontorets kontakt ved Driftskontoret hver mandag. Prioritet og ønsket om å utnytte produksjons- og brukerdatasett mest mulig effektivt når de ligger på disk, avgjør i hvilken rekkefølge kjøringene blir utført, og når og i hvor lang tid det enkelte datasett ligger på disk.

Magnetbandene blir lagret i et spesielt arkiv ved Driftskontoret. Det er en arkivar som er ansvarlig for det, og det føres journal for all ut- og innlevering. Arkivaren har imidlertid ingen kontroll med hva magnetbandet brukes til, bare hvem som har fått det utlevert.

I den tida datasettet ligger på platelager, kan det benyttes av alle som har et passord som blir godtatt for vedkommende datasett (jf. kap. 2). Dersom noen med et ikke godkjent passord prøver å få tilgang til platelageret, blir dette registrert.

4. Hvordan stemmer praksis med regelverk for lagring av data i Byrået?

Regelen om at registre som inneholder identifiserte persondata bør anonymiseres, har som nevnt blitt praktisert slik at en beholder fødselsnummeret så lenge en trenger det for statistiske formål. Det er flere spørsmål som kan stilles i denne forbindelse:

a) Hvordan skal en forstå betegnelsen "statistiske formål"?

b) Hvor lang tid er "så lenge en trenger det"?

c) Når går et personregister over til å bli et nytt register som det må sendes melding om til Datatilsynet?

a) Byråets datainnsamling kan grovt sett deles i fire:

1) Utvalgsundersøkelser som engangsundersøkelse eller som

2) panelundersøkelse,

3) totalundersøkelse som et engangsprosjekt og

4) totalundersøkelse som en løpende (månedsvis, kvartalsvis, årlig eller tiårlig) statistikk. Til dette punktet kan også regnes oppdatering av Byråets registre og arkiver.

Før et datainnsamlingsprosjekt starter, skal det foreligge et prosjektskriv som skal beskrive formålet med undersøkelsen, og hvordan den skal gjennomføres. (En del innsamlingsrutiner har gått i så mange år, at det ikke foreligger formelle prosjektskriv som definerer formålet. Som regel er imidlertid formålet definert i notater el.l.). Formålet med undersøkelser er som regel å gi statistikk som gir tall for de kjennemerkene som blir registrert. Behovet for statistikk som for det første knytter sammen flere kjennemerker som er registrert om personer på samme tidspunkt (f.eks. utdanning og arbeid), og som for det andre gir opplysninger om kjennemerkene på ulike tidspunkter, har blitt sterkere med årene. Denne statistikken brukes til å studere såvel sammenhengen mellom

kjennemerkene som tidsutviklingen i de enkelte kjennemerker eller kombinasjoner av dem. Dersom en slik statistikk skal bli brukbar, må det for det første være en indre sammenheng i dataene. Det betyr at alle kjennemerker og enheter må ha samme definisjon både når en ser på tvers av flere datakilder som viser til samme tidspunkt, og når en følger datakildene over tid. For å kunne bearbeide dataene med sikte på å oppfylle disse kravene, vil det være nødvendig å lagre dem over en lengre periode med identifikasjon. Behovet for slik statistikk og de kravene den stiller, har hittil vært lite nevnt i prosjektskrivene i Byrået. Det kommer for en stor del av at den tekniske kapasiteten (i form av system- og maskinkapasitet) for å lagre slik statistikk, hittil ikke har vært så stor at kontorene har ansett slike deler av prosjektet som realistisk.

Formålsspesifikasjonene i Byråets prosjektskriv viser alltid til et behov som prosjektet skal dekke. Behovet kan fra brukernes side ofte være nokså vagt definert. (Erfaring viser at brukere ofte må se en mulighet for data før de kan definere et behov). De som fremmer prosjektet må derfor ofte bruke formuleringer som "antatt behov" og "regner med etterspørsel etter" med referanse til beslutninger eller utviklingstrekk som tilsier at et behov kommer til å oppstå. På grunn av de begrensede ressursene vi har i Byrået, er det bare i de tilfellene det er brei enighet om slike antakelser, eller det koster lite å dekke slike antatte behov, at prosjekt av dette slaget blir godkjent utført.

b) Av det som er sagt ovenfor, følger at "så lenge en trenger det" betyr inntil formålet med prosjektet er oppfylt. Når ett av formålene med prosjektet er å følge grupper av individer og overganger mellom gruppene over tid, og dette ikke kan oppfylles uten at en har datasett med personidentifikasjon, må slike datasett arkiveres uten tidsbegrensning og i en versjon som gjør slike oppfølginger mulig i praksis (jf. det som er sagt under c) om kvalitet). Andre versjoner med personidentifkasjon kan avidentifiseres.

c) Byrået har alltid passet på ikke å lagre identifiserbare data på en slik måte at en kan se endringer i verdien på mange kjennemerker for en person over tid. For å kunne gjøre det, må det være vedtatt et statistikkprosjekt som krever en kobling mellom årgangene evt. på tvers av årganger, og denne koblingen skal avidentifiseres så snart resultatet er kontrollert. Prinsippet er altså at hver årgang lagres for seg, selv om de tilsammen utgjør et register. Når en i meldingen til Datatilsynet har

sagt at registeret oppdateres årlig, er det ikke nødvendig å melde nye årganger som nye register, selv om de ulike årgangene er lagret på fysisk adskilte magnetbånd. Ny melding kreves i første rekke dersom det er endringer i omfang eller innhold i registeret, eller det skal brukes til annet enn statistikk. Dersom Byrået hadde koblet årgangsfilene sammen og lagret dem med fødselsnummer, ville altså dette ikke være i strid med Datatilsynets konsesjonsbetingelser. Men det ville være i strid med det som hittil har vært Byråets praksis, en praksis som altså har vært strengere enn Datatilsynets vilkår.

Når det gjelder administrative data må det presiseres at når disse dataene er kommet til Byrået fra det administrative organet, er det Byråets regler for lagring som gjelder, og ikke reglene hos det administrative organet. Ettersom Byrået beholder den gradering på dataene som det administrative organet har gitt dem, betyr det i praksis at reglene for lagring vil være like strenge i Byrået som i det administrative organet.

5. Endringer i regelverket for lagring av data som kan bli aktuelle fram mot 1990

Det ser ikke ut til at det vil komme nye retningslinjer de nærmeste åra som vil kreve endringer i det lagrings- og sikkerhetssystemet som er beskrevet foran. De internasjonale konvensjoner og rekommendasjoner som Norge har undertegnet, går ikke lengre i sine krav enn det Byrået allerede har lagt opp til. Det er heller ikke kommet signaler fra Datatilsynet som går utover de reglene Byrået allerede har gjennomført.

Utvalget som utreder hvor sårbart det norske samfunnet er i spesielle politiske, sosiale og tekniske situasjoner ("Sårbarhetsutvalget") kan muligens komme med forslag som berører Byrået.

6. Krav til lagring av data utfra Byråets behov

"Statistisk Sentralbyrå har som oppgave å:

- klarlegge de samfunnsmessige behov for statistikk og analyse,
- utvikle og holde ved like et statistisk system og samtidig ha en viss beredskap for supplerende undersøkelser etter behov,
- utnytte statistikken til analyse av viktige samfunnsspørsmål."

(Side 9 i "Statistisk Sentralbyrå Perspektiv for 1980-årene" av Arne Øien, RAPPORTER 82/28)

I arbeidet med å klarlegge de samfunnsmessige behovene for statistikk og analyse, har Byrået registrert at det stilles en rekke krav til statistikken, f.eks. at den skal gi både detaljer og oversikt, at begrepene som brukes må standardiseres og at den må gi sammenliknbare tall over tid. Det er to krav som har fått økt betydning i de siste årene. Det ene er at statistikken skal være integrert. Statistikk fra forskjellige områder skal kunne stilles sammen slik at den samla informasjonen gir et helhetsbilde av samfunet. Men kravet omfatter også at statistikken skal kunne integreres på en slik måte at en skal kunne følge grupper av enheter (f.eks. grupper av personer) på tvers av de ulike statistikkområdene. En er f.eks. ikke interessert i bare opplysninger om hvor mange personer som har en bestemt utdanning, eller arbeider i en bestemt næring, men også i hvilke næringer eller yrker personer med bestemte utdanninger er sysselsatt (jf. pkt. 2 om livsløpsanalyser).

Det andre kravet er at statistikken skal gi et bilde av samfunnet som i størst mulig grad kan gi hjelp til å løse de problemene samfunnet allerede arbeider med å finne løsninger på, og hjelp til å oppdage nye problemer i tide.

Selvsagt har disse kravene alltid vært til stede, men når samfunnet blir stadig mer sammensatt og når det raskt endrer seg, får kravene større tyngde.

Dette stiller Byrået overfor flere nye problemer.

Noen av disse problemene skriver seg fra at statistikkbrukerens situasjon også er blitt endret. Endringene i samfunnet er vanskelig å forutsi, og dermed har brukerne vanskelig for å forutsi hvilke statistiske sammenstillinger de vil få behov for. Det fører igjen til at de tabellønskene vi får, ofte vil være lite konkret formulert. En del brukere har erkjent dette. Istedet for (eller ofte i tillegg til) de tabeller Byrået måtte ønske å publisere, ønsker de å ha muligheten til å kunne bestille fra Byrået de tabeller de til enhver tid har behov for, eller selv kunne ta ut slike tabeller fra terminaler med tilknytning til Byrået. Vi må regne med at flere brukere vil se seg best tjent med slike muligheter.

Svaret på den nye utfordringa er sannsynligvis en blanding av standardisering og fleksibilitet i Byrået. Standardisering vil være nødvendig når det gjelder innsamling av data og bearbeidinga av dem. Den vil også være nødvendig for utkjøring av tabeller til faste publikasjoner og abonnementsordninger for å holde kontinuitet, og dermed gi mulighet til å lage tidsserier. Standardkrav vil også være nødvendig for dokumentasjon av data (beskrivelse av enheter, definisjoner, inndelinger, innsamlings- og bearbeidingsmetoder o.l.) og for publisering av data på en slik måte at opplysninger om enkeltpersoner som vi ønsker å beskytte, ikke kan leses ut av statistikken.

Mulighetene for sammenstilling av data bør være slik at data om samme enhet må kunne knyttes sammen uansett hvilke datakilder det gjelder. Dersom det kan knyttes forbindelse mellom enhetene, må det være mulig å knytte opplysninger om enhetene sammen for så å lage statistikk. Ett eksempel er å kunne knytte opplysninger til en person om andre personer i samme familie eller husholdning. Et annet eksempel er å knytte opplysninger om bedriften en person arbeider i til personen selv. Forutsetninger for å knytte forbindelsen, er at det eksisterer en identifikasjon som knytter personen og familien/husholdningen eller bedriften sammen.

For statistikkbrukeren, enten vedkommende arbeider ved fag- eller forskningsavdeling i Byrået, eller arbeider som ekstern bruker ved en dataterminal knyttet til Byrået, vil det være ønskelig å kunne ha til

disposisjon en oversikt over hvilke type enheter Byrået har data om, hvilke kjennemerker som er knyttet til disse enhetene, hvordan kjennemerkene er definert, hvilke tilknytninger som finnes mellom kjennemerkene osv. På grunnlag av denne oversikten skal brukeren kunne formulere sitt ønske om statistikkuttak i form av tabell, grafisk framstilling, temakart, beregninger e.l. Ønsket skal kunne etterkommes som om dataene lå tilgjengelig samlet, og produktet skal være uten sensitive informasjoner om den enkelte person når det kommer ut, dersom ikke hensynet til videre bearbeiding gjør dette nødvendig.

7. I hvor stor grad er de tekniske forutsetninger til stede?

I avsnitt 6 konkluderer vi med at det for brukeren av Byråets data enten vedkommende arbeidde i Byrået eller ikke, var viktig å ha et effektivt dokumentasjonssystem, og et system for lagring og uttak av data som på en rask og fleksibel måte kan gi statistikk i ulike former. Vi skal nå se på i hvor stor grad det tekniske utstyret og den tekniske kunnskapen vi har eller vil få i Byrået, dekker dette behovet.

Byrået har i de siste åra arbeidd med å lage et system for datadokumentasjon. Arbeidet har være konsentrert om to prosjektgrupper. En gruppe har arbeidd med dokumentasjon av publikasjoner. Den andre har arbeidd med et system for metadata ("data om data") i Byrået. Vi skal her konsentrere oss om det siste prosjektet.

Metadatasystemet i Byrået vil, dersom prosjektgruppas forslag blir vedtatt, være lett tilgjengelig for brukeren på en dataskjerm. Det skal inneholde alle opplysninger om data som er relevante for statistiske undersøkelser. Samtidig skal det være knytta til statistikkproduksjonen. Det gjelder såvel bruken av metadata som ajourføring av opplysningene. En førsteutgave av et slikt system (ISDS) er alt i bruk i samband med den kommunaløkonomiske databasen i Byrået. Vi regner derfor med at Byrået om kort tid vil ha til disposisjon et dokumentasjonssystem som vil tilfredsstille de brukerbehovene som er nevnt i avsnitt 6.

Kravet om at brukeren skal kunne formulere sitt ønske om statistikkuttak og få det utført i den form (tabell, grafikk, temakart, beregning) som brukeren ønsker, er delvis dekket i dag. (Vi har systemer for framstilling av tabeller og beregninger ved hjelp av EDB, men savner system for framstilling av grafikk og kart.)

Problemet i dag er tiden og kostnadene ved framstillingen av det ønskede produktet.

Dersom kravet er at produktet skal være klart nokså umiddelbart etter at oppdraget er formulert, er det i dag ikke mulig å få dette til i Byrået. Det vil det heller ikke være i nærmeste framtid, bortsett fra produksjon på grunnlag av "råtabeller". (Et system av detaljerte tabeller som kan "summeres" sammen til det uttaket en ønsker). Dersom en må ha tak i data for det enkelte individ for å lage tabeller, er produksjonstida bl.a. avhengig av hvor mange data som er samla om det enkelte individ, og hvor stor kapasitet datamaskinen har til å håndtere disse dataene. (Fordi om dataene er "samla" trenger de ikke være kobla. Det vil være en teknisk vurdering i hvor stor grad det lønner seg å lagre opplysningene kobla. Å samle data på en slik måte, krever en stor lagerkapasitet, fordi en også må lagre de ikke-koblede dataene). I tillegg kommer alle de praktiske problemene som ofte vil dukke opp i samband med koblinger. F.eks. alle enheter som en forutsetter skulle koble, kobler likevel ikke, og en får dermed "uforklarlige rester" av enheter. Eller: Noen av enhetene mangler de opplysningene en ønsker å ha med i koblingen. Byrået kan ikke regne med å ha kapasitet til å kunne lagre og håndtere en samling av alle data for hvert individ, selv om det teknisk sett er mulig å få til. I praksis vil det bare kunne være ei kjerne av data som vil bli lagt til rette på denne måten. Så lenge en holder seg innenfor denne kjerna, skulle det være mulig å levere produktet nokså umiddelbart etter at det ble bestilt. Omfanget av denne kjerna blir utredet i samband med prosjektet Datasystem for integrert personstatistikk (DIPS).

Med den kapasiteten Byrået kan håpe på å ha i de nærmeste åra, vil det være optimistisk å regne med en produksjonstid på 1-2 dager på de fleste produkt som krever sammenstilling av individdata fra ulike kilder. Måten bestillinga blir gjennomført på, kan også bli noe annerledes når produksjonstida blir økt fra noen minutter eller timer til 1-2 dager. Ved kort produksjonstid, må en ha et standard databehandlingssystem som brukeren sjøl kan styre uten særlig store forkunnskaper utover de som står i dokumentasjon (ofte kalt for "4.generasjonsverktøy"). Slike verktøy kan sjølsagt nyttes også når en må regne lengere tid på oppdraget. Her vil imidlertid tradisjonelt systemarbeid kunne bli et alternativ til bruk av 4.generasjonsverktøy. Avgjørende for hvilket alternativ som skal velges, vil være hvilken løsning som vil bli mest maskinkrevende, og kostnadene ved den manuelle systeminnsatsen. Uansett hvilken metode som blir benyttet ved gjennomføringen av oppdraget, kan brukerens bestilling av oppdraget gå for seg på samme viset som for oppdrag med kort produksjonstid. Etter at oppdraget er formulert, kan systemet gi informasjoner som samlet vil si noe om hvor lang tid det vil ta å produsere det.

I tillegg til de to produksjonsalternativene som er nevnt foran, må en regne med å holde oppe noe av dagens system med spesialprogrammering for uttak av statistikkprodukt, men omfanget av slike oppdrag blir trolig langt mindre enn i dag.

Et av kravene til produksjonssystemet, var at produktet skulle være uten informasjoner om den enkelte personen som skal være beskyttet når det kom ut av prosessen. Til nå har en ikke noen fullgod løsning på hvordan en kan undertrykke tall i tabeller automatisk, men det er mulig å lage systemer som kan skille ut de tabelldelene som er helt fri for beskyttet informasjon om enkeltpersoner, slik at det er et begrenset antall som må igjennom en manuell kontroll.

8. Vurdering av de tekniske forutsetningene mot dagens regelverk og forholdet til opinionen

Av avsnitt 7 ser vi at det er ønskelig å kunne lagre ei kjerne av data om det enkelte individ. Diskusjonen om hvilke data dette skal være er ikke avslutta, men det ser ut til at iallfall fire typer av data peker seg ut:

- demografiske data
- utdanningsdata
- sysselsettingsdata
- inntektsdata

For å lage en slik kjerne av data, må en koble flere registre. I punkt 4,2 i Byråets konsesjon fra Datatilsynet står det at "Registre .... bør bare kobles med andre registre dersom det er absolutt nødvendig". Byrået har tolket det siste slik at dersom det er absolutt nødvendig for å gjennomføre et statistikk- eller analyseprosjekt, kan en foreta kobling av registre. Men i samsvar med praksis for lagring av data i Byrået og med konsesjonens punkt 4,1 Anonymisering, er koblinger blitt avidentifisert så snart det er mulig. Og uten personidentifikasjon kan ikke koblingen brukes som en kjerne av persondata, slik en forutsatte i avsnitt 7.

En måte å løse dette problemet på, og samtidig holde seg innenfor den ramma som konsesjonen setter, er å lagre de koblede dataene med et kryptert nummer istedenfor personnummer. Hvordan dette kan gjøres, er beskrevet i de to notatene som følger som vedlegg 2 og 3. Metoden forutsetter at nøkkelen som omformer personnummer til kryptert nummer, blir lagret slik at det er mulig å koble nye data til "kjernen", både for åjourføring av dataene i kjernen, og for å kunne koble "kjernedata" mot andre data. Metoden forutsetter altså at det finnes en sikker måte å

lagre og forvalte denne nøkkelen på, slik at ikke uvedkommende som måtte ha fått tak i krypterte data kan identifisere personer.

Både krypteringsmetoden, forvaltningen av nøkkelsen og virkningen av metoden for formuleringene i Byråets konsesjon, er nå til vurdering i Datatilsynet. Hittil har vi fått positive reaksjoner på metoden, men spørsmålet om forvaltningen av nøkkelen og eventuelle endringer i konsesjonen, må det arbeides videre med. En må regne med at dersom vi finner fram til en ordning som Datatilsynet godkjenner, vil den bli akseptert av opinionen.

9. Konklusjoner

a) Dagens system for lagring av data ser ut til å fungere tilfredsstillende utifra de personvernreglene som gjelder både i Byrået og ellers i samfunnet.

b) Det ser ikke ut til at nye regler for lagring av data vil komme i de nærmeste åra.

c) Statistikkbrukernes krav og de tekniske mulighetene vi kan regne med å rå over de nærmeste åra, kan gjøre det ønskelig å opprette en permanent lagret kjerne av persondata.

d) Lagring av en kjerne av persondata av forskjellig type med fødselsnummer, bryter med Byråets praksis og Datatilsynets konsesjon. Kryptering av fødselsnummer ser ut til å være en metode som både tilfredsstiller brukernes behov, Byråets lagringspraksis og Datatilsynet.

e) Dersom Datatilsynet finner å kunne godkjenne denne ordningen, er det ikke nødvendig med ytterligere tiltak fra Byråets side når det gjelder tiltak for personvern i samband med lagring av persondata.

Vedlegg 1

Til Driftskontoret i Oslo/Kongsvinger

Fra kontor/brukerident_____ SSB den _____

TILLATELSE TIL BRUK AV DATASETT PÅ IBM

For å kunne utføre prosjekt_____ og statistikknummer_____ gis herved

brukerident:_____ tilgang til følgende datasett:

Datasettidentifikasjon	Til-gang	Tilgangen gjelder: i perioden	Tilgangen gjelder: for program	Kjøre-type
	L/			
	L/			
	L/			
	L/			
	L/			
	L/			
	L/			

VED DETTE SKJEMA SKAL DET LEGGES KOPI AV SIGNERT KJØREDIAGRAM OG/ELLER PROSJEKT-SKRIV SOM BESKRIVER BRUKEN AV DE PRODUKSJONSDATASETT TILGANGEN GJELDER.

PRODUKSJONSDATASETT ER IKKE TILLATT KOPIERT TIL BRUKERDATASETT.

Spesielle betingelser ved bruken av datasettene_____

_____

Andre opplysninger_____

_____

_____

_____
Kontorleders underskrift

Tilgang: L=Les, S=Skriv (gjelder bare brukerdatasett)

Kjøretype: B=Satsvis (BATCH), T=TSO, C=CICS

Dersom det ikke settes noen begrensning for hvor lenge tilgangen gjelder, blir den satt til 3 dager.

Utført:_____ Sign:_____

Vedlegg 2

EAu/BjF, 5/2-85

KOBLING AV REGISTRE UTEN BRUK AV KJENT IDENTIFIKASJON

1. Generelt om statistikkproduksjon og datavern i Statistisk Sentralbyrå

Statistisk Sentralbyrå (SSB) har alltid hevdet at Byråets statistikkproduksjon ikke innebærer noen trussel mot personvernet.

Selv om statistikkproduksjonen i seg selv ikke representerer noen trussel mot personvernet, vil imidlertid den store samlingen av data i SSB's dataarkiv kunne representere en trussel dersom den kommer uvedkommende i hende. Det er derfor nødvendig å føre en forholdsvis streng kontroll med bearbeiding og lagring av Byråets data. Viktig i denne sammenheng er at det ikke lagres data på identifiserbar form som det ikke er bruk for eller med en detaljrikdom som ikke utnyttes. En tidlig avidentifisering av dataene vil imidlertid ødelegge mye av den statistikkproduksjonen vi har i dag og de potensielle muligheter til ny statistikk og analyser som ligger i arkiverte data. Dette skyldes at den permanente identifikasjon av personer, bedrifter og foretak vi har hatt siden 1964 gjør det mulig å koble data fra forskjellige perioder eller forskjellige kilder.

SSB's konsesjon gir anledning til slik kobling når det er "strengt nødvendig". Dette er forstått slik at kobling bare skal foretas i sammenheng med vedtatt statistikkproduksjon og oppheves når arbeidet er avsluttet. På denne måten begrenses den faktiske sammenstilling av data om den enkelte oppgavegiver til enhver tid til et minimum. Dette fører til at koblinger må gjentas hver gang det er behov for det. Dette er noe sløsing med ressurser og nedsetter det vi kaller databeredskapen noe.

Etter hvert som datamaskinressurser blir billigere og maskinene raskere blir denne ressurssløsingen og nedsatte beredskap av mindre betydning. For statistikkproduksjonen er dette en fordel, men

det representerer samtidig en betenkelig utvikling når det gjelder datavern ved at misbruk av data gjennom kobling også blir enklere.

2. Nye krav til datavern under kobling

Utviklingen beskrevet under pkt. 1 har fått SSB til å søke etter en løsning som tillater en økning av databeredskapen med samme datagrunnlag som tidligere, men samtidig oppnå et bedre vern mot misbruk av data under kobling. Det er rimelig å stille følgende krav til en slik løsning:

1) Det må være like gode muligheter for å utføre godkjente koblinger som i dag

2) Behovet for å lagre identifiserte data i SSB's dataarkiv må reduseres

3) Der en finner det nødvendig å levere data med kryptert identifikasjon fra en dataleverandør til SSB, skal dette være mulig uten å tape mulighetene for å koble slike data til andre data i SSB på individnivå

4) Dataene skal ha bare kryptert identifikasjon under statistikkproduksjonen og lagring av koblede data

5) Det skal være forholdsvis enkelt for en utenforstående kontrollør, f.eks. Datatilsynet, å kontrollere at kravene 1-4 er oppfylt.

Det at dataene bare har kryptert identifikasjon betyr at de ikke kan identifiseres uten gjennom en nøkkel som oppbevares av en kontrollør. Kontrolløren kan være en institusjon utenfor SSB.

3. Kryptering av identifikasjonsnummer med fortsatt mulighet for kobling

Vi skal her beskrive en løsning som oppfyller kravene nevnt under pkt. 1 ovenfor. Beskrivelsen er knyttet til persondata som i utgangspunktet er identifisert med fødselsnr.

Det vil bli laget en EDB-rutine som oversetter fødselsnummer til et kryptert identifikasjonsnr. Rutinen vil kunne oversette til flere serier av krypterte nummer avhengig av hvilken nøkkel en velger. (Hver serie vil ha en nøkkel.) Oversettingen foregår i to trinn.

I første trinn blir fødselsnummer erstattet av et tilfeldig nummer (T-nr) i en serie fra 1 til n (hvor n er tallet på personer som har fødselsnr. i dag. Det blir dessuten trolig også avsatt plass i serien til fødselsnr. som vil bli tatt i bruk i de nærmeste 5 eller 10 år). Personer som gjennom tiden har hatt flere fødselsnr. vil få bare ett T-nr. Denne oversettelsen utføres gjennom en katalog som inneholder alle fødselsnr. i stigende orden og ett T-nr. for hver person. Katalogen er framstilt ved først å tildele hvert fødselsnr. et tilfeldig tall fra en såkalt random number generator. Deretter sorteres fødselsnummerene i den rekkefølge de tilfeldige tallene bestemmer. Når dette er gjort tildeles T-nr. som en fortløpende nummerering fra 1 til n og det tilfeldige tallet sløyfes. (Den siste operasjonen er nødvendig fordi generatoren kan tildele samme tilfeldige tall to ganger selv om den arbeider innenfor et tallområde større enn 1 til n.) Til slutt sorteres den katalogen som nå er framkommet i stigende orden på fødselsnummer og for personer som har hatt mer enn ett fødselsnr. blir de øvrige fødselsnr. føyd til i katalogen med T-nr. lik de T-nr. som allerede er tildelt for disse personene.

I annet trinn blir T-nr. oversatt til et nytt kryptert nummer som vi kaller K-nr. Denne oversettingen skjer gjennom to tabeller som programmet genererer ved bruk av en random number generator. Generatoren startes opp ved hjelp av en nøkkel og tabellene som genereres er avhengig av nøkkelen. Forskjellige nøkler gir forskjellige tabeller, men samme nøkkel gir alltid samme tabell. Programmet ødelegger den innlagte kopi av nøkkel og tabeller straks disse er brukt. Det blir ikke tatt utskrift av tabellene.

I prinsippet ville det vært tilstrekkelig å bruke en tabell til oversettingen. Tabellen måtte i så fall hatt 5-6 millioner verdier og blitt kostbar i bruk. Det blir derfor brukt en tabell med 1000 verdier til å oversette tre siffer av K-nr. og en tabell med 10 000 verdier til å oversette de resterende fire siffer av K-nr. Programmet som generer tabellene må sørge for at alle verdier forekommer, men i tilfeldig orden. (I praksis kan noen verdier utelates fordi vi ikke har så mange som 1000 x 10 000 K-nr. til oversetting.) I tillegg til krypteringen blir postene (records) i det krypterte registeret ordnet i tilfeldig rekkefølge i forhold til det opprinnelige registeret for å hindre at fødselsnr. kan bestemmes ut fra en

kjent rekkefølge. Dette blir gjort før registeret skrives ut med K-nr. ved å bestemme rekkefølgen innen blokker på f.eks. 100 records ved bruk av tilfeldige tall som genereres for hver blokk.

Denne formen for oversetting av fødselsnr. til K-nr. tillater at flere registre, samtidig eller til forskjellige tidspunkter, oversettes til samme K-nr.-serie slik at de kan kobles mot hverandre ved bruk av K-nr. En kontrollør må passe på at det brukes riktig nøkkel under krypteringen, at nøkkelen ikke kommer uvedkommende i hende og at bare de data som tillates koblet slipper igjennom oversettingsrutinene med en bestemt nøkkel. Det vil ikke være mulig å oversette fra K-nr. tilbake til T-nr. og videre til fødselsnr. uten å få tak i nøkkelen som ble brukt ved krypteringen. Ved oversetting mellom to forskjellige K-nr.-serier er det nødvendig å kjenne begge nøkler.

4. Krypteringens betydning for datavernet

Ved en vurdering av krypteringens betydning for datavernet er det hensiktsmessig å skille mellom tre områder:

1) Utlevering av data fra SSB
2) Levering av data til SSB
3) Kobling av data i SSB.

4.1. Utlevering av data fra SSB

Etter SSB's konsesjon kan statistiske individualdata utleveres til forsknings- og planleggingsformål. Ved slik utlevering blir dataene avidentifisert så langt prosjektet tillater det. Ved forskningsprosjekter forekommer det at identifikasjonen må beholdes fordi det skal foregå en kobling hos mottakeren med andre data, eller data skal overføres fra SSB til det samme prosjektet på et senere tidspunkt, f.eks. for ajourhold av et register. Et slikt behov er registrert ved et par større forskningsprosjekter. Bruk av K-nr. vil i slike tilfelle klart redusere bruken av data identifisert med fødselsnr. uten å redusere forskernes adgang til data. Dersom alter-

nativet til K-nr. hadde vært å utlevere fullstendig avidentifiserte data i slike tilfeller ville prosjektet krevd mer datakapasitet i SSB fordi senere ajourhold av det avidentifiserte registeret måtte vært utført ved en gjentatt kobling ved bruk av fødselsnr. (Dette har hittil vært tilfelle ved utlevering av data fra folketellinger til Kreftregisteret for å analysere sammenheng mellom yrke og kreft.)

Dersom ikke en tredje institusjon bidrar med data til koblingen kan krypteringsnøkkelen i slike tilfelle oppbevares av SSB.

4.2. Levering av data til SSB

Ved levering av data (på maskinlesbar form) til SSB kan krypteringen foregå hos dataleverandøren som må få utlevert krypteringsnøkkel dersom SSB skal koble mot data som har kommet fra andre leverandører. Kryptering hos dataleverandøren kan være aktuelt ved folketellinger basert på administrative registre. Dette vil kunne redusere bruken av fødselsnr. under en folketelling betydelig.

4.3. Behandling av registre internt i SSB

Kryptering av fødselsnr. i registre internt i SSB vil være aktuelt enten dataene skal inngå i en kobling eller ikke. Krypteringen kan ikke utføres før dataene er kontrollert og eventuelle feil er rettet. Under dette arbeidet er det nødvendig å bruke fødselsnr. for å finne tilbake til nødvendig dokumentasjon under feilrettingen. Under datakontrollen kan det også være nødvendig å foreta maskinelle koblinger mot andre registre ved hjelp av fødselesnr.

Etter at tilstrekkelig mange feil er rettet, kan fødselsnr. erstattes med K-nr., og registeret med fødselsnr. kan som regel slettes. Dette betyr at lagringen av data med fødselsnr. i Byråets dataarkiv blir betydelig redusert. Det bør være mulig å fjerne registre med fødselsnr. etter ett til to års lagring. Dette

forutsetter at koblinger for produksjon av statistikk normalt blir utført ved bruk K-nr.

Det kreves en grundig drøfting før en kommer fram til om alle registre i SSB skal ha samme K-nr.-serie eller om de skal deles mellom flere serier. Vi skal her gjøre noen betraktninger rundt tre forskjellige alternativer.

1) En K-nr.-serie som SSB har nøkkel til

2) En eller flere K-nr.-serier som SSB ikke har nøkkel til

3) En K-nr.-serie som SSB har nøkkel til og en eller flere serier som SSB ikke har nøkkel til

4.3.1. En K-nr.-serie som SSB har nøkkel til

Dette gir samme muligheter som i dag til å gjøre koblinger i SSB, men de koblede data vil ikke være lette å identifisere. En slik identifisering forutsetter at det lages et program for oversetting fra K-nr. til T-nr. og videre til fødselsnr. Dessuten må den som vil bruke dette programmet få tak i krypteringsnøkkelen. Denne bør kunne beskyttes meget godt i SSB. Det største problemet ved en slik løsning vil være å vinne tillit hos publikum til at nøkkelen ikke blir brukt til å oversette tilbake til fødselsnr. Det forutsettes her at de registre som blir kryptert ikke samtidig blir lagret med fødselsnr.

4.3.2. En eller flere K-nr.-serier som SSB ikke har nøkkel til

Dersom krypteringsnøkkelen ikke er tilgjengelig for SSB, må det antas å være lettere å overbevise publikum om sikkerheten i systemet. Om det brukes en eller flere K-nr.-serier, vil avhenge av hvilke restriksjoner Datatilsynet vil legge på SSB's muligheter for kobling.

4.3.3. *En K-nr.-serie som SSB har nøkkel til og en eller flere serier som SSB ikke har nøkkel til*

Under dette alternativet vil det være mulig for SSB å oversette alle sine registre ved bruk av sin nøkkel og koble disse. Men registre kan leveres til SSB med andre K-nr.-serier som ikke kan kobles til SSB's serie. Dersom det senere blir tillatt å koble data mellom registre med forskjellige K-nr.-serier, må SSB's serier oversettes til den andre K-nr.-serien.

4.3.4. *Vurdering av alternativene*

Den største sikkerheten oppnår vi ved ikke å gi SSB adgang til krypteringsnøklene, slik som beskrevet under 4.3.2. Ulempen ved å holde krypteringsnøklene utenfor SSB kan være at det fører med seg mye arbeid for kontrolløren å overvåke alle krypteringer. Dersom det viser seg å bli for mye arbeid, kan SSB tildeles nøkkel for de registre som SSB samler inn med fødselsnr slik som beskrevet under 4.3.3. Når slike registre skal kobles mot registre levert med en annen K-nr.-serie må det skje en oversetting fra SSB's serie til den serien det skal kobles mot. Oversettingen kan bare skje dersom kontrolløren gir riktig nøkkel til datamaskinens program.

5. Hvordan skal krypteringen kontrolleres

Kryptering av T-nr. til K-nr. foregår i datamaskinen styrt av et program som må få tilført et tall, krypteringsnøkkel, for å settes i gang. Programmet vil bli laget slik at krypteringsnøkkelen bare kan gis fra en bestemt dataskjerm. Denne kan plasseres utenfor SSB, f.eks. i Datatilsynet. SSB må på forhånd varsle kontrolløren om at en kryptering skal foretas og oppgi hvilke data registeret inneholder. Kontrolløren må avgjøre om SSB skal få lov til å koble dette registeret mot de registre som tidligere er overført til denne K-nr.-serie. SSB har som tidligere nevnt, konsesjon for kobling når dette er "strengt nødvendig".

Dersom kontrolløren finner at krypteringen kan settes i gang,

oppgir han riktig nøkkel på dataskjermen. For å sikre at det blir brukt riktig nøkkel vil det fra SSB bli oppgitt et nøkkelnr. på to siffer som skal oppbevares sammen med nøkkelen. Nøkkelnummeret blir overført til registeret slik at en kan hindre kobling mellom registre med forskjellig K-nr.-serie. En slik kobling vil være teknisk mulig, men vil gi en tilfeldig sammensetning av personer fra de to registrene. Ved første gangs bruk av en K-nr.-serie må kontrolløren velge nøkkel. Hvilket tall som velges vil være likegyldig, men det må ha det riktige antall siffer. Nøkkelen vil trolig bli satt til 10 siffer. Dersom det er ønskelig kan den deles i to, slik at to personer setter inn hver sin halvpart.

Kontrolløren kan be om å få skrevet ut et lite utvalg av registeret etter krypteringen. Denne utskriften kan brukes til å kontrollere at det ikke kommer andre opplysninger (f.eks. fødselsnr.) inn i registeret enn det som er oppgitt av SSB. En fullstendig kontroll av dette forutsetter at kontrolløren leser utskriften mot de dokumenter (f.eks. statistiske spørreskjema) som opplysningene er hentet fra. Dette kan være en tidkrevende prosess som kanskje ofte vil bli sløyfet. Datatilsynet kan imidlertid på et hvilket som helst tidspunkt ta de kontroller det finner nødvendig for å slå fast at SSB ikke har overført til krypterte registre opplysninger som det ikke har konsesjon til å overføre.

Kontrolløren må ha muligheter for å overbevise seg om at det programmet som utfører krypteringen avsluttes normalt slik at de katalogene som er brukt av programmet er slettet. Han kan også be om utskrifter av programmet for å kontrollere virkemåten.

6. Andre forhold

SSB må til enhver tid ha en beredskapsplan for tilintetgjøring av registre som en ikke ønsker skal falle i en fiendes hender. For registre som er overført til K-nr. bør det være tilstrekkelig å slette nøklene. Dette er en meget enkel operasjon og vil forenkle gjennomføringen av en beredskapsplan betydelig.

Dokumentasjonen av registrene vil bli lagret i datamaskinen i en metadatabase tilgjengelig for lesing fra dataskjermen, også fra

kontrollørens dataskjerm. Metadatabasen vil vise om registrene er lagret i maskinen eller på magnetbånd i arkivet.

Ansvaret for lagring av registre og rapportering til Datatilsynet har hittil ligget på det enkelte fagkontor. Registre som overføres til K-nr. bør planlegges og kontrolleres av et sentralt organ i SSB. Dette organet bør ha all kontakt med kontrolløren og Datatilsynet.

Vedlegg 3

Systemkontoret - Kongsvinger.
KKv, 10/4-85

RUTINE FOR OVERSETTING AV FØDSELSNUMRE TIL ET KRYPTERT NUMMER VIA ET TILFELDIG REKKEFØLGENUMMER

1. Innledning

Statistisk Sentralbyrå har alltid lagt stor vekt på personvernet i sin statistikkproduksjon. Statistikk skal aldri publiseres på en slik måte at følsomme opplysninger kan føres tilbake til personer eller bedrifter. Dataregistre i Byrået som blir brukt til produksjon av personststatistikkk inneholder som regel fødselsnummer. Det er nødvendig for å kunne koble dataregistre med forskjellige referansetidspunkter sammen. Ofte er det også nødvendig å koble dataregistre med forskjellige emner sammen. Dataregistre som er koblet for å lage en bestemt statistikk skal kun forbli koblet til statistikken er produsert. Deretter vil registeret bli fjernet. Det finnes også situasjoner der det ikke uten videre er tillatt å koble registre.

Etter hvert som datatekniken utvikles med organisering av data i databaser, mere effektive koblingsverktøy og datautstyr desentralisert til brukerne, vil det bli enklere å koble dataregistre. Slik det fungerer i Byrået nå, må flere personer fra flere kontorer engasjeres for å få til en kobling. I framtida kan koblingen styres fra en dataskjerm. Selv om dataregistrene fysisk er organisert som egne filer, vil koblingsprosedyren være så enkel at det for brukeren kan virke som registrene er koblet. Vi kan da fort miste kontrollen med hvilke registre som kobles og hvem som kobler.

Dette notatet beskriver en rutine for hvordan vi kan erstatte det offisielle personummeret med et kryptert nummer (koblingsnummer). Rutina innebærer at fødselsnummeret først blir oversatt til et tilfeldig valgt løpenummer, heretter kalt T-nummer. T-nummeret vil ikke få noen spesiell beskyttelse, og katalogen som viser sammenhengen mellom fødselsnummer og T-nummer kan være tilgjengelig for alle brukere av Byråets data. Deretter vil T-nummeret krypteres til et nytt nummer, heretter kalt K-nummer. Det vil ikke være mulig å identifisere en person på grunnlag av et K-nummer, forutsatt at man ikke kjenner nøkkelen som genererte tilfeldige tallrekker og som igjen ble brukt til å kryptere T-nummeret til et K-nummer. Det vil aldri ekstistere noen sammenheng mellom T-nummer og K-nummer på lister eller maskinlesbare media, sammenhengen vil kun være tilstede i EDB-maskinens minne mens krypteringen utføres. Det vil heller ikke være mulig å bryte noen krypteringskode da det egentlig ikke er noe noe system i krypteringen. Det er kun tilfeldig valgte tallverdier som brukes og man må kjenne nøkkelen (the seed) som brukes for å generere tilfeldige tall for å kunne oversette fra K-nummer til T-nummer og derfra til fødselsnummer. I vår rutine vil nøkkelen være et ti-sifret tall som igjen er et tilfeldig

valgt tall og som kjennes kun av de personer som er autorisert for å bruke krypteringsrutina.

2. Formål

Formålet med å erstatte fødselsnummeret med T-nummer og K-nummer er følgende:

- Hindre at personer i dataregistre for statistikkproduksjon kan identifiseres med et identifikasjonsnummer.
- Hindre kobling av dataregistre som vi ikke har anledning til å koble.
- Vi vil i tillegg oppnå EDB-tekniske fordeler ved at et individ får tildelt kun et T-nummer uavhengig av hvor mange personumre vedkommende kan ha hatt, og at vi får en fortløpende og tett nummerserie.

3. Overgang fra fødselsnummer til T-nummer

Dette vil være en meget enkel rutine som tildeler hvert fødselsnummer et tilfeldig valgt nummer, T-nummeret. T-nummeret kan genereres med en rutine som genererer tall mellom 2 grenseverdier, og som aldri lager to like tall. Vi bør likevel legge inn en maskinell kontroll på at det ikke finnes dubletter blant T-numrene.

En annen metode er å bruke en rutine som genererer tilfeldige tall på grunnlag av dato og klokkeslett når rutina kjøres. År, måned, dag, time, minutt og sekund når genereringen starter brukes som startnøkkel (the seed). Den kan lage to eller flere like verdier. Vi må derfor sortere katalogen etter det tildelte tallet og deretter tildele T-nummer fortløpende fra 1 og oppover. T-nummeret vil bestå av 7 sifre, og det vil ikke inneholde noen informasjon.

Input til T-nummerrutina vil være Byråets linkfile som inneholder alle tildelte fødselsnumre gjennom tidene. En person som har fått nytt fødselsnummer en eller flere ganger vil få tildelt kun et T-nr, dvs. at to eller flere fødselsnumre vil bli oversatt til samme T-nummer forutsatt at det refererer til samme person.

Output fra rutina blir en katalog som viser sammenhengen mellom fødselsnummer og T-nummer og som vil bli brukt seinere for å erstatte fødselsnumre i dataregistre for statistikkproduksjon med T-nummer.

T-nummerkatalogen trenger ingen spesiell beskyttelse og kan være tilgjengelig for alle som har bruk for den.

Vi må hindre at de som blir født eller har innvandret til

Norge etter at systemet er satt i drift, får tildelt T-nummer fra f.eks. fem millioner og oppover. Det kan vi få til ved å lage en T-nummerserie som også inkluderer tilveksten i fødselsnumre for 10 - 15 år framover. Fra denne serien med T-nummer kan vi trekke et tilfeldig utvalg som skal være ledige T-nummer for framtidige fødselsnumre. Det vil gi oss anledning til å å bruke et dataregister sortert etter T-nummer til utvalgstrekking.

T-nummerkatalogen må føres ajour med ledige T-numre etter hvert som nye fødselsnumre blir tildelt fødte og innvandrede. De som får korrigert sine fødselsnumre må rettes direkte i i katalogen da de skal beholde det opprinnelige T-nummeret.

4. Overgang fra T-nummer til K-nummer (kryptering)

I denne rutina vil T-nummeret bli erstattet av et tilfeldig valgt K-nummer. Det vil ikke eksistere noen sammenheng mellom T-nummer verken på lister eller maskinlesbare media. Sammenhengen vil kun eksistere i datamaskinens minne mens krypteringen pågår, og da som to tilfeldige tallrekker hvor tallenes rekkefølge bestemmer K-nummeret. Dessuten vil sammenhengen mellom et T-nummer og tilsvarende K-nummer være til stede i maskinenes minne, men bare et og et om gangen.

Med hjelp av en nøkkel (the seed) vil det bli generert to tallrekker, en i intervallet 0 - 9999 og en i intervallet 0 - 999 Alle tall i de to intervallene vil bli generert, men rekkefølgen av de genererte tallene vil være tilfeldig, dog avhenging av av nøkkelen. Nøkkelen kan f.eks. være et ti-sifret tall. Samme nøkkel vil alltid gi samme tallrekker.

Eksempel på tallrekker:

Tallrekke A		Tallrekke B	
Plass i tallrekken.	Tilfeldige tall	Plass i tallrekken	Tilfeldige tall
1	1589	1	853
2	0267	2	025
3	3851	3	387
.	.	.	.
.	.	.	.
3856	0785	289	518
.	.	.	.
.	.	.	.
9999	2375	999	325

Hvis disse tallrekkene hadde blitt generert, ville T-nummer 3856289 bli oversatt til K-nummer 0785518. De fire første sifrene i T-nummeret ville bli erstattet av det tilfeldige tallet som står på det plassnummeret i tallrekke A som er lik de første fire sifrene i T-nummeret. De tre siste sifrene vil bli erstattet av det tilfeldige tallet fra tallrekke B på tilsvarende måte.

Det vil bli umulig å komme tilbake til fødselsnummeret for en person som har fått et K-nummer, uten å kjenne nøkkelen som generererte tallrekkene som k-nummeret er utledet fra.

Vi kan operere med forskjellige K-nummerserier, dvs. at samme individ kan ha forskjellig K-nummer i dataregistre som vi ønsker å beskytte spesielt. Det er ikke mulig å koble to registre som har fått tildelt K-nummer etter forskjellige serier.

Etter dette kan vi operere med to typer koblingsnumre på Byråets dataregistre med personopplysninger. På registre som ikke trenger spesiell beskyttelse kan vi bruke T-nummer. På registre som vi ønsker å legge til rette for en spesiell koblingsberedskap kan vi bruke K-numre fra en spesiell serie. Registre som aldri skal kobles kan få K-nummer fra hver sine serier.

5. Overgang fra en K-nummerserie til en annen

Det vil være mulig å konvertere fra en K-nummerserie til annen. Det kan være aktuelt hvis to registre som har fått forskjellige K-nummer allikevel skal kobles. Framgangsmåten vil være den samme som for overgang fra T-nummer til K-nummer, men begge startnøklene må oppgis til programmet.

6. Hvordan kan eventuell rekonstruksjon av sammenhengen mellom T-nummer og K-nummer forhindres

Programmet som erstatter T-nummer med K-nummer vil være et ordinært program med sekvensielle input- og outputfiler. Recordene vil normalt ligge i nøyaktig samme rekkefølge på input- og outputfile. Det vil da naturligvis være meget enkelt å lage en katalog som viser sammenhengen mellom T-nummer og K-nummer ved å sammelikne en og en record fra henholdsvis input- og outputfile. Dette kan imidlertid forhindres ved å samle opp en viss mengde outputrecorder i maskinens minne , og deretter skrive dem ut i en tilfeldig rekkefølge. Hvor mange records som skal samles opp og rekkefølgen de skal skrives ut i, kan endres tilfeldig for hver gang bufferen skrives ut.

7. Hvordan kan eventuelle feilkoblinger forhindres

Dataregistre med individopplysninger som har fått erstattet T-nummer med K-nummer fra forskjellige serier kan teknisk sett kobles. I utgangspunktet er det umulig for både mennesker og maskiner og se forskjell på K-nummer fra forskjellige serier og også T-nummer for den saks skyld.
Det kan også ligge en fare i at programmet som erstatter T-nummer med K-nummer blir matet med feil nøkkelverdi. Resultatet av det vil bli et register med "ville" K-numre, men det er ikke

lett å oppdage det uten videre.
Følgende kan gjøres for å forhindre dette:

. Hver K-nummerserie får et eget prefix (kode) som sier hvilken serie det er. Prefixet kan enten ligge som en egen record først i dataregistrert eller sammen med K-nummeret i hver record. Poenget er imidlertid at prefixet må tas med i koblingsnøkkelen (ikke det samme som startnøkkel for tildeling av K-nummer) for kobling av registre. Det vil forhindre at registre med K-nummer fra forskjellige serier blir koblet.

. Prefixet nevnt i punktet ovenfor kan være kontrollsifferet til startnøkkelen , beregnet etter modulus11-regelen. Det vil gi en god forsikring mot at programmet blir matet med feil startnøkkel. Hvis man bruker et kontrollsiffer gir det maksimum 9 K-nummerserier, 2 kontrollsifre gir mulighet for 81 K-nummerserier. Antagelig vil det være tilstrekkelig med 9 serier.

Startnøkler kan lages maskinelt. Man kan lage et program som f.eks lager 1000 tall som gir kontrollsiffer 1, 1000 tall som gir kontrollsiffer 2 o.s.v. Fra hver serie på 1000 tall kan man trekke tilfeldig et tall som skal være startnøkkel for den serien. Man kan også skrive ut alle tall fra hver serie på ei liste og la den personen som skal ha ansvaret for nøklene, manuelt trekke et tall fra hver serie.
Startnøklene skal kun være kjent av de personene som er autorisert for å bruke avidentfiseringsrutina.

8. Rutine for hindre "skjulte" fødselsnumre i inputrecordene

Det vil selvsagt være mulig å gjemme et fødselsnummer i en statistikkfile som skal avidentifiseres v.h.a. K-nummer. Fødselsnummeret kan splittes opp og plasseres i forskjellige posisjoner på recorden, der det egentlig er spesifisert andre kjennemerker. Dette kan det være vanskelig å gardere seg 100 % mot, men følgende tiltak er det mulig å gjennomføre:

. Alle inputfiler må dokumenteres på forskriftsmessig måte med filebeskrivelser og kodelister. Et tilfeldig utvalg av recordene kan kontrolleres manuelt mot dokumentasjonen.

. Innholdet på recorden kan kontrolleres manuelt mot skjemaene med grunndata, hvis de finnes. Under bearbeidingen av dataene, vil det ofte bli laget avledededede kjennemerker og det vil vanskeliggjøre kontrollen. Etter hvert som stadig mere grunndata for

statistikkproduksjon blir hentet fra administrative registre, vil denne type kontroll få mindre betydning.

- Man kan foreta maskinell kontroll av validitet av forskjellige kjennmerker. I kode for kjønn kan det som kjent bare fore- komme kodene 1 og 2. Kommunekoder kan kontrollers mot en katalog med gyldige kommunekoder. Dersom det ligger deler av et fødselsnummer i felter som kan kontrolleres på denne måten, vil vi få tilfeller av ugyldige koder.

- Vi kan lage tabeller med marginalfordelinger av kjennemerke-verdier. Marginalfordelingene vil i de fleste tilfellene være kjent fra før. Hvis vi får store differanser kan det indikere at det står noe annet i kjennemerkene enn det som er oppgitt.

Det vil naturligvis bli veldig kostbart og tidkrevende å gjennomføre de to siste punktene. De må sannsynlivis bare ligge som et "ris bak speilet" og gjennomføres som stikkprøvekontroller.

9. Hindre uautorisert bruk av rutina

Det vil være påkrevet å beskytte bruken av programmene i avidentifiseringsrutina. Algoritmen som blir brukt for å generere tilfeldige tall kan godt være kjent av alle. Men selve programmet må beskyttes, slik at uvedkommende ikke kan gå inn å gjøre endringer i programmet. En flink programmerer kunne f.eks. lage en tilleggsrutine for å skrive ut sammenhengen mellom T-nummer og K-nummer på en egen file.

Vi bør også forsikre oss mot at ikke uvedkommende ikke får love å kjøre programmet. Selv om det er helt nødvendig å kjenne startnøklene for å kunne oversette fra T-nummer til K-nummer, vil nøklene foreligge et eller annet sted. Det vil da naturligvis alltid være en mulighet for at uvedkommende kan få tak i dem.

Det finnes mange gode muligheter for å beskytte program og data gjennom sikkerhetssystemer i datamaskinen.

- Det vil være kun en bruker som har tilgang til programmene i rutina. Brukerens identifikasjon som må være "kjent" for datamaskinen, er beskyttet med passord som tildeles tilfeldig av maskinen. Passord blir skiftet automatsisk etter x antall dager. Man kan også be om at det blir laget et nytt passord for hver gang man bruker maskinen.

- Startnøklene skal aldri oppbevares sammen med programmet. Programmet må mates med startnøkkelen i det øyeblikket programmet startes (ikke når programmet eventuelt lastes inn i maskinen). Startnøkkelen må gis fra en bestemt skjermterminal og av en bestemt bruker. Nøkkelen kan også deles opp og fordeles på flere brukere.

. Startnøkkelen kan fjernes automatisk fra programmet når det første tallet er generert. Det vil si at nøkkelen vil befinne seg i maskinen kanskje bare brøkdelen av et sekund.

10. Beskyttelse av nøklene

Det må treffes tiltak for å beskytte nøklene. Det vil være like viktig å beskytte nøklene mot å gå tapt som å beskytte mot misbruk. Det finnes mange måter å lagre nøklene på fra i forseglet konvolutt i en safe til lagring på magnetstriper i plastkort.