Interne notater

STATISTISK SENTRALBYRÅ

82/27

10. august 1982

STATISTISK SENTRALBYRÅS LANGTIDSPROGRAM FOR 1980-ÅRENE

Bakgrunnsnotat nr. 6

DATATILGJENGELIGHET FOR FORSKNINGSFORMÅL

INNHOLD

BYRÅETS LANGTIDSPROGRAM FOR 1980-ÅRENE

BAKGRUNNSNOTAT NR. 6: DATATILGJENGELIGHET FOR FORSKNINGSFORMÅL

I. Innledning

Emnet for dette notatet er lagring og gjenhenting av informasjon. Informasjon kan lagres på mange forskjellige måter og en gitt, større informasjonsmasse vil vanligvis kunne inndeles i delmasser som det kan være rasjonelt å lagre separat, etter ulike systemer og på ulike medier. Selv en ekstrem utnytting av moderne databaseteknikk vil neppe i overskuelig framtid føre til at hele Statistisk Sentralbyrås samlede datatilfang bør organiseres i en enorm database. Men dette betyr ikke at ikke hele datamassen bør innordnes i et rasjonelt og avstemt, enhetlig system.

Lagringssystemet og dets enkelte elementer må være tilpasset den bruk av informasjonen som skal dekkes. Avgjørende for lagringsformen vil være en avveining av ordnings- og lagrings-kostnader mot gjenhentingskostnader ved påregnelig bruk av data.

For mange formål vil det være behov for å hente fram og omgruppere statistisk materiale som ikke nødvendigvis er tilrettelagt for den bruken som nå er aktuell. Formålene kan være beskrivelser eller ha et sterkere eller svakere preg av analyse. I alle fall vil det være viktig at eksisterende informasjon er lagret på en slik måte at den er tilgjengelig uten unødige kostnader i tid og ressurser, når det er bruk for den. Det er neppe fruktbart å trekke noe skarpt skille mellom bruk til forskningsformål og andre mer beskrivende anvendelser. Det er grunn til å tro at bruk til forskningsformål vil stille de største krav til tilgjengelighet.

I den tradisjonelle statistikkproduksjonen blir behandlingen av den innsamlete informasjon styrt av hensynet til å frembringe et produkt, som vanligvis vil være en tabellpublikasjon eller et sett tabeller. Vi kunne betrakte et slikt tabellprodukt som et sett av avledete informasjoner, produsert ved hjelp av de primære informasjoner som gikk inn til statistikkprodusenten. Vanligvis ville da primærinformasjonen og eventuelle mellomprodukter kunne betraktes som en slags slagger, som var tømt for sitt økonomisk verdifulle innhold. Leilighetsvis ville nye behov eller nye metoder kunne gjøre det lønn-

somt å gå tilbake til slaggene for å ekstrahere residualt informasjonsinnhold. Av den grunn burde slaggene tas vare på, men siden ingen kunne vite om og i tilfelle hvordan de kunne komme til nytte, var det lite lønnsomt å legge omtanke og kostnader i en rasjonell lagringsprosedyre.

Tabellsettet som gav det aggregerte informasjonsinnhold fra et gitt sett av primærinformasjoner var statistikkproduksjonens sluttprodukt. Hvis dette settet var godt nok var det det perfekte utgangspunkt for videre bearbeiding for observatører og analytikere som var avtakere av produktet og atskilt fra produksjonsprosessen.

Disse forholdene førte til det typiske produksjonsmønster med et stort antall separate "statistikker" som hver aggregerte et begrenset sett av primærinformasjoner frem til et begrenset sett av tabeller.

II. Situasjonen har forandret seg

Fem forhold har brakt viktige nye trekk inn i bildet:

1. Utviklingen av nasjonalregnskapet, og i de senere år ressursregnskapet og arbeidet med levekårsmålinger og et system for sosiodemografisk statistikk (SSDS), har forlenget produksjonsprosessen i retning av samordning og samkobling av produktene fra flere "statistikker". Statistikkprodusentens engasjement i analytisk bruk av produktene har trukket i samme retning.

2. Etableringen av statistisk-administrative registre som oppdateres ved faste rutiner og der nøkkelopplysninger er knyttet til fast identifiserte enheter har skapt viktige datareservoarer som i prinsippet er løpende tilgjengelige. Registrene tjener også som anker- og koblingspunkter for tilleggsinformasjon om de enhetene de omfatter.

3. Utviklingen i datateknologi har gjort at trykte tabeller ikke lenger er det naturlige utgangspunktet for analytisk bruk av et statistikkprodukt. I stor utstrekning er det nå mulig og formålstjenlig å basere den analytiske bearbeidingen direkte på primærinformasjonen, i sin mest avanserte form, gjennom databaser. Også aggregert informasjon i den form vi finner den i tradisjonelle, trykte tabeller vil teknisk sett ligge best til rette for videre analyse når den er lagret på andre media.

4. Nyere analyseteknikker har sammen med utviklingen i datateknologi gitt en enorm utviding i den massen av informasjon som kan taes ut av et gitt sett av primærinformasjon. (Dette har også ført til store besparelser og gitt nye muligheter i produksjonsprosessen for de tradisjonelle "statistikker".)

5. Viktigst av alt er det kanskje at utviklingen i datateknologi har gjort det naturlig å produsere aggregert informasjon ved å koble primærinformasjon fra forskjellige sett av primære kilder. Kobling kan gjennomføres på flere nivåer, helt ned til den enkelte statistiske enhet, og den gir en stor utvidelse i informasjonsverdien av en gitt samling av primær statistisk informasjon.

Det er imidlertid ikke slik at de nye aggregatprodukter kan framstilles kostnadsfritt. Som rimelig er vil kostnadene ved å hente ut et gitt, spesialtilpasset produkt i tillegg til et tradisjonelt være avhengig av hvordan prosessen for bearbeiding og lagring av primær-informasjon, eventuell "mellomproduktinformasjon" og "sluttproduktinformasjon" er lagt opp.

For brukerne av data betyr den nye situasjonen at de i prinsippet kan få adgang til data på et mye mer detaljert nivå enn før og at de kan spesifisere sine egne aggregater og kombinasjoner. Men for at dette skal kunne realiseres kreves det en ganske høyt utviklet ekspertise hos brukeren. Mens det ofte er slik at det er relativt enkelt å gjennomføre mange slags analyser på grunnlag av publiserte data, så langt disse rekker, ser vi stadig eksempler på store og små "katastrofer" når uerfarne analytikere skal utføre analyser på data som må hentes ut fra moderne lagringsmedia, der de i prinsippet skal være lett tilgjengelige. For at fordelene ved den nye teknologien skal kunne utnyttes kreves det opplæring av brukere og dataeksperter, og det kreves at en finner fram til gode og operasjonelle rutiner for datalagring. Det er det siste vi er opptatt av i dette notatet.

Situasjonen gjør det naturlig å tenke seg et nytt system for å tilveiebringe statistiske data til beskrivelse og analyse. En kan tenke seg at det bygges opp et systematisk arkiv av statistiske opplysninger. Dette arkivet ajourføres og suppleres fortløpende med primærdata som hentes inn gjennom ulike rutiner: dels tradisjonelle "skjemaundersøkelser", dels tapping av registre som også - og i alle fall til dels primært - tjener andre formål. Arkivet tappes delvis i regulære rutiner, delvis etter behov for statistikk til publikasjoner, direkte beskrivelser og analyser.

En omlegging i den retningen vil være av større betydning for den analytiske utnytting av data, og for den mer analytiske beskrivelse av fenomenene enn for den direkte statistiske beksrivelse i tradisjonell forstand. For forskningen vil en vellykket omlegging åpne kilder som en idag vet finnes, men som i praksis er utilgjengelige eller svært tungt tilgjengelige. Det må her presiseres at innordningen av data i et samordnet system vil være av begrepsmessig og organisasjonsmessig art. Forutsetninger om bruken, samt praktiske og økonomiske overveielser må være avgjørende for hvordan begrepene skal fastlegges, etter hvilke prinsipper data skal grupperes og identifiseres, i hvilken

grad det skal være mer eller mindre uavhengige delarkiver og deler av delarkiver og hva slags fysiske lagringsmedia som skal brukes for de forskjellige deler av arkivet. Poenget er at alle deler av arkivet skal være vurdert i forhold til helheten og ha en plass innenfor den. Alle lagrete data skal i prinsippet være tilgjengelige. Men graden av tilgjengelighet skal selvsagt ikke være den samme for alle typer av data. Her må en bygge på en avveining av lagringskostnader mot gjenhentingskostnader, sett i forhold til påregnelige behov for gjenhenting.

Det opplegget som er skissert her kan betegnes som et arkivstatistisk system, og i det følgende vil denne betegnelsen bli anvendt. Det som er av vesentlig betydning for databrukerne er at arkivfunksjonen gjøres til gjenstand for en samlet gjennomtenkning og systematisk planlegging.

Kanskje kan en si at vi med et arkivstatistisk system her mener et opplegg av produksjonsprosessen for statistikk, der uthenting av statistikkprodukt fra systematisk arkivert og samordnet informasjon er et vesentlig element.

Det er ikke nye tanker som legges fram her. Allerede i 1969, for 11 år siden, ble et "arkivstatistisk system" foreslått og ganske grundig beskrevet av Aukrust og Nordbotten (Artikler nr. 34). Disse forfatterne konkluderte sin gjennomgåelse med følgende uttalelse: "Byrået ville forsømme sin plikt hvis det ikke pekte på de muligheter som nå foreligger for å oppnå en radikal forbedring av datagrunnlaget for norsk samfunnsforskning og -planlegging. Hvor langt planene skal realiseres, avhenger av hvor langt administrasjonen og de bevilgende myndigheter vil finne at pengene vil være vel anvendt."

Det må være klart at et arkivstatistisk system i den forstand det er definert her eller i Aukrust og Nordbottens artikkel ikke eksisterer i Statistisk Sentralbyrå.

(Noen vil allikevel hevde, at så sant det finnes lagrete, statistiske data, og disse på en eller annen måte kan hentes ut av lagrene, har en et arkivstatistisk system. Det er lite fruktbart å strides om ord, og det får være nok å understreke at i dette notatet er betegnelsen gitt et langt mer krevende innhold. Det betyr noe annet og mer enn et arkivsystem for statistikk).

III. Situasjonen i dag

Vi kan trolig slå fast:

- Produksjonen av statistikk foregår stort sett etter tradisjonelt mønster. I noen grad har tapping av registre erstattet tradisjonell datainnhenting, men uten at dette har endret rammen og målsettingen i de seksjonerte produksjonsprosessene.

- I tillegg til tabellpublikasjoner produseres det i økende grad filer med primærdata for analyser og spesialbearbeiding.

- Det gjennomføres situasjonsbetingete koblingsoperasjoner av engangsnatur.

- Det er gjort en innsats for å bedre tilgjengeligheten av bestemte typer lagret informasjon, nemlig de demografiske data som organiseres under beredskapsfilen.

- Det er satt i gang arbeid for å produsere en bedre dokumentasjon av lagrete data.

- Arbeidet med Tilleggsundersøkelsen til FOB-1980 og framtidige folketellinger trekker i retning av et arkivstatistisk opplegg.

- Det er etablert systematiske dellagre av spesielle datamasser i form av spesialfiler, databanker og databaser.

Tilsammen peker ikke dette mot en snarlig overgang til et arkivstatistisk system etter de linjer som er skissert foran, men noen av de prosesser som en slik overgang vil kreve er påbegynt.

IV. Muligheten for en utvikling mot et arkivstatistisk system

En kan si at overgang til et arkivstatistisk system vil kreve vesentlige omlegginger på to punkter:

1) Produksjonsprosessene for primærstatistikken må frigjøres fra sin ensidige innretning mot de sektorvise publikasjoner, og trekkes inn i en større sammenheng.

2) Den systematiske lagring av data måtte bli en hovedoppgave, med den nødvendige ressurstilgang.

Kommentar til 1): Frigjøring og integrering av produksjonsprosessene behøver ikke gjennomføres som brudd med tidligere opplegg. Det er snarere spørsmål om en bevisst og gradvis sterkere satsing på utviklingslinjer som tildels allerede har tvunget seg fram. (For å unngå misforståelser: Det er ikke spørsmål om å avskaffe sektorstatistikken, men om å se den i en større sammenheng; hvordan informasjon om en sektor både kan trekke på og bidra til å utfylle informasjonen om en annen, jfr. nasjonalregnskapet.)

Kommentar til 2): Den systematiske lagring av data er trolig den del av oppgaven som vil stille størst krav til omstilling både hos produsenter, brukere og datateknisk personale. Oppgaven er ikke løst med utbyggingen av godt ajourholdte statistikkregistre og lagerrutiner for "slaggene" fra den løpende statistikkproduksjonen.

Det vil kreves:

- at dataarkiveringen defineres som et sentralt mål i den statistiske informasjonsbehandlingen
- at det utarbeides en utførlig, overordnet og samordnet plan for dataarkivets innhold og funksjoner
- at det innføres rutiner og normer for standardisering og dokumentasjon av arkiverte data
- at det utarbeides et logisk og konsistent referansesystem for lagrete data
- at det innføres rutiner for å fastlegge hvordan det enkelte, konkrete sett av data skal ordnes og innpasses i arkivet
- at det innføres rutiner for retting, ajourhold og komplettering av lagrete data og den tilhørende dokumentasjon, herunder rutiner for behandling av inbyrdes motstridende opplysninger
- at det innføres rutiner for hvordan uttaksdata skal kunne produseres
- at det utarbeides en plan for gradvis overgang fra nåværende system til et arkivstatistisk system
- at ansvaret for det arkivstatistiske system legges til en bestemt organisatorisk enhet i Byrået
- at den ansvarlige enhet får nødvendige ressurser.

Vi skal gå nærmere inn på disse punktene.

V. Dataarkiveringen som mål

Et rasjonelt dataarkiveringssystem kan bare utvikles gjennom et samarbeid mellom tre grupper: statistikkprodusentene, EDB-spesialistene og brukerne. Dersom de arkiverte data blir sett på som hovedprodukt i produksjonsprosessen, og ikke den enkelte produksjon, vil det trolig føre til omlegginger i produksjonsprosessen og til at produsenter og brukere blir tvunget til å ta mer aktivt del i å planlegge hvordan data skal gå inn i og organiseres i arkivet.

VI. En overordnet og samordnet plan

Det er mulighetene for utnytting på tvers av de tradisjonelle inndelinger som gjør det nødvendig med en overordnet samordning. Her spiller innordningen i samordnete systemer som nasjonalregnskap og SSDS en viktig rolle. Men det er også viktig at statistikk som er samlet inn med et gitt primært formål ofte også kan gi data til belysning av andre tradisjonelle statistikkområder: Kjente eksempler er forbruksdata som kan utnyttes til omsetningsdata, fordelingsanalyse, levekårsdata etc. Dette er forhold som kan få betydning for inndelingen i del-arkiver.

Rent organisasjonsmessig vil det også være behov for at dataarkivet bygges opp etter et felles og samordnet mønster for alle sine delmasser.

Et viktig problem blir samordningen mellom totalmasser og utvalgsmasser, og mellom de forskjellige utvalgsmasser.

VII. Standardisering og dokumentasjon

En betingelse for at dataarkivet skal kunne funksjonere som et hele er at det arbeidet med å innføre tverrgående standarder som lenge har vært en hovedoppgave i Byrået, og som er kommet særlig langt i den økonomiske statistikken, videreføres, særlig i personstatistikken.

Standardiseringen må omfatte de grunnleggende enheter som kjennetegnene knyttes til, ikke bare personer og bedrifter, men også enheter som familie, husholdning, region på ulike nivåer, foretak, konsern osv.

Når det gjelder standardisering av kjennetegn og grupperinger kan det trolig sies at mye er gjort og mere står igjen.

Kravene til dokumentasjon er mangeartet. Det må være mulig å finne fram til:

- hvilke data som er arkivert
- detaljerte spesifikasjoner av arkiverte data
- hvordan en kan hente ut arkiverte data
- hvordan en kan bearbeide og samkople arkiverte data
- mangler og forbehold ved arkiverte data
- sammenliknbarhet og overgangsnøkler for data som ikke er direkte sammenliknbare.

VIII. Ordning og lagring av data

Det må avgjøres hva som skal lagres og hvordan det skal lagres.

Det er ikke uten videre gitt hvilke informasjoner som skal lagres; informasjon om en fødsel vil normalt inneholde informasjon om egenskaper ved selve fødselen: levendefødsel-dødfødsel, enkelfødsel-flerfødsel; informasjon om morens alder, ekteskapelig status, barnetall, varighet siden forrige fødsel, eller inngåelse av ekteskap, informasjon om faren, søsken osv.

Informasjon om et salg vil inneholde informasjon om varenes art, pris, oppgjørsmåte, forsendelsesmåte, leverandør, mottaker osv. Det er også spørsmål om i hvilken utstrekning aggregert og avledet informasjon skal lagres: spesielt produkter og mellomprodukter fra bearbeidingsprosesser.

Det er altså spørsmål om data på mange stadier:

(i) Direkte opplysninger fra oppgave- og meldingsskjemaer
(ii) Avledete opplysninger om grunnenhetene
(iii) Opplysninger ordnet i statistiske registre
(iv) Grupperte rådata
(v) Opplysninger ordnet i "analysefiler"
(vi) Opplysninger ordnet i databaser
(vii) Bearbeidete data, tabeller.

De fleste bearbeidinger av rådata vil skje for å skaffe frem spesifikke produkter, og ikke for generelle arkiveringsformål. Men når bearbeidingen skjer, må arkivet ha plass til å lagre resultatene.

Hvordan skal data lagres? Data kan lagres som tidfestede kjennetegn ved grunnenhetene, personer, familier, ..., bedrifter, foretak, ... Men lagringen kan også bygge på de statistiske hendinger som enheter: fødsler, trafikkulykker, salg. Når grunnenhetene f.eks. er personer, kan vi ha kjennetegn som kjønn, fødselsår, fullført utdanning, ekteskapelig status, dato og kjønn for barnefødsler osv. Når grunnenhetene f.eks. er fødsler, kan vi ha kjennetegn som morens alder, ekteskapelige status og bosted, fødselens karakter som enkel eller flerfødsel, medisinske kjennetegn, (f.eks. keisersnitt) kjennetegn ved barnet (barna) død - levende, kjønn, vekt osv.

For begge lagringsformer blir det spørsmål om valg av ordningsprinsipper ved lagringen: kronologi, arten av kjenntegn, enheter som er berørt osv.

En avveining av tre typer av kostnader vil være viktige for valg av lagringssystematikk: kostnader ved å legge data inn i arkivet, lagringskostnader og kostnadene ved uttak. Opplegg må velges på grunnlag av en bedømmelse av alle tre kostnadselementer.

Fysiske lagringsmedia

Teknisk spenner de mulige lagringsmedia fra arkiverte, utfylte statistikkskjema og meldingsblanketter over EDB-media som datafiler på bånd, plater og disketter, databaser, maskintabeller, trykte publikasjoner til mikrofotograferte tabeller.

Fordeling på og samordning av bruken av de ulike madia er i mye et teknisk problem, men de løsninger som velges for de ulike datamasser må være basert på en samlet vurdering av alle mulighetene.

IX. Retting, ajourhold og komplettering

a) Det må innarbeides rutiner som sikrer at informasjon om rettinger tilføres alle lagrete data, enten gjennom direkte oppretting eller på annen måte.

b) Endringer i definisjoner og omfang må dokumenteres og rutiner for samordning av data med manglende homogenitet må innarbeides, f.eks. overgangsnøkler, markeringer av brudd i serier etc.

c) Nye opplysninger må flyte inn i systemet på de riktige punkter.

X. Produksjon av uttaksdata

Arkiveringen må planlegges og organiseres fram til og med prosessene for å hente ut data, og uttaksdata må kunne være lagt til rette for den bruk som skal gjøres av dem.

Det er også under dette punktet mange alternative former, som må være rimelige, i varierende grad for de ulike typer av data: Det betyr at uttaksprosessen må omfatte også ordning og aggregering av data, og i det minste tilretteleggingen for beregning av sumtall, rater og fordelingsmål.

Uttak kan gjøres i mange former, f.eks. maskintabeller eller tekstede manuskripttabeller, mer eller mindre standardiserte bånd (f.eks. SPSS-filer), plater og film. Det må også vurderes hvem som skal ha kompetanse til å hente ut data på de forskjellige måter.

XI. Plan for overgang

At Byrået innfører et arkivstatistisk system vil ikke nødvendigvis bety noe brudd med de produksjonsmetoder som praktiseres idag og de utviklingstendenser som gjør seg gjeldende. Det vil først og fremst bety at arkivfunksjonene vies en mer systematisk gjennomtenkning og tildeles en økende rolle som mål og retningsgiver for de ulike ledd i produksjonsprosessen.

Til en viss grad kan en si at elementer av et arkivstatistisk opplegg tvinger seg inn i statistikkproduksjonen i økende tempo. Men hvis vi ikke bruker omtanke og ressurser til å styre prosessen, kan vi få et resultat som er mye dårligere enn det behøvde å være.

Det vil kreves:

- en systematisk gjennomgåelse med sikte på å utarbeide et arkiveringsprogram som i prinsippet løser problemer av den typen som er tatt opp i dette notatet. Prinsippløsningen vil tildels gi konkrete behandlingsforskrifter, og tildels gi regler for hvordan slike forskrifter skal fastlegges i det enkelte konkrete tilfelle (for en gitt datamasse).
- en gradvis omlegging av perspektivet i statistikkproduksjonen.

Det sier seg selv at innføringen av et arkivstatistisk system henger nøye sammen med Byråets EDB-plan.

XII. Ansvaret for et arkivstatistisk system

De generelle retningslinjer og planer for arkivfunksjonen så vel som den løpende styring kan bare utformes på en rasjonell måte hvis både emnespesialister i statistikkproduksjonen, dataspesialister og brukere sammarbeider om oppgavene.

Trolig må det skapes spesielle samarbeidsprosedyrer og kanskje spesielle samarbeidsorganer for den løpende styring. Men selve arkivet må ligge under en av Byråets enheter, og vel i produksjonsavdelingen. Det kan være naturlig å se dette i sammenheng med forslaget om organisasjonsmessige tiltak i notatet om administrative datasystemer.

For å utarbeide de generelle retningslinjer og planer for gjennomføringen av et arkivstatistisk system vil det trenges en spesiell prosjektgruppe med representanter for de tre grupper som er nevnt ovenfor.

XIII. Nødvendige ressurser

I sin Artikkel fra 1969 (loc.sit.) skriver Aukrust og Nordbotten:

"En gjennomføring av de prosjekter som er omtalt ovenfor vil ta mange år, fordi det vil kreve arbeidskraft som ikke kan skaffes på kort varsel. Kostnadene vil bli forholdsvis store. Grove overslag viser at selve etableringen av et dataarkiv som skissert vil kreve beløp av størrelsesordenen 10-20 mill. kr. fordelt over en årrekke, og at det senere vil kreve 6-8 mill. kr. pr. år å holde datamassen løpende vedlike.

... I tillegg kommer utgifter - men etter hvert også store besparelser - hos andre etater som Byrået vil samarbeide med for innhenting av data."

Med dagens priser skulle disse anslagene mer en fordobles. På den annen side er det vel slik at mange av de mest kostnadskrevende av de prosessene som er forutsatt hos Aukrust og Nordbotten kanskje allerede er kommet ganske langt, og at en videre utvikling er innbakt i eksisterende planer og budsjetter.

I dag kan det se ut som om det viktigste er å få avsatt personressurser til å utarbeide den samlete arkivplan, og til å gjennomføre den løpende styring av arkivfunksjonene. Men det er klart at ordningen også bl.a. vil kreve betydelig bruk av tekniske ressurser. Konsekvenser for kostnader ellers i produksjon og arkivering av statistiske data er det vanskelig å ha noen mening om.

XIV. Handlingsprogram

A. Et punktvis handlingsprogram

For å komme i gang med innføringen av et samordnet og produksjonsrettet arkivstatistisk system av den typen det er snakk om her bør det på et tidlig stadium utpekes en enhet i Byrået som (eventuelt inntil en mer permanent ordning innføres) skal være ansvarlig for systemet.

Det bør også nedsettes en arbeidsgruppe som har til oppgave å stå for den praktiske gjennomføringen. Arbeidsgruppen kan bestå av representanter fra system- og driftskontor, fagavdeling og forskningsenheter, og den må disponere de nødvendige ressurser.

Arbeidsgruppens oppgave kan omfatte følgende punkter:

1) Fase I Gjennomgåelse av tilstanden i dag

(i) En registrering av eksisterende lagringsrutiner for primærdata og statistikkprodukter. Oversikten må omfatte rutiner for arkiveringen, dokumentasjon og mulighetene for uttak av data. Ideelt burde den beskrive hvordan data går fra primærkilden, gjennom produksjonsprosessen til arkiv, og eventuelt til datauttak.

(ii) En oversikt over aktuell og potensiell gjenbruk av data, særlig for forsknings- og analyseformål.

(iii) En vurdering av eksisterende lagringsrutiner for hver enkelt "datamasse"; og en vurdering i forhold til behovet for samordningen mellom datamassene.

2) Fase II. Utarbeiding av generelle retningslinjer for et arkivstatistisk system (Se også B nedenfor):

(i) Inndelingen i delarkiver.

(ii) Utarbeiding av et generelt referansesystem (eventuelt med "underavdelinger").

(iii) Utarbeiding av dokumentasjonsstandarder

(iv) Utarbeiding av ajourholds- og avstemmingsrutiner.

3) Fase III. Iverksetting (Se også C nedenfor):

(i) Prosedyre for innpassing av eksisterende, lagrete data.

(ii) Prosedyre for innpassing av eksisterende registre.

(iii) Prosedyre for innpassing av data fra løpende statistikk.

(iv) Prosedyre for innpassing av ny statistikk.

B) Nærmere om fase II, generelle retningslinjer

(i) Om inndelingen i delarkiver

Stort sett ligger eksisterende data lagret i delmasser bygget opp i tilknytning til de sentrale registre, eller i delmasser som refererer til den enkelte innsamlings- eller bearbeidingsprosess. På samme måte kommer nye data inn i systemet. Det blir derfor i første rekke spørsmål om i hvilken utstrekning det er behov for integrering av grupper av slike delmasser, og hvordan en slik integrering skal gjennomføres. Integreringen kan være fysisk, eller bare dokumentasjons- og eller definisjonsmessig. Den mest vidtgående integrering representeres (for tiden) av databaser.

Det synes åpenbart at det vil være mye å vinne på en økt integrering. På den annen side kan dette bli kostbart og reise problemer i forhold til dataloven. En bør nok derfor gå forsiktig og skrittvis fram.

Arbeidet må organiseres sentralt og gjennomføres i samråd med fagkontorene og produksjonsenhetene.

(ii) Om et generelt referansesystem

Det systematiske referansesystem vil være kjernen i hele arkivsystemet, og vi ville være kommet et langt stykke hvis vi allerde nå kunne skissere strukturen i det.

Men vi må nok regne med at det vil være vanskelig og krevende å komme fram til et brukbart system, der bl.a. behovet for "veier" inn i systemet og for kryssreferanser er tilfredstillende ivaretatt. Det får være nok her å vise til at det har vært arbeidet en god del i Byrået med å finne fram til et referansesystem for lagret statistikk, og at det må være mulig å sette inn ressurser for å bygge videre på dette arbeidet.

Arbeidet må organiseres sentralt.

(iii) Utarbeiding av dokumentasjonsstandarder

Det er viktig at dokumentasjonen må gjelde såvel innhold som tekniske spesifikasjoner. Det må legges opp et system som er utformet generelt for bruk på alle slags data og alle typer av lagringsmedier. Systemet må være så fullstendig at en bruker får nøyaktig beskjed om hvilke enheter som dekkes, hvilke opplysninger som gis om enhetene og hvilke avgrensinger og definisjoner som er brukt. Videre må den nødvendige informasjon for å hente ut og tolke data finnes. Dokumentasjonen må være løpende ajour og tilgjengelig. Erfaringer i Byrået tyder på at det vil være rasjonelt å legge opp dokumentasjonen "trinnvis" på to eller flere nivåer. Brukeren vil da i første omgang få grove, men oversiktlige opplysninger, antakelig gjennom skjermoppslag med utskrifter. I ett eller flere trinn kan han så gå videre til den fullstendige grunndokumentasjon av de opplysninger han skal bruke.

Selve standardene må utarbeides sentralt.

(iv) Rutiner for ajourhold og avstemming

Rutinene må sikre:

- at nye opplysninger føyes inn i arkivet på en måte som er konsistent med det som allerede finnes, og slik at den samlede masse av opplysninger kommer best mulig til nytte
- spesielt må det sikres at endringer i omfang, definisjoner, eller spesifikasjoner blir registrert og dokumentert på en slik måte at brukeren blir gjort oppmerksom på hva de betyr for tolking og bruk
- Feil og suppleringer til allerede lagrede data skal helst rettes opp i arkivet, men dette kan ikke alltid gjennomføres, f.eks. av kostnadshensyn. Men enten feil rettes eller ikke, bør deres eksistens og hva som er gjort eller ikke gjort med dem registreres i standard dokumentasjonen.
- I noen tilfelle må data justeres før de kan brukes i visse anvendelser, f.eks. oppblåsning av utvalgstall eller tilpasning av eldre data til definisjoner som endres over tid (endring i kommunegrenser). Direktiver og spesifikasjoner for slike justeringer må også dokumenteres.
  Metodene må utarbeides sentralt.

C. Nærmere om fase III, iverksetting.

(i) Innpassing av eksisterende, lagrete data

Dette kan bli en omfattende jobb. På den annen side vil det vesentlige av den dokumentasjon som finnes kunne utnyttes når en gang systemet er utarbeidet (jfr. 2) og B)).

Alt behøver selvsagt ikke tas i en jafs. Det er en mulighet å innpasse mindre sentrale datamasser bare ettersom de faktisk skal hentes fram for bruk og særlig ressurskrevende lagringsformer, som f.eks. databaser må bare tas i bruk i den utstrekning det er økonomisk grunnlag for det. På den annen side kan vi på denne måten miste en del verdifullt materiale fordi det vil være gått for lang tid når vi endelig prøver å aktivisere informasjonen.

Arbeidet kan utføres på hver enkelt delmasse og krever innsats av emnespesialister og dataspesialister.

(ii) Innpassing av eksisterende registre

De eksisterende registre må trolig innpasses snarest mulig. Men her vil det etterhvert være opparbeidet det meste av den dokumentasjon som trenges. Det som ennå måtte mangle må etableres i alle fall.

Arbeidet vil trolig kunne gjennomføres av registerkontoret, i en viss kontakt med brukerne.

(iii) Innpassing av løpende statistikk.

Den løpende statistikken krever grundig gjennomtenking, særlig med sikte på samkobling og samordning av komplementære datakilder. Ordningen bør komme igang så snart som mulig. En god gjennomføring krever innsats på høyt nivå fra fagkontorene og datateknisk medvirkning.

(iv) Innpassing av ny statistikk.

Innpassingen i arkivsystemet med de nødvendige justeringer og full dokumentasjon må gjøres til en integrert del av produksjonsprosessen for ny statistikk.

D. Ressurs- og tidsplan

Skal prosjektet kunne gjennomføres må arbeidsgruppen settes i gang og få ressurser til å gjennomføre fase I og II av sitt oppdrag relativt raskt. Kanskje kan fase I gjennomføres tilstrekkelig grundig ved innsats av ca. 1/2 personårsverk.

Til fase II kunne en forsøksvis avsette 1 personårsverk, om nødvendig etter at fase I var gjennomført.

En vurdering av ressursbehovet for fase III må inngå i fase II. Her vil det måtte utarbeides særskilte faseplaner for innarbeiding av de ulike deler av Byråets datamasse. Ressursbehovet vil gjelde samordningen, men kanskje i særlig grad være knyttet til de enkelte delmasser. For delmasser som er kompliserte, og hvor det kreves kostbar lagringsteknikk (databaser) kan det bli spørsmål om betydelig ressursinsats,mens kanskje bare ukeverk eller timeverk er tilstrekkelig for andre deler av systemet.