Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20, 41 36 60

IO 73/8

5. mars 1973

METODEHEFTE NR. 6

Forarbeider til et variabelkatalog-prosjekt i Statistisk Sentralbyrå

INNHOLD

FORORD

Metodehefter i serien Arbeidsnotater

I tilknytning til mange prosjekter i Statistisk Sentralbyrå utarbeides det mindre, upretensiøse notater for avklaring av spørsmål av metodisk interesse. Det kan dreie seg om utvalgsteknikk, alternative spørsmålsformuleringer, presentasjonsmetoder, begrepsavklaringer, diskusjon av "funn" i data, systemidéer eller andre temaer. Selv om mange slike notater bare har begrenset interesse i ettertid, vil det blant dem være noen som kunne fortjene å bli mer alminnelig tilgjengelig enn de har vært hittil. Det kan også være nyttig å ha dem registrert sentralt slik at det blir lettere å få oversikt over det stoffet som foreligger, og å referere tilbake til det. Byrået har innført en publikasjonsordning for stoff av dette slaget ved at en leilighetsvis publiserer et passende antall slike notater samlet i et metodehefte i serien Arbeidsnotater.

Inneværende hefte er det sjette av denne typen. Det inneholder notater vedrørende den variabelkatalog som overveies innført i Byrået. Det første notatet i heftet (EH/IH, 8/2-72) beskriver idéene bak prosjektet. Sett i sammenheng gjenspeiler notatene prosjektets nylige forhistorie.

Forsker Jan M. Hoem er oppnevnt som redaktør av metodeheftene. Assistent Liv Hansen er redaksjonssekretær. Medarbeidere i Byrået som lager stoff som kan være aktuelt, bes sende dette til redaksjonen etter hvert som det blir ferdig.

Kontorlederne bes holde øynene åpne for denne nye publiseringsmuligheten.

EH/IH, 8/2-72

VARIABELKATALOG-PROSJEKTET - FORELØPIGE SYNSPUNKTER

av Eivind Hoffmann

Innledning

Av referatet [2] fra møtet hos Bjerve 12. oktober 1971 der arbeidet med et system for sosio-demografisk statistikk (SSDS) ble diskutert, går det fram at arbeidet med oppbyggingen av en variabelkatalog "må komme godt igang i løpet av 1972". I mitt notat av 28/9-71 [1] som var utgangspunktet for diskusjonen på møtet, uttalte jeg at noe arbeid burde begynne allerede i 1. eller 2. kvartal 1972. Dette notatet er skrevet med håp om at det kan være en hjelp ved planleggingen og igangsettingen av arbeidet.

Målsetting

Variabelkatalogen skal være en veiviser i norsk statistikk - bygget opp på en systematisk måte og slik at den kan være et hjelpemiddel for dem som skal bruke statistisk informasjon og samordne statistisk informasjon fra ulike kilder. Ved hjelp av variabelkatalogen bør man være i stand til å svare på de fleste spørsmål av typen: Hvilke data har Byrået om vårt samfunn? Hvor og hvordan fins de lagret? Hvilke enheter omfatter de? Kan enhetene identifiseres? Hvilke kilder har data? Hvilke bearbeidinger er gjort av disse data? Hvilke standarder er benyttet? Hvordan er kvaliteten av data? Kan de brukes sammen med andre data om de samme enheter på samme tidspunkt eller på andre tidspunkt? Hvorfor kan de eventuelt ikke brukes sammen med bestemte andre data? Hvordan kan man få adgang til data? Denne listen over spørsmål som variabelkatalogen bør kunne gi svar på, viser at oppbyggingen av en variabelkatalog vil være et stort og tidkrevende prosjekt.

Tidligere arbeid

Uten at det - såvidt jeg vet - foreligger noe skriftlig, har Nordbotten og Øien diskutert en del av de prinsipper som en slik variabelkatalog må bygges opp etter - kanskje særlig den delen av katalogen som jeg nedenfor betegner med del III. Selsjords notat av 6/1-70 til Nordbotten med "Oversikt over en del grupperinger som nyttes i personstatistikken" kan anses å være et første forsøk på å lage en del av det jeg nedenfor har betegnet som variabelkatalogens del II. I Sverige har

det pågått noe variabelkatalogarbeide i forbindelse med prosjektet ARKSY (gjelder særlig del III) og i forbindelse med arbeidet med et SSDS (gjelder særlig del II). I Danmark har man laget en oversikt av samme type som Selsjords notat.

Skisse til opplegget av katalogen

Etter mitt syn må det være rimelig å tenke seg katalogen bygget opp med tre deler:

Del I vil spille omtrent samme rolle overfor publikum som Veiviser i norsk statistikk gjør idag. Den bør inneholde et stikkordregister med henvisninger til katalogens deler II og III, samt en knapp beskrivelse av de mest sentrale deler av statistikken. Publikum skal på grunnlag av del I kunne stille spørsmål av den type som er referert ovenfor, og få svar fra delene II og III. Det er neppe hensiktsmessig at de har direkte adgang til II og III.

Del II bør inneholde en systematisk og detaljert katalog over den statistiske informasjon som er lagret i bearbeidet form - for det meste som tabeller - i publikasjoner, på kontorer, og på mikrofilm og på magnetbånd.

Del III bør inneholde en systematisk og detaljert katalog over den statistiske informasjon som er lagret på maskin-media på individ-(enhets)-basis. Både del II og del III bør gi slike opplysninger at man kan svare på flest mulig av de spørsmål som er nevnt ovenfor.

Omfanget av katalogen

I prinsippet bør katalogen omfatte all statistisk informasjon i Byrået. I praksis bør man antakelig begynne med å la den omfatte all statistisk informasjon som er kommet til Byrået etter et bestemt tidspunkt, for så senere å utvide katalogen til å omfatte eldre data i Byrået og data utenfor Byrået.

Utgangspunktet for arbeidet

Antakelig bør arbeidet med delene II og III starte noenlunde samtidig og gå parallelt. Utgangspunktet for arbeidet med del II kan kanskje være Selsjords notat, kodelister, Bjørnstads skjemaregister, oversikter ved fagkontorene og det arbeidet som er utført og blir utført i Sverige (se ovenfor). Det bør foregå i nær kontakt med SSDS- og SNA-arbeidet. Utgangspunktet for arbeidet med del III kan kanskje være

Systemkontorets filebeskrivelser, skjemaregisteret og kanskje ARKSY-arbeidet i Sverige. Antakelig bør det også her være nær kontakt med SSDS- og SNA-arbeidet.

Organisering av arbeidet, ressurser

Arbeidet vil kreve minst én "voksen" person på full tid. Vedkommende bør ha godt kjennskap til Systemkontorets rutiner og filebeskrivelser, og han bør ha hatt god kontakt med fagkontorenes arbeid. Det bør etableres en støttegruppe for arbeidet med en representant fra fagkontorene, en representant fra de som utnytter statistikken analytisk og en person med god innsikt i systemarbeid. Støttegruppens oppgave bør være å gi impulser, idéer og kritikk til dem som har det daglige arbeid og å sørge for å ivareta de hensyn som den representerer. Støttegruppen bør også ha som oppgave å holde kontakten med arbeidet i Norsk Samfunnsvitenskapelig Datatjeneste, som driver et beslektet arbeid - for det meste for data fra Byrået.

[1] Arbeidet med et system for sosio-demografisk statistikk(SSDS): Forsøk på konkretisering. Notat EH/San, 28/9-71

[2] Referat fra møte hos Bjerve, 12. oktober 1971. Notat EH/GH, 23/10-71

SN/WA, 31/5-72

PROSJEKTET VARIABELKATALOG OG ET REFERANSESYSTEM TIL ARKIVERT NUMERISK INFORMASJON

av Svein Nordbotten

1. Innledning

Begrepet variabelkatalog har gjennom flere år gått igjen i arbeidsprogrammer og prosjektbeskrivelser. I den senere tid har det bl.a. vært trukket fram i forbindelse med Byråets opplysningsvirksomhet og markedsføring av statistikkprodukter, som hjelpemiddel for å orientere om innholdet av Byråets dataarkiver, utvikling av sosiodemografiske regnskapssystemer, sosialstatistikk, maskinelle framhentingsmetoder, m.m. (jmfr. Eivind Hoffmanns prosjektbeskrivelser, arbeidsprogram og hans notat av 8/2-72 om: Variabelkatalogprosjektet - Foreløpige Synspunkter). Utenfor Byrået begynner flere også å bli opptatt av å lage et system som gjør det mulig å holde oversikt over informasjon som lagres i forskjellige systemer før det hele blir for sent på grunn av det raskt voksende volum. Rasjonaliseringsdirektoratet har derfor tatt initiativ til å få utredet spørsmålet om en registrering av data som finnes lagret omkring i de forskjellige statlige institusjoner. Det utvalg som skal utrede spørsmål om personlighetsvern i tilknytning til offentlige dataarkiver vil naturlig nok også ta opp spørsmålet i en videre sammenheng. Norsk Samfunnsvitenskapelig Datatjeneste, finansiert av NAVF, har som et av sine hovedformål å samle statistiske data og tilby de på en slik måte at de lettest mulig vil kunne utnyttes av samfunnsvitenskapelige forskere. Dette har ført til at NSD har satt i gang og utarbeidd en katalog over statistikk dette organ har arkivert.

Betegnelsen "variabelkatalog" representerer imidlertid ikke noe entydig begrep med alminnelig akseptert innhold og peker i de fleste tilfelle bare på en enkel komponent i et større system. Jeg skal i det følgende prøve å klarlegge hva jeg legger i variabelkatalog og hvordan denne inngår i et referansesystem for de statistiske arkiver.

2. Problemstilling

Når vi arbeider med et samfunnsvitenskapelig problem følger vi ofte en framgangsmåte som skissert i figur 1. I to faser av arbeidet, illustrert ved boks 2 og 4, søker vi å eliminere unødig arbeid ved å gjøre bruk av resultater av arbeid andre tidligere har oppnådd.

Når det gjelder fase 2, er framgangsmåten å gå til et fagbibliotek og søke i dets kataloger og indekser etter referanser til eventuelle metoder og løsninger som kan nyttes i det foreliggende problem. Bibliotekenes kataloger er fra gammelt av systematisk oppbygd og vedlikeholdt, og gir et fint grunnlag for søking. Det kan vel føyes til at det for tiden også foregår stor aktivitet med sikte på å forbedre bibliotekenes katalogsystemer og indekseringer.

I fase 4 søker vi etter relevant numerisk informasjon, dvs. statistikk og individualdata. Til tross for at slik informasjon er enklere strukturert og mer standardisert enn tekstlig informasjon, finnes det likevel ikke noe utbygd sidestykke til bibliotekenes katalogsystemer for numerisk informasjon. Når selv ikke Byrået, som må anses som det nasjonale skattkammer for numerisk informasjon om samfunnsforhold, har noe systematisk opplegg for søking og eventuell framhenting av numerisk informasjon, skyldes dette at en tidligere oppfattet all informasjon utenom de trykte publikasjoner på grunn av de betydelige kostnader som framhenting ville medføre, som utilgjengelig i praksis. Den statistiske årsproduksjon som ble gitt ut i 10-20 publikasjoner pr. år, ble videre betraktet som så oversiktlig at det ikke var behov for noe søkesystem for å finne fram til den ønskede informasjon. Etter at moderne databehandlingsutstyr ble tatt i bruk og arkivstatistiske metoder innført, har situasjonen endret seg vesentlig. Tallet på publikasjoner fra Byrået pr. år er nå i størrelsesorden 100, dvs. det utgis 3 000 - 4 000 tabeller pr. år, og et stort antall datasett på lavere bearbeidingsnivå enn publiserte tabeller er lettere tilgjengelig for spesialbearbeiding.

Den måte framleting av numerisk informasjon har foregått på hittil kan kanskje illustreres ved figur 2. En bruker med behov for numerisk informasjon starter vanligvis ved å studere Vegviseren, Publikasjoner fra Statistisk Sentralbyrå, etc. eller ved personlig å konsultere et fagkontor vedkommende mener bør være det riktige å henvende seg til. Dette fører trolig til at det vises til en eller flere publikasjoner. Ved oppslag i tabellregisteret ledes brukeren videre til en eller flere tabeller, og til en eller flere tabellkolonner og -linjer som forhåpentligvis gir relevant informasjon. Ulempene med et slikt system er at det er stor risiko for at det også i andre publikasjoner og tabeller kan finnes relevant informasjon som publikasjons- og tabellnavn ikke avslører. Bortsett fra spredte fotnoter gir systemet heller ikke noen referanse til utrykte informasjonssett, og er derfor utilstrekkelig.

Det ovenfor beskrevne system kan karakteriseres som en systematisk fortegnelse over informasjonssett - publikasjoner og tabeller - som i tabellhoder og forspalter gir referanse til den numeriske informasjon for de begreper (objekttype, objektsamling, kjennemerke, tidsrom) de enkelte informasjonssett belyser. Oversikt over hva som finnes av trykt statistikk om et bestemt emne krever en mer eller mindre fullstendig gjennomgang av alle tabeller. Når det gjelder innholdet i uttrykte datasett som holdes på et eller annet maskinlesbart medium, er situasjonen enda mindre tilfredsstillende. Selv for en byråfunksjonær kan det by på betydelige problemer å få presise opplysninger om innholdet av et gitt informasjonssett. Å få oversikt over hva som finnes om et bestemt emne er i dag ikke mulig.

Det system vi har behov for innebærer at hvert enkelt informasjonssett - publikasjon, tabell, aggregatfiler, og samlinger av individualdata - beskrives på en slik måte at vi kan lage en systematisk liste over de forskjellige begreper som forekommer med henvisninger til de informasjonssett som på en eller annen måte bidrar til å belyse emnet.

3. Et system for å referere arkivert numerisk informasjon

3.1 Systemstruktur

En variabelkatalog er en del av et "information retrieval system". I figur 3 er dette systemet, her betegnet referansesystemet, avgrenset av det prikkede rektangel som igjen kan oppfattes som en del av det arkivstatistiske system representert ved hele figuren.

I det arkivstatistiske system arkiverer statistikkprodusenten numerisk informasjon, individualdata og statistikk, i et data- og statistikkarkiv, F. Denne informasjonen er, eller bør være, nærmere begrepsmessig og operasjonelt definert i et sett med dokumenter - planleggingsnotater, skjemaer, gjennomføringsrapporter, kvalitetsvurderinger m.m. - som er/bør være systematisk satt opp i standardisert form og arkivert tilgjengelig i et dokumentarkiv, P. Referansesystemets sentrale del er en ordnet innholdsfortegnelse, referanseregisteret, R, over hva som finnes arkivert av numerisk informasjon og dokumenterte spesifikasjoner i det arkivstatistiske systemet. Referanseregisteret holdes løpende ajour ved produktbeskrivelser av nye informasjonssett som settes inn i arkivene. Brukere som søker informasjon, formulerer en etterspørselsbeskrivelse og søker etter tilsvarende produktbeskrivelse i referanseregisteret. De funn som gjøres, kommer fram som referansemeldinger med henvisning til hvor i

arkivene det eventuelt finnes informasjon og dokumentasjon av den type som søkes og som gir grunnlag for uttak. Hvor vellykket søkingen vil være, avhenger blant annet av om produsent og bruker uttrykker seg på en noenlunde ensartet måte. For å ta vare på dette krav, må både produsent- og etterspørselsbeskrivelser utformes i et kontrollert, felles beskrivelsesspråk, S.

I det skisserte system er data- og statistikkarkivet den komponent som nå er mest utviklet i Byrået. Dokumentasjonen av innsamling og bearbeiding er varierende, ustandardisert og spredt på forskjellige steder. Oppretting av et systematisert dokumentarkiv bygd på standardisert dokumentasjon av begrepsdrøftinger og -definisjoner, og operasjonelle definisjoner bygd på beskrivelser av innsamling og bearbeiding, bør være en viktig oppgave å løse. Det sentrale skjemaarkivet som foreløpig er det eneste som finnes i systematisert form, kan her danne et naturlig utgangspunkt.

I dette notatet skal vi gå litt nærmere inn på referansesystemet og dets komponenter.

3.2 Beskrivelsesspråk

La oss betrakte en gitt produktbeskrevet mengde informasjon symbolisert ved den opptrukne sirkelen A i figur 4. Vi tenker oss videre at den prikkede sirkelen B representerer den mengde som en bruker får henvisning til når han på sin måte lager en etterspørselsbeskrivelse av det han tror er mengden A.

Vi får da tre delmengder:

- mengden (1) er relevant informasjon som bruker ikke får,
- mengden (2) av relevant informasjon som brukeren får,
- mengden (3) av irrelevant informasjon som brukeren får.

Den ideelle situasjon vil være inntruffet om sirklene A og B er sammenfallende. I praksis vil imidlertid produsent og bruker uttrykke seg på forskjellig måte og brukeren vil både få henvisning til irrelevant informasjon og mangle henvisning til relevant informasjon. Forholdet mellom mengdene (2) og (A) betegnes ofte "recall", mens forholdet mellom (2) og (B) kalles "precision". Begge forholdstall vil ligge i intervallet 0 til 1, med 1 som den verdi en søker å oppnå. Byråets nåværende referansesystem som skissert i avsnitt 2, vil ha en relativ lav "recall" verdi.

Målsettingen vil derfor være å lage et beskrivelsesspråk som bidrar til at de to nevnte forholdstall begge blir så nær 1 som mulig.

Forbedring av "recall" kan gjøres ved å karakterisere hvert informasjonssett med flere begreper (f.eks. mer omstendelige publikasjons- og tabellnavn). Forbedring av "precision" oppnås ved å bryte opp hvert informasjonssett i flere og mer homogene sett. Begge framgangsmåter fører med seg økte kostnader. Kostnadsfunksjonen vil vanligvis være slik at med gitte ressurser vil en forbedring av det ene forholdstall føre til en reduksjon i det andre. Uten å ha noen vurderingsfunksjon for de to egenskaper, vil målsettingen derfor i praksis modifiseres til å lage et beskrivelsesspråk som for en gitt "precision" maksimerer "recall", med andre ord, vi godtar en viss prosent av irrelevante referanser, men krever flest mulig henvisninger til relevant informasjon.

Et beskrivelsesspråk vil bestå av to hovedkomponenter, en ordliste og et regelverk for å knytte ordene sammen i beskrivelser. Vi kan tenke ordlisten som sammensatt av fire dellister som representerer "ordklasser":

- liste over objekttyper
- liste over navn
- liste over kjennemerker ved objekter
- liste over tidsspesifikasjoner.

Listen over typer av objekter vil omfatte alle de statistiske enhetstyper som inngår i Byråets undersøkelser, slik som personer, familier, husholdninger, bedrifter, foretak, kommuner, biler, eiendommer, osv. Disse objekttypene angir at beskrivelsen angår et informasjonssett med informasjon om individuelle enheter. Tilsvarende omfatter listen også statistiske aggregatenheter som personaggregater, bedriftsaggregater, osv. og som angir at det pekes på et informasjonssett med statistikk for klasser av personer, bedrifter, osv. For å kunne skille mellom situasjoner når et enkelt objekt beskrives og når en samling objekter beskrives, vil listen omfatte både entalls- og flertallsformer på de forskjellige objekttyper.

Den vanskeligste av dellistene er listen over navn på de enkelte objekter og objektsamlinger.

Navnet beskriver informasjonssettets utstrekning. Vi kan tenke oss navnelisten bygd opp på:

- registre over individualenheter, og
- statistiske klassifikasjoner.

Navnene på de forskjellige registre - inklusive spesialregistre over utvalgsenheter etc. - vil, avhengig av objekttypen:

a. avgrense en samling individualenheter,
b. identifisere en objektklasse.

De enkelte enhetsnavn - identifikasjonsnummer - tillater den enkelte individualenhet beskrevet.

Objekttype: Personer, og navn: Personregister, vil eksempelvis peke på en samling av objekter, mens objekttype: Personaggregat, og navn: Personregister, på den annen side vil angi en statistisk opplysning om bestanden i personregisteret. Objekttype: Person, og navn: xx xx xx xxx xx, vil peke på det enkelte individ.

Navnet på en klasse i en klassifikasjon vil, avhengig av objekttype, enten

a. avgrense en samling individualenheter,
b. avgrense en samling av subklasser,
c. identifisere en objektklasse.

Objekttype: Personer, og navn: Kontorfunksjonærer, objekttype: Personaggregater, og navn: Yrkesklassifikasjon, og objekttype: Personaggregat, og navn: Kontorfunksjonærer, er eksempler på de tre alternative avgrensninger.

Listen over kjennemerker er det vi vel vanligvis tenker på når vi snakker om en variabelkatalog. Denne listen omfatter både de kjennemerker, datakatalogen, vi observerer i tilknytning til de individuelle primærenheter og de kjennemerker, statistikk-katalogen, vi beskriver objektklasser med. Datakatalogen vil også omfatte navnene på våre klassifikasjoner fordi disse også gir uttrykk for kjennemerker ved den enkelte individualenhet.

Variabelkatalogen vil delvis bestå i objektspesifikke kjennemerker. Ekteskapelig status, utdanning o.l. er personspesifikke kjennemerker, mens bruttoproduksjonsverdi og bearbeidingsverdi er bedriftsspesifikke kjennemerker. Andre kjennemerker som alder kan være felles for flere objekttyper. Det er rimelig å tenke seg variabel- eller kjennemerkekatalogen redigert etter kjennemerkenes objekttilknytting.

Listen over tidsspesifikasjoner omfatter alle punkter og intervaller på tidsaksen som nyttes i statistikkproduksjonen. Vi angir punktene på tidsaksen ved åtte siffer, f.eks. 1970.11.01 som tidspunktet for Folketellingen.

Ved å holde en stram kontroll med de betegnelser som innføres i ordlistene slik at det bare er de mest entydige betegnelser som tillates

nyttet, kan dette bidra til å høyne "precision". En måte til å høyne "recall" er å føre inn de vanligste synonymer som likeverdige ord i listene.

Reglene som styrer sammensetningen av ordene i setninger er beskrivelsesspråkets annen komponent. I 1960-årene ble spørsmålet om fritt kontra fast setningsformat inngående drøftet. Fast format er det enkleste og sannsynligvis vil det være en fordel å starte med det. Vi forutsetter derfor at en setning alltid er skrevet slik:

objekttype/navn/kjennemerke/tidsspesifikasjon.

Mellom objekttype, navn og kjennemerke har vi følgende sammenhenger:

Objekttype	Navnetype	Kjennemerke	Angir
Individualenhet			
Entallsform	Individnavn	Individual-kjennemerke	Individual-enhet
Flertallsform	Registernavn Klassenavn	Individual-kjennemerke	Individual-enheter i registeret eller klassen
Aggregatenhet			
Entallsform	Registernavn Klassenavn	Aggregat-kjennemerke	Aggregat-enhet
Flertallsform	Klassenavn	Aggregat-kjennemerke	Aggregat-enheter for subklasser

3.3 Beskrivelser

En beskrivelse av et informasjonssett som arkiveres eller søkes omfatter to eller flere linjer. Hver linje er delt i felter adskilt med tegnet /. Første linje kan vi tenke oss slik:

l.nr. / beskrivelsestype / identifikasjon / tekniske data.

Linjenummeret lar vi alltid være 0 i første linje. Beskrivelsestype er enten Produktbeskrivelse eller Etterspørselsbeskrivelse, identifikasjon kan være navnet på den som har laget beskrivelsen. I Produktbeskrivelsen gir tekniske data referanseadresser til dataarkiv og dokumentarkiv for den lagrede informasjon som beskrives. I Etterspørselsbeskrivelser kan feltet tekniske data spesifisere om det er referanser til dokumentarkiv,

dataarkiv eller begge som ønskes.

De etterfølgende linjer i beskrivelsen er setninger som beskriver informasjonssettet som arkiveres eller lagres knyttet sammen med operasjonssymboler. Formen for hver linje kan vi forestille oss slik:

l.nr. / objekttype / navn / kjennemerke / tidsspesifikasjon / operator.

De operasjoner vi har behov for vil være:

E : logisk "eller",
O : logisk "og",
N : logisk "ikke",
. : stopp.

For å redusere unødig skriving tillater vi gjentakelsestegnet " i felter fra og med tredje linje. Det indikerer at innholdet er det samme som i tilsvarende felt i linjen over.

La oss ved eksempler illustrere hvordan vi kan tenke oss produktbeskrivelser utformet. Et utsnitt fra beskrivelse av personfilen i Folketellingen 1970 kan vi tenke oss slik:

```
0   / Produktbeskrivelse / ......          / ......
.....................................................................
N   / Personer          / Personreg.nov.70 / Fødselsdato/ 1970.11.01/ 0
N+1 /     "            /          "       / Yrke       /      "    / 0
.....................................................................
```

"Personer" er et ord i listen over objekttyper og angir at det er en beskrivelse av en samling personer. I listen over navn, forekommer Personreg. nov. 70 som en avgrensing av samlingen personer. Fødselsdato og Yrke er kjennemerker i variabelkatalogen og 1.11.70 et punkt på tidsaksen. Operatoren 0 angir at data for både Fødselsdato og Yrke pr. 1.11.70 forekommer koblet i tilknytting til hver person i samlingen. Slik vi her tenker oss Produktbeskrivelsene - en for hvert informasjonssett - vil E og N ikke forekomme i denne beskrivelsestype.

Når Folketellingen er bearbeidd vil det bl.a. foreligge en tabell over antall personer i kryssklassifisering mellom alders- og yrkesklasser. Vi kan forestille oss denne tabellen beskrevet ved:

```
0 / Produktbeskrivelse / ..........  / ..........
1 / Personaggregater   / Utdanningsklassifikasjon / Sum / 1970.11.01/ 0
2 /        "           / Yrkesklassifikasjon      / "   /     "     / .
```

Beskrivelsen angir at det er en samling aggregater som beskrives. Målet på hver klasse er Sum personer.

Brukere vil ha behov for å ha et symbol som betegner at innholdet i et bestemt felt er likegyldig. Vi vil bruke tegnet - for dette formål. Vi tenker oss først en bruker som vil undersøke om det foreligger informasjon som gir yrke, kjønn og alder på samlingen av alle personer som var med i Folketellingen 1960 og om dette kan kombineres med deres yrke og inntekt i 1970.

Beskrivelsen vil være:

```
0 / Etterspørselsbeskrivelse / ..... / .....
1 / Personer / Personreg. nov. 1960 / Yrke        / 1960.11.01 / 0
2 /    "     /          "          / Kjønn       /      "     / 0
3 /    "     /          "          / Fødselsdato /      "     / 0
4 /    "     / Personreg. nov. 1970 / Yrke        / 1970.11.01 / 0
5 /    "     /          "          / Inntekt     / 1970.- .-  / .
```

Personer angir at vi ønsker referanse til en eventuell samling av individer som både finnes i Personregister nov. 60 og i Personregister nov. 70. For hvert individ ønsker vi de angitte kjennemerker. At operatoren 0 forekommer her betyr ikke at den søkte informasjon nødvendigvis må foreligge i ett informasjonssett.

En annen bruker er interessert i tallet på døde etter yrke eller etter næring i 1960-årene, og vi kan tenke oss at følgende beskrivelse som illustrerer blandet bruk av operatoren:

```
0 / Etterspørselsbeskrivelse / ..... / .....
1 / Personaggregat / Yrke   / Sum / 196-.-.- / 0
2 /      "        / Døde   / "   /     "    / E
3 /      "        / Næring / "   /     "    / 0
4 /      "        / Døde   / "   /     "    / .
```

Bruk av likegyldighetstegnet betyr i virkeligheten at vi ber om å få referanser svarende til alle de mulige spesifikasjoner som kunne ha stått i feltet.

3.4 Koding

For at beskrivelsene skal kunne behandles effektivt av systemet, må de gjøres mer "behandlingsvennlige" ved en kodeprosess hvor ordene omgjøres til koder.

Kodeprosessen har tre formål:

- å standardisere alternative synonymbetegnelser
- å introdusere relasjoner mellom klasser i et klassifikasjonshierarki
- å effektivisere søkningen

Kodeprosessen foregår ved hjelp av en kodeliste hvor hvert ord i ordlisten har en linje med en kode. Hvert begrep har sin særskilte kode som alle synonyme betegnelser for begrepet har.

Kodens generelle form kan være:

A.B.C....

hvor A er koden for et begrep mens A.B og A.C er koder for underordnede begreper. For objekttyper kan følgende tjene som eksempel:

A : Personer

A.B: Person

C : Personaggregater

C.D: Personaggregat.

Personregisteret er en monotont voksende samling personer, A. Personregisteret nov. 60 er en delsamling med kode A.B, mens Personregisteret nov. 70 er en annen delsamling, A.C, av A.

For klassifikasjoner er dette en velkjent hierarkisk opplysning:

A : Næringen Industri

A.B : Næringsgrenen Næringsmidler

A.B.C: Næringsgruppen Bakerier

osv.

Kodeoppbygningen er viktig for at systemet blant annet skal kunne identifisere subklasser i en setning av typen

objektaggregater / navn / ..... / .....

Dersom navnet har koden A vil de objektaggregater en søker informasjon om ha koder A.x hvor forekommende verdier av x spesifiserer et aggregatobjekt i den søkte samlingen.

Når det gjelder kjennemerker og tidsspesifikasjoner forutsetter vi en tilsvarende oppbygning.

Kodingen kan enten foregå manuelt under utarbeiding av beskrivelsene eller ved maskinelle tabelloppslag på grunnlag av tekstlig utformede beskrivelser.

3.5 Referanseregisteret

Referanseregisteret er systemets sentrale komponent og omfatter alle produktbeskrivelser. Mot dette registeret sammenliknes alle Etterspørselsbeskrivelser.

Vi kan betrakte registeret som en matrise eller tabell som i Figur 5. En kolonne i en slik tabell utgjør en beskrivelse av et informasjonssett, mens hver linje svarer til et begrep. Avhengig av

hvordan en vil søke, kan et slikt register ordnes etter:

- beskrivelse x begreper
- begrep x beskrivelser

Dersom formålet for søkingen er å finne ut hva spesifiserte undersøkelser (representert ved beskrivelser) omfattet, vil den første organisasjonsform være å foretrekke. Det systemet vi hittil har hatt, må i stor utstrekning karakteriseres som et system beslektet med den første typen. Vår problemstilling er imidlertid å finne hvilke informasjonssett (beskrivelser) som inneholder gitte begreper og vi vil derfor foretrekke den andre organisasjonsformen, som kan illustreres ved et system som vist i figur 6.

Ved søking i registeret slås det først opp i tabell B, forspalten, på de begreper som inngår i etterspørselsbeskrivelsen. Hvert oppslag gir en adresse til en linjetabell, L.a., som inneholder en ordnet kjede med referanser til de beskrivelser, S.b, S.c, S.d. ..., som omfatter vedkommende begrep. Kjedene kolleres og ferdige referanser ekstraheres i samsvar med etterspørselsbeskrivelsens operatorer.

En produktbeskrivelse legges inn ved at det i tabell B slås opp på hvert begrep som det vises til og at beskrivelsens adressereferanse legges inn i hver linjetabell som oppslagene i tabell B viser til.

Dersom produktbeskrivelsen inneholder et nytt ord kan dette være:

- et synonym til et allerede eksisterende begrep, eller
- en betegnelse på et nytt begrep.

I det første tilfellet må ordet føres inn i kodelisten med koden for begrepet. Når ordet betegner et nytt begrep må dessuten et nytt felt opprettes i B-tabellen med referanse til en ny linjetabell for vedkommende begrep.

3.6 Referansemeldinger

Resultatet av behandlingen av en etterspørselsbeskrivelse er utskrift av en referansemelding. Referansemeldingen skal gi opplysning om:

- hva som finnes av søkt informasjon i dokumentarkiv og data- og statistikkarkiv,
- den form informasjonen finnes på,
- hvor den finnes,
- hvilke betingelser som er knyttet til informasjonen og
- eventuelt hvorfor søkningen ikke har gitt noen referanser.

De fire første kategorier opplysninger er basert på det som finnes i feltet Tekniske opplysninger i første linje i Produktbeskrivelsene. Den siste kategori hentes fra systemet som standardmeldinger på situasjoner som oppstår, som feil i utforming av beskrivelsen, etc.

Når et system av denne type er operativt vil det også være naturlig å tilby individualisert informasjonstjeneste på løpende basis om nye innlegg i arkivene av interesse for brukeren. En slik tjeneste vil bygge på faste etterspørselsbeskrivelser som gir uttrykk for de enkelte brukeres "interesse-profil" og som regelmessig vil bli sammenholdt med referanseregisteret.

4. Oppbygging av et referansesystem i Byrået

4.1 Arbeidsoppgaver og arbeidsprogram

Oppbygging av et referansesystem av den type som ble skissert i avsnitt 3 vil være en omfattende oppgave og sannsynligvis kreve større innsats enn hva de fleste kanskje forestiller seg.

Hoffmann har i sitt notat av 8/2-72 skissert en oppbygging i tre deler. Del I svarer vel stort sett til min ordliste, mens Del II og Del III til beskrivelser for henholdsvis statistikkarkiv og dataarkiv. Hoffmann foreslår at en lar arbeidet med Del II og III gå parallelt etter at Del I er etablert (?).

Jeg vil foreslå følgende framdrift som kan betraktes som en videre utbygging av Hoffmanns forslag.

Fase 1: Begrenset referansesystem

Oppgave 1.1: Ordliste for kjennemerker med referanse til NOS. Arbeidet forutsetter en systematisk gjennomgang av NOS-publikasjonene (begrenset f.eks. til 2-3 siste år) med nedtegning av kjennemerkenavn og tilhørende publikasjons- og tabellidentifikasjoner (jmfr. det som ble gjort ved utarbeiding av norsk-engelsk ordliste), med etterfølgende påføring av "se også" for synonyme eller nesten synonyme betegnelser.

Oppgave 1.2: Ordliste for kjennemerker med referanse til arkiver på maskinmedia. Utgangspunktet vil være filebeskrivelsene med tilleggsinformasjon fra det sentrale skjemaregisteret. Denne ordlisten må samarbeides med den som er nevnt under oppgave 1.1.

Ordnet på en hensiktsmessig måte vil ordlistene med referanser fungere som et begrenset referansesystem. Oppslag i ordlisten vil gi henvisninger til i hvilke publikasjoner/tabeller og i hvilke filebeskrivelser/arkivfiler vi vil finne informasjon om de kjennemerker vi

søker, eventuelt supplert med henvisninger til andre kjennemerkenavn som kan dekke de begreper vi søker å belyse.

En invertering av disse ordlistene vil gi en oversikt over hvilke kjennemerker hver publikasjon og file omfatter (jmfr. kolonnene i figur 5).

Fase 2: Innføring av tidsspesifikasjoner

Oppgave 2.1: Ordliste over tidsspesifikasjoner. Etterhvert som referansesystemet vokser og dekker flere årganger, vil det være ønskelig å kunne skille mellom henvisninger til gammelt og nytt materiale. En ordliste over tidsspesifikasjoner med referanse til publikasjoner og filer lar seg mest hensiktsmessig utarbeide ved å ta utgangspunkt i den inverterte kjennemerkeordliste (dvs. listen over publikasjoner/tabeller og filer) og tidfeste de enkelte kjennemerker som forekommer.

Isolert vil ordlisten med referanser gi oversikt over hvilke informasjonssett som refererer seg til de enkelte tidspunkter eller -perioder. Kombinert med ordlisten for kjennemerker gir den muligheter til å selektere referanser til publikasjoner/tabeller og filer hvor bestemte kjennemerker forekommer for bestemte tider.

Fase 3: Utbygging av et dokumentarkiv

Oppgave 3.1: Dokumentarkivet må bygge på en samling notater i en standardisert form for hver undersøkelse. Prosjektskisser, -beskrivelser, planleggingsnotater, skjemaer, kodelister, revisjonsinstrukser, programhenvisninger, klassifikasjoner, tabellspesifikasjoner, kvalitetsvurderinger, kostnadskomponenter m.m. er komponentene i det som bør utgjøre dokumentasjoner av en undersøkelse og som skal gi brukerne presis informasjon om statistikk og data.

Det må først utarbeides en modell for hvordan en komplett dokumentasjon skal se ut, og deretter må en prøve å få dokumentert de viktigste undersøkelser. Deretter må hver dokumentasjon gis entydig identifikasjon og kjennemerkelisten påføres referanser til disse identifikasjoner.

Fase 4: Etablering av objekttype- og navnelister

Oppgave 4.1: Ordlister for objekttyper og navn. For å redusere omfanget av henvisninger til informasjonssett for objekttyper, objekter og objektsamlinger (øke "precision"), innføres objekttype og navnelister.

Oppretting av navnelistene forutsetter at det foreligger detaljerte beskrivelser av de bestander som er observert eller av de klasser

som inngår i informasjonssettet. Det er derfor rimelig at de opprettes etter fase 3 som vil gi de nødvendige opplysninger.

Fase 5: Automatisering og integrering med arkivsystemet

Oppgave 5.1: Automatisering. Det er rimelig å anta at det gjennom alle faser vil bli gjort bruk av tekniske hjelpemidler i større eller mindre grad, men i den avsluttende fase må referansesystemet automatiseres som et "information retrieval" system for rask reaksjon på forespørsler.

Oppgave 5.2: Integrering. Hittil har vi betraktet systemets output som referanser til den numeriske informasjon vi søker. Det endelige mål vil være å få systemet integrert i det arkivstatistiske system for direkte output av numerisk informasjon.

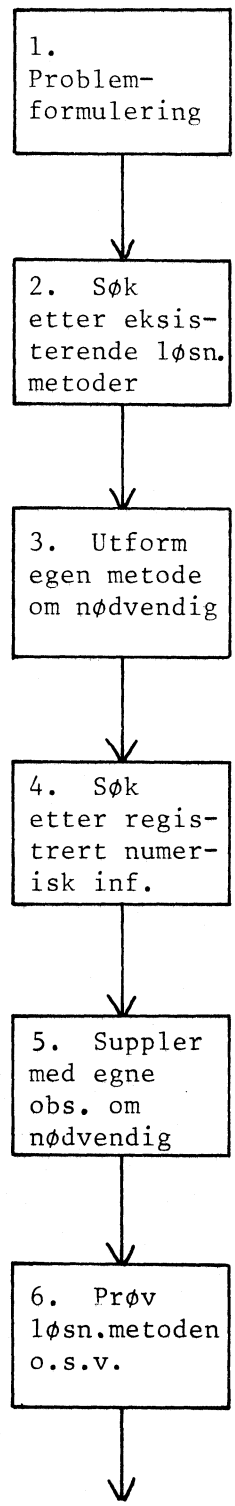
5.2 Arbeidsdeling

Utviklingen av et system som skissert vil forutsette team-work av representanter fra flere grupper. Jeg kan se følgende med sentrale interesser og oppgaver:

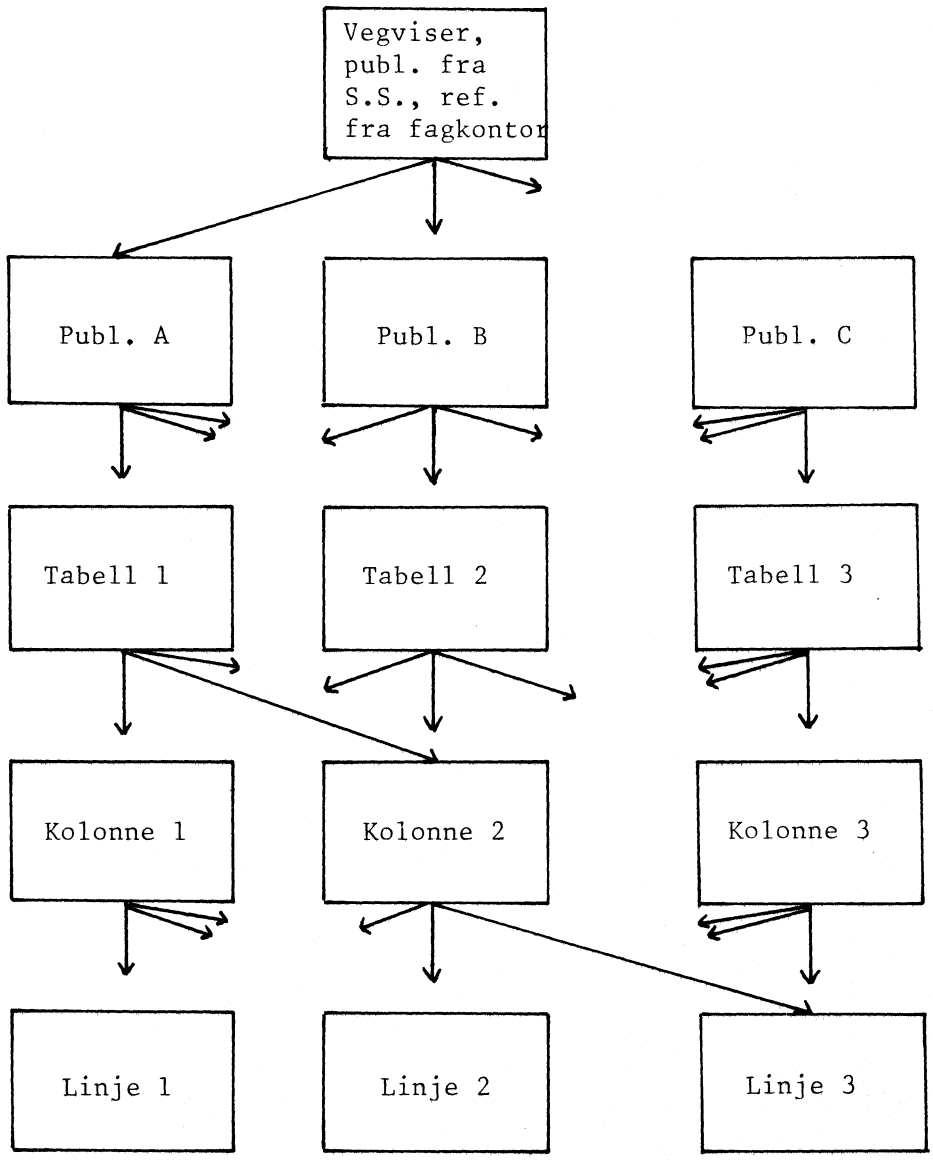
Informasjonskontoret (for Fagavdelingen)
Sosio-demografisk forskningsgruppe
Nasjonalregnskapskontoret (for Forskningsavdelingen)
Systemkontoret (for Produksjonsavdelingen)

Det bør sannsynligvis være to personer, en med faglig-metodisk bakgrunn og en med system-EDB bakgrunn, som samarbeider om prosjektet med en gruppe av representanter fra de nevnte organer som et rådgivende utvalg.

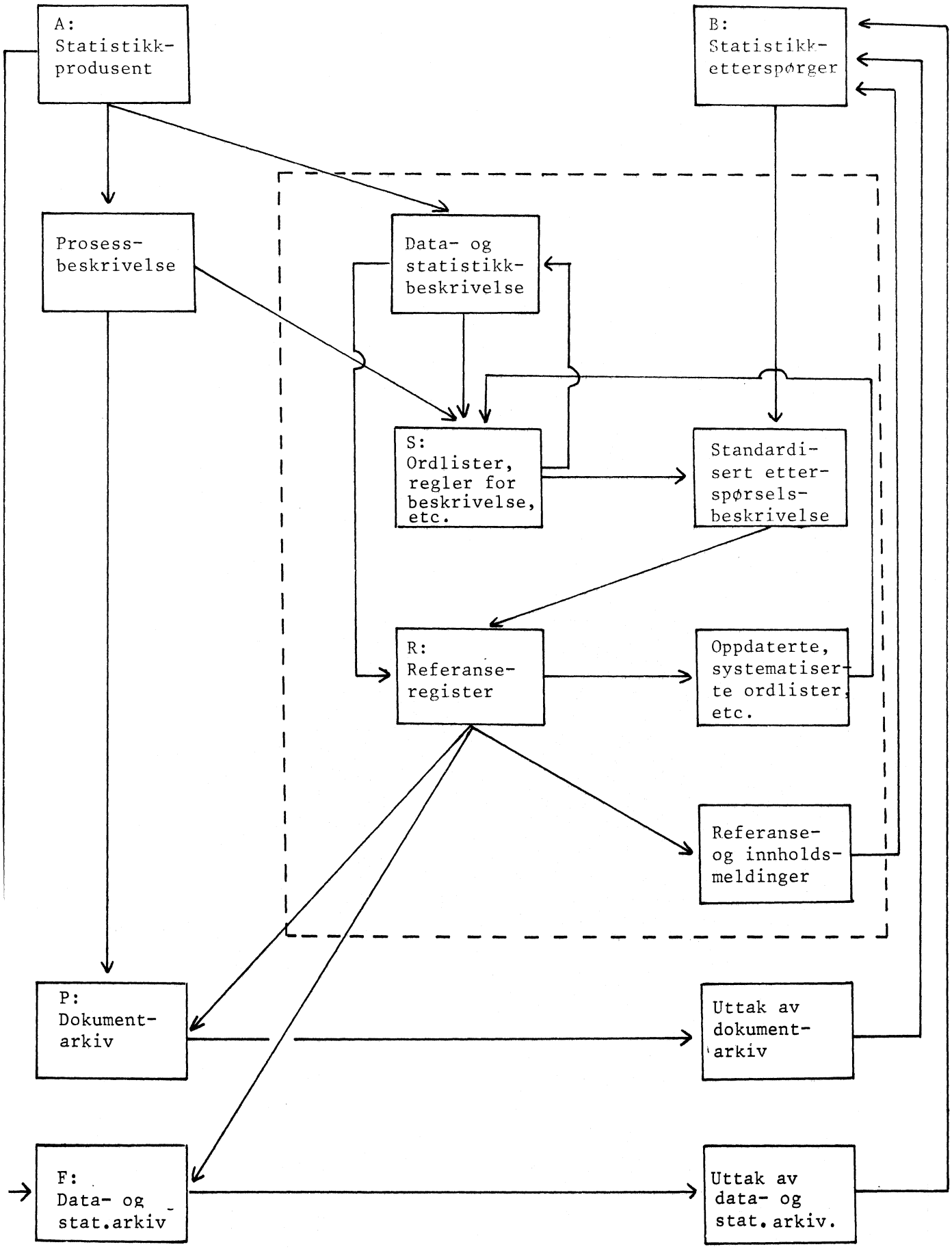
Jeg vil tro det vil være mest hensiktsmessig å legge prosjektet under Produksjonsavdelingen.



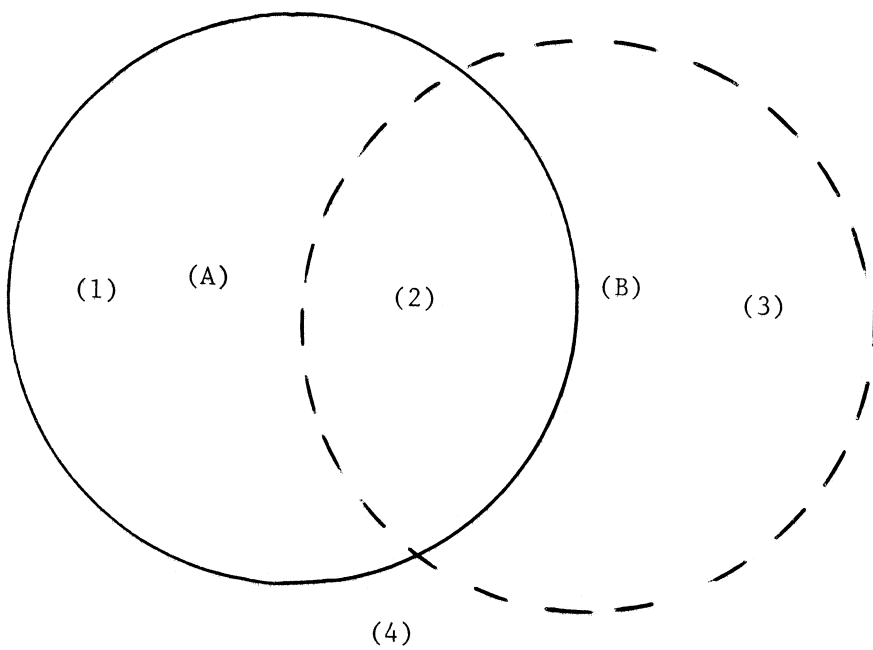
Figur 1: Faser i løsning av problemer



Figur 2: Søking etter statistisk informasjon



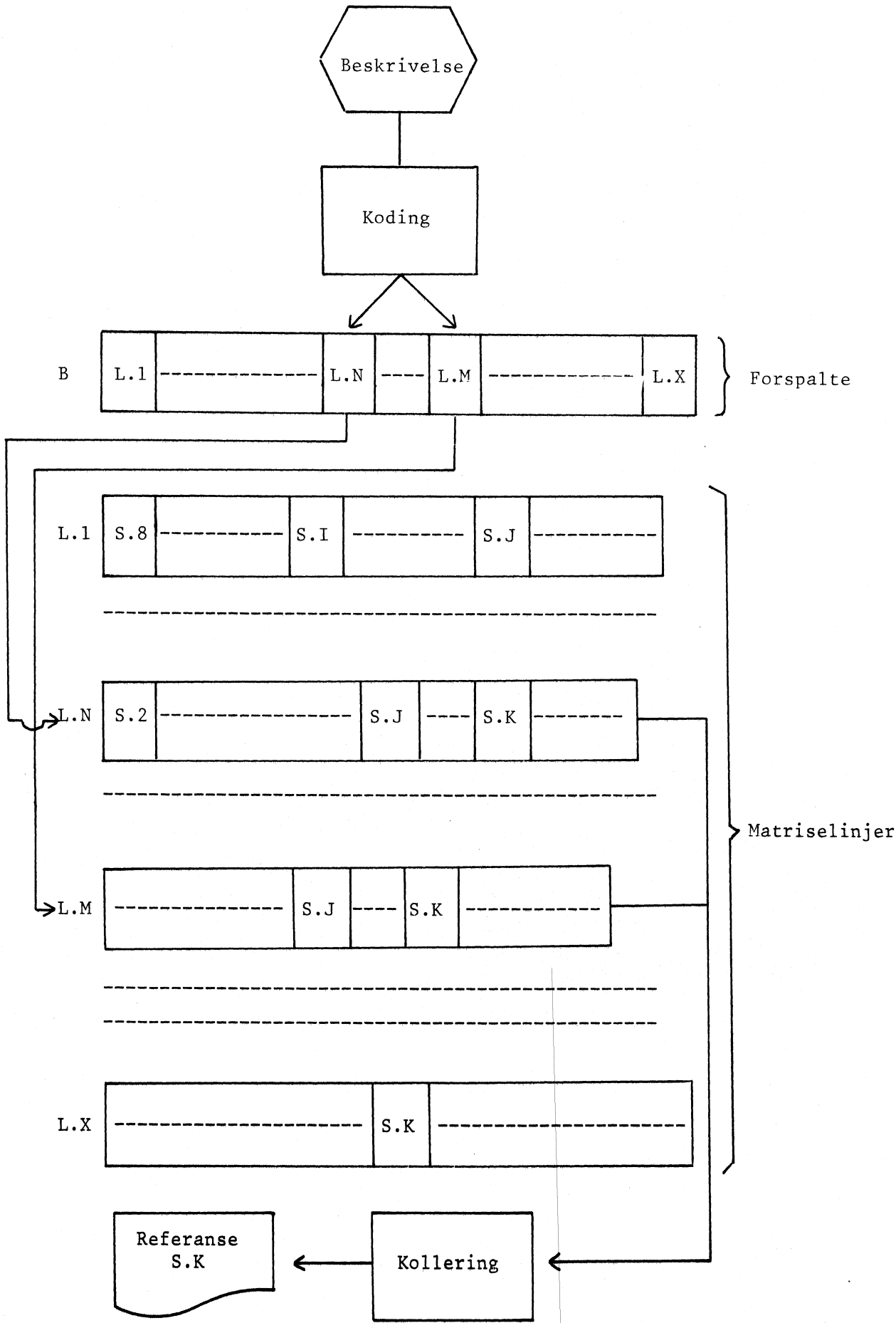
Figur 3: Informasjonssystem for framhenting av data og statistikk



Figur 4: Lagring og framhenting av informasjon

	Beskrivelser						
	Folketelling	Kommuneregnskaper	Arbeidskraftund.	Nasjonalregnskap	Industristatistikk	Data for konsumprisindeks	Pers.reg.
OBJEKTTYPE:							
Personer	x		x				
Familier	x						
Bedrifter					x	x	
Foretak(er)					x		
Kommuner		x					
Transaksjonsaggregater				x			
Bedriftsaggregater					x		
Foretaksaggregater					x		
NAVN:							
Befolkn. i Norge	x						x
Norges kommuner		x					
Utvalg arbeidsaktive			x				
Store bedrifter					x		
Utvalg av detaljomsetningsforr.						x	
Nasj.regn. sektorer				x	x		
Næringsklassifikasjon					x	x	
Kons.vare klassifikasjon						x	
KJENNEMERKE:							
Alder	x		x				
Kjønn	x		x				
Næring	x		x		x		
Utgift		x		x			
Inntekt		x		x			
Aktivitetsstatus	x		x				
Bruttoproduksjonsverdi				x	x		
Gj.pris				x		x	
Pris						x	
TIDSSPESIFIKASJON:							
1.11.1960	x						
1.11.1970	x						
1969		x		x	x		
1970		x		x	x		
4. kv. 1971			x				
1. kv. 1972			x				
Jan. 1972						x	
Feb. 1972						x	

Figur 5: Begrep x beskrivelse matrise



Figur 6: Lagerorganisasjon i et begrep x beskrivelser system

JMH/EH/ES, 16/6-72

VARIABELKATALOGEN

av Jan M. Hoem og Eivind Hoffmann

1. Framdriften og gjennomføringen av SSDS-prosjektet vil etterhvert være sterkt avhengig av variabelkatalogen og dens etablering - særlig når det gjelder eksperimentering med å få fram tabell-former og -systemer på grunnlag av et SSDS som kan tenkes å gi noe mer enn det personstatistikken gir i dag. Hoffmann har under arbeid et notat om SSDS-arbeidets nåværende situasjon og en mulig framdriftsplan for det fortsatte arbeidet. I det notatet vil samspillet med variabelkatalogen bli berørt, men noen inngående behandling av variabelkatalogprosjektet vil ikke bli gitt der. Når notatet foreligger, vil vi imidlertid be om å få et møte med Byråets ledelse for å diskutere det videre SSDS-arbeidet og det gjensidige forhold mellom dette og variabelkatalog-prosjektet. De følgende merknader er dels reaksjoner på Nordbottens notat SN/WA, 31/5-72 og dels en del av forarbeidet til et slikt møte.

2. Vi forestiller oss at en sentral del av en variabelkatalog må bestå av et fullstendig sett med variabelbeskrivelser. Rundt dette settet organiseres det så et referansesystem som gjør det mulig for brukeren å finne fram til de variabelbeskrivelsene han vil ha tak i

(a) uten å gå glipp av relevante beskrivelser, og

(b) uten å måtte sortere vekk for mange irrelevante beskrivelser, slik Nordbotten angir i avsnitt 3.2.

Det blir da viktig å bestemme seg for hva som skal gå inn i en variabelbeskrivelse. Vi skjønner ikke alt Nordbotten skriver om dette i sitt notat, og vi savner en innholdsliste for en variabelbeskrivelse, framstilt i klartekst. For å gjøre dette klarere for oss selv, har vi derfor utarbeidet nedenstående diskusjon.

3. Vi minner da først om at en variabel defineres som en tilordning som til ethvert element i en definisjonsmengde tilordner akkurat én verdi i en verdimengde. (Ordet "kjennemerke" er synonymt med "variabel", som er det samme som det matematiske begrepet "funksjon".) I en variabelbeskrivelse må det derfor inngå tre grunnelementer, nemlig definisjonsmengde, verdimengde, og tilordningssystem.

Anvendt på den konkrete situasjon i Byrået, betyr dette at en variabelbeskrivelse minst må inneholde følgende:

(i) Definisjonsområde. Våre variable vil typisk være definert over objekter (institusjoner; personer; grupper av personer, som familier og husholdninger; eiendommer; gjenstander) og hendelser (handlinger, transaksjoner, begivenheter, osv.).

(ii) Verdimengde. Dette kan være en standard (utdanningsstandard, yrkesstandard, el.l.), en kodeliste, eller noe annet.

(iii) Tilordningssystem. Dette kan være en beregningsprosedyre, en algoritme, en formel, eller noe annet. I noen tilfeller vil tilordningssystemet følge umiddelbart av pkt. (i) og (ii) ovenfor, men det vil ofte være bruk for utfyllende forklaringer.

(iv) Opprettelses- og opphørstidspunkt. Verdiene av Byråets variable framkommer gjennom datainnsamling og -bearbeidelse. Med jevne mellomrom endres innsamlings- eller bearbeidelsesrutinene. I prinsippet opprettes det en ny variabel ved en slik endring, og en tidligere variabel vil kanskje opphøre.

For å gjøre det mulig å følge en problemstilling over tid, bør en variabelbeskrivelse inneholde

(v) henvisning til eventuelle forgjengere og etterfølgere til en variabel. (Sml. pkt. (iv) ovenfor.)

For å knytte sammen statistikkområder, bør beskrivelsen også inneholde

(vi) henvisning til andre beslektede variable.

En del av disse formålene, og andre formål, kan ivaretas gjennom en

(vii) dokumentliste. Her inngår henvisning til prosjektskisser, prosjektbeskrivelser, planleggingsnotater, skjemaer/blanketter, kodelister, revisjonsinstrukser, programhenvisninger, klassifikasjonsbeskrivelser, tabellspesifikasjoner, skriftlige kvalitetsvurderinger osv. Likeledes bør man referere til kjente, sentrale anvendelser.

(viii) Tilgjengelighet og arkiveringsform for variabelverdiene bør også tas med i beskrivelsen. Foreligger de som publiserte tabeller, på maskinmedia, eller i upubliserte tabeller? Hva er arkiveringssted,

hvor er eventuell filebeskrivelse, hvem skal kontaktes for nærmere opplysninger, er det noen konfidensialitetsbeskrankninger, osv.

4. Med utgangspunkt i det ovenstående vil vi gjerne komme med følgende betraktninger over Nordbottens notat.

5. Hvis vi har forstått notatet riktig, vil et fullt sett av variabelbeskrivelser foreligge når Nordbottens tre første faser er gjennomført. Det ser ut for at hans opplegg innebærer at det i prinsippet ikke vil foreligge et fullstendig sett av opplysninger innenfor noe enkelt statistikkområde før katalogen er ferdig etablert. Man bør overveie om det ikke vil være mer hensiktsmessig, både utfra fagkontorenes behov og medvirkning og ut fra ønsket om å prøve katalogen i praksis, først å etablere mest mulig fullstendig informasjonssett for variablene innenfor ett eller to avgrensede statistikkområder, og så utvide katalogen "i bredden" etter hvert - framfor å bygge "i høyden" slik Nordbottens forslag synes å innebære. (I parentes kan bemerkes at katalogdelene I til III i Hoffmanns notat EH/IH, 8/2-72 gjelder organiseringen av katalogen mer enn gangen i arbeidet med å etablere den.)

6. Nordbotten omtaler ikke hendelser. En hendelse kan naturligvis oppfattes som en endring i verdien til en variabel definert over en samling objekter, og den vil derfor i prinsippet kunne gjenfinnes i systemet. (Eksempel: antall levendefødte barn i Norge i 1972 til x-årige mødre med paritet 0.) I noen tilfeller kan det vel allikevel være fornuftig å ta med i katalogen variable definert over hendelser. (Eksempler: kjønnsproporsjonen blant fødslene, andel av tvillinger blant de levendefødte, ulykker (branner) med personskader og materielle skader.)

7. Er tanken at variabelkatalogarbeidet skal starte f.eks. med NOS-årgangen 1970 (og siden med maskintilgjengelige data og statistikk med registreringsdato i 1970), og senere arbeide seg utetter gjennom 1971 og 1972, med ajourføring etter hvert som tiden går? Hva da med eldre materialer? (Sml. Nordbotten, s. 13, oppg. 1.1.)

8. Skal hånd- og maskinskrevne tabeller helt utelates fra systemet? Eller kommer de inn under oppgave 3.1? (De nevnes ikke der.)

9. For oss ser det ut som om tanken er at fase 2 skal gi en katalogdel der første "oppslagsord" er en dato (f.eks. 1/11 1970) eller en periode (f.eks. 1970). Man skal så få en liste over data og statistikk som er tilgjengelige for datoen eller perioden. (Sml. Nordbotten, s. 13-14.)

Hvis dette ikke kan framkomme som et enkelt biprodukt av fase 1, f.eks. ved en grei omsortering av recordene, vil vi synes at andre deler av prosjektet, som nå er stilt opp i senere faser, eller som det hentydes til under punktene ovenfor, bør ha høyere prioritet.

10. Vi merker oss at den nødvendige personellmengde som angis for utviklingen av variabelkatalogen, er øket til to, formodentlig to personer i fullt arbeid. Når kan dette arbeidet komme i gang? Hvor stor belastning kan det tenkes å få for Sosiodemografisk forskningsgruppe p.g.a. samspillet med SSDS-prosjektet? Hvor store personellressurser trengs for den årlige ajourføringen når variabelkatalogen er opprettet?

Vi er enige i at det må være mest hensiktsmessig å legge prosjektet under produksjonsavdelingen og at det forøvrig legges opp administrativt som skissert av Nordbotten på side 5.

EH/KS, 29/6-72

REFERAT FRA MØTE OM VARIABELKATALOGPROSJEKTET 26/6-72

ved Eivind Hoffmann

1. Formålet med møtet var å avklare uklare punkter og eventuelle uoverensstemmelser i Nordbottens og Hoem/Hoffmanns notater om variabelkatalogen, [1] og [2], med særlig vekt på spørsmål vedrørende innhold, omfang og arbeidets gang.

2. Nordbotten innledet med å kort redegjøre for prosjektets forhistorie, som går tilbake til 1965, og for oppbygningen og innholdet av sitt notat. Dette notatet er både omfattende og mer teknisk i sitt innhold enn Hoem/Hoffmanns. Det siste behandler de elementene av figur 3 i Nordbottens notat som er betegnet "Data- og statistikkbeskrivelse", "Prosessbeskrivelse" og "Dokumentarkiv", og det var enighet om at det var disse delene av systemet som i første rekke skulle behandles på møtet, da resten av systemet på figur 3 har mer "teknisk" interesse.

3. Man var enige om at en variabelkatalogs innhold dekkes av punktene (i) - (viii) på s. 2-3 i Hoem/Hoffmanns notat, og at det bare var terminologiske uoverensstemmelser med Nordbottens notat på dette punkt. I første linje i punkt (vii) skal det dog stå: "(vii) dokumentliste. Her inngår henvisning til prosjektbeskrivelser, ..."

4. Det var enighet om at variabelkatalogen når den er etablert må omfatte all statistikk utarbeidet av Byrået fra og med en bestemt årgang - f.eks. 1970. I særlige tilfeller bør man ta sikte på å få med også eldre årganger, men generelt sett er dette en vanskelig oppgave som ikke kan gis høy prioritet.

5. De to notatene skisserer to prinsippielt forskjellige framgangsmåter for å etablere variabelkatalogen. Nordbotten foreslo en serie faser i arbeidet der man i hver fase skal dekke hele Byråets statistikkområde. Hoem/Hoffmann konstaterer at variabelkatalogen ikke vil være operasjonell før de tre første fasene er fullført, og foreslår at man istedet velger ut et par statistikkområder og starter med å lage en fullstendig katalog for disse. Det var enighet om at denne framgangsmåten ville være å foretrekke både utfra mulighetene til å få utprøvet

hele variabelkatalog-systemet, og utfra hensynet til samarbeidet med fagkontorene, som særlig i den egentlig dokumentasjonsfasen, Nordbottens fase 3, må være meget nært. For i størst mulig grad å unngå sektorspesifikke løsninger vil det være en fordel om arbeidet starter med ett økonomisk og ett individstatistisk område.

6. Det var enighet om den organisering av arbeidet som er skissert i Nordbottens notat (s. 15) og om at det bør ligge under Produksjonsavdelingen. Hvis det imidlertid viser seg å være vanskelig på kort sikt å finne begge de stillingene som Nordbotten foreslår, ved avdelingen, er Hoem innstilt på å la en av de framtidige stillingene i Sosiodemografisk forskningsgruppe bli overført til dette arbeidet.

7. Når variabelkatalogen er etablert, vil det vesentlige av ajourføringsarbeidet ligge hos fagkontorene i samarbeid med system- og driftskontor. Det vil inngå som et integrert ledd i arbeidet med de ulike statistikkområdene. Variabelkatalogen som sådan vil derfor neppe kreve store ressurser. I en overgangsperiode etter etableringen av systemet må vi allikevel vente at fagkontorene og andre vil trenge en del rettledning fra dem som har utarbeidet systemet.

8. Det ble understreket at når etableringen av en variabelkatalog må forventes å bli et forholdsvis stort prosjekt, så skyldes det i stor grad at dokumentasjonsarbeidet synes å ha vært lavt prioritert ved fagkontorene. Det vil i første rekke være dem som vil ha fordel av systemet og arbeidet med å etablere det, f.eks. i form av lettere overføring av oppgaver til nye funksjonærer og bedre muligheter for å besvare forespørsler - som det antakelig vil kunne bli færre av.

9. Selv om det i første omgang vil være Byrået selv som har glede av variabelkatalogen, vil den også bidra til å forbedre Byråets service overfor publikum. Sammen med Biblioteket og Informasjonstjenesten vil Variabelkatalogens "ytre" del sterkt kunne øke Byråets informasjonstilbud, f.eks. ved tilbud om et meldingssystem for statistiske data tilsvarende slike som er etablert for tidsskriftartikler på mange fagområder.

10. De fleste spørsmål som er reist i Hoem/Hoffmanns notat er besvart ovenfor, men det kan tilføyes at Nordbottens notat egentlig også tar i

betraktning hendelser og at det er meningen at hånd- og maskinskrevne tabeller skal med i systemet.

11. Det var enighet om at med prosjektets viktighet, dets lange forhistorie og det foreliggende materiale, bør en anbefale Byråets ledelse snarest å ta opp til vurdering om hvorvidt og når prosjektet kan settes igang.

Henvisninger:

[1] Nordbotten, Svein: Prosjektet variabelkatalog og et referansesystem til arkivert numerisk informasjon. Notat. SN/WA, 31/5-72.

[2] Hoem, Jan M. og Eivind Hoffmann: Variabelkatalogen. Notat. JMH/EH/ES, 16/6-72.

DOKUMENTASJON AV DATA I STATISTISK SENTRALBYRÅ

av Erik Aurbakken

1. INNLEDNING

Det har gjennom de senere år til en viss grad blitt reist kritikk mot den beskrivelse av data som eksisterer i Byrået og uttrykt bekymring om den framtidige situasjon etter hvert som dataarkivene vokser og utnyttelsesgraden øker. Problemene er foreslått løst ved innføring av en "variabelkatalog" som er ment å være et oppslagsverk eller nøkkel til alle data, både individualdata og aggregerte data. Søkingen i dette oppslagsverket kan tenkes mer og mindre automatisert. En automatisering forutsetter innføring av standardiserte språk for data- og produktbeskrivelse som må være kjent av brukeren. Professor Svein Nordbotten har i et notat 31/5-72 "Prosjektet variabelkatalog og et referansesystem til arkivert numerisk informasjon" skissert et relativt avansert opplegg for en slik variabelkatalog.

På den annen side bør det reises spørsmål om den dokumentasjon vi har i dag blir utnyttet i den grad det er mulig, om den når fram til og blir forstått av de brukergrupper som har behov for å finne fram i data.

Formålet med dette notatet er å gi en oversikt over dokumentasjonen av data i Byrået i dag, peke på svake sider og mulige tiltak på kort sikt for å utbedre disse.

2. KRAV TIL DATADOKUMENTASJONEN

Det generelle krav til datadokumentasjonen synes å være at den skal gi en total oversikt over alle tilgjengelige data med muligheter for lett søking etter klare og entydige beskrivelser av de data vi er interessert i til enhver tid. Systemet må samtidig gi den nødvendige beskyttelse mot illegal bruk av data.

Mer spesielt kan vi stille disse krav:

- Dokumentasjonen skal være rettet mot både bruker, statistikkprodusent og programmerer og lette kommunikasjonen mellom disse.
- Dokumentasjonen skal være mest mulig ajour til enhver tid.
- Dokumentasjon bør framkomme som nødvendige produkter under planlegging og bearbeiding av statistikken. Dette antas å sikre en mest mulig feilfri dokumentasjon.

- Dokumentasjonsteknikken bør være godt beskrevet og enkel å lære.
- Dokumentasjonsteknikken bør være mest mulig ensartet for alle data. Dette betyr at vi må være forsiktige med å foreta unødige endringer i teknikken. (Dokumentasjonssystemet kan utvides uten at teknikken endres.)

3. SITUASJONEN I DAG, MED FORSLAG TIL ENDRINGER

3.1. Maskinlesbare data

De fleste maskinlesbare data er lagret på magnetbånd. Dokumentasjonsteknikken for disse data har vært noe endret siden de første data ble lagret på bånd i 1962. De siste 5 årene har imidlertid teknikken vært den samme og praktisk talt alle data i arkivet er beskrevet ved hjelp av den teknikk som brukes i dag. Teknikken kan også brukes for data på hullkort, men dette har i dag liten betydning.

Den største dataenheten i dokumentasjonssystemet er file. Hver file blir tildelt et identifikasjonsnr. ved Systemkontoret. Dette nummeret er satt sammen av 3 ledd. Første ledd identifiserer det vi kan kalle en hovedfile. Fra en slik hovedfile kan vi danne flere underfiler ved å ekstrahere et utvalg av kjennemerker for alle enheter i filen. Slike underfiler tildeles nummer ved å variere annet ledd i nummeret. Hver slik underfile kan brytes videre ned ved å ekstrahere et utvalg av enheter. Disse filene tildeles nummer ved å variere tredje ledd i nummeret. Ut fra identifikasjonsnummeret er det altså mulig å se hvordan de enkelte filer er ekstrahert fra filer av høyere nivå.

Dessuten er det for hver file registrert hvilke filer som er brukt når filen er etablert. Dette kan betraktes som en referanse til filens "foreldre". Det tidsintervall eller tidspunkt dataene refererer til er registrert som periode eller generasjon i et fastlagt kodesystem. I tillegg til dette er det også mulig å registrere hver file i flere versjoner. Dette blir bare brukt når det på grunn av feil i bearbeidingsprosessen er nødvendig å framstille samme file flere ganger.

Hver file er beskrevet ved navn, enhet, omfang og en liste over de kjennemerker den inneholder. Hvert kjennemerke har et navn og en teknisk beskrivelse som er nødvendig for programmeringen.

For hver file er det registrert de nødvendige tekniske data for Driftskontoret, hvilken rull magnetbånd de er lagret på, når filen ble etablert, når den er tenkt slettet i arkivet, og hvem som er ansvarlig for slettingen.

Dokumentasjonen av dataene er delt i følgende deler:

- Fileoversikt

- Filebeskrivelse for produksjon (vedlegg 1)
- Filebeskrivelse for programmering (vedlegg 2)

Disse delene skal tjene forskjellige formål og oppdelingen er også av sikkerhetsmessig betydning.

3.1.1. Fileoversikt

Driftskontoret produserer maskinelt en liste som viser alle filer og hvilken periode (årgang, kvartal osv.) av filen som nå står i arkivet. Lista er ordnet etter statistikknummer og filens identifikasjonsnummer. Lista blir brukt til å vise hvilke data som finnes i arkivet og til markering av hvilke data som skal slettes.

Ansvaret for sletting av filene fastlegges av Systemkontoret etter følgende retningslinjer:

- Individualdata: Byråets ledelse.
- Kopier og ekstrakter av individualdata og aggregerte data i rutiner som er under innkjøring: Systemkontoret.
- Kopier og ekstrakter av individualdata og aggregerte data i innkjørte rutiner: Driftskontoret.
- Filer som er etablert i forbindelse med spesialoppdrag: Fagkontoret.

Det kan pekes på følgende svakheter ved fileoversikten i dag:

- Den inneholder en del filer som har gått ut av arkivet og aldri vil bli etablert på nytt. Disse kan selvfølgelig fjernes, men dette må gjøres av folk som har god kjennskap til statistikkområdet, helst fagkontoret.
- Oversikten er tung å lese, særlig fordi navnet på filene ofte er lite opplysende for en som ikke kjenner rutinen. Dette bør kunne rettes på ved at fagkontoret kommer sterkere inn i bildet ved navnsettingen.
- Oppdelingen på identifikasjonsnummer er ikke alltid utført etter reglene. Dette skyldes dels at reglene for nummerering kanskje har vært noe uklare når det gjelder kobling av data fra forskjellige filer til en ny file, dels at en rutine iblant bygges opp av uavhengige delrutiner slik at en ikke tidsnok har oversikt over hva som egentlig skal være hovedfile i en rutine. Dette kan bare rettes på ved å legge mer arbeid i fileidentifiseringen under systemarbeidet.

Det bør arbeides videre med fileoversikten med tanke på å gjøre den mer lesbar. Den opprinnelige tanken var at fagkontorene skulle bruke oversikten til å følge med i og planlegge sin del av dataarkivet. Hittil har oversikten i liten grad blitt presentert for fagkontorene. Etter hvert som det blir nødvendig å reise spørsmålet om sletting av individualdata, må imidlertid oversikten legges fram for fagkontorene og ledelsen.

3.1.2. Filebeskrivelser for produksjon

Dette er blanketter som føres manuelt ved Driftskontoret med utgangspunkt i en blankett som fylles ut av Systemkontoret når rutinen planlegges (se 3.1.3.). Denne dokumentasjonen inneholder alle data fra fileoversikten, dessuten signatur for sletting av data.

Denne dokumentasjonen brukes til å rekvirere data fra arkivet og fylle ut kontrollkort for kjøring på datamaskinen. I kontrollkortet blir det satt inn en del data som maskinen kontrollerer mot tilsvarende data registrert på magnetbåndet. Dette sikrer oss at riktig data blir kjørt på riktig program.

Blanketten gir også plass for utfylling av filens omfang og kjennemerker, men dette skal normalt av sikkerhetsmessige grunner ikke fylles ut. De første årene systemet var i bruk ble også disse data fylt ut.

Det har også her vært et problem å få plukket ut blanketter som etter hvert har blitt foreldet.

Ellers later det ikke til å være noen nevneverdige svakheter ved denne delen av dokumentasjonen når en ser den i relasjon til det formål den skal dekke.

3.1.3. Filebeskrivelser for programmering

Dette er blanketter som føres manuelt ved Systemkontoret og inneholder alle data om filen og hvert enkelt kjennemerke, men viser ikke hvilken periode eller generasjon av filen som er arkivert. Det er derfor ikke mulig fra denne dokumentasjonen å hente fram data fra båndarkivet uten å gå om fileoversikten eller filebeskrivelser for produksjon. Dette forhold er av sikkerhetsmessig betydning.

Denne dokumentasjon er ordnet etter statistikknummer og fileidentifikasjonsnummer slik at det er lett å finne fram til dokumentasjonen av de enkelte kjennemerkene når en har funnet en file i fileoversikten.

Denne dokumentasjonen blir først og fremst brukt under programmeringen. Kopier kan leveres ut til fagkontorene på forespørsel. Dette har hittil skjedd i liten utstrekning.

Denne delen av datadokumentasjonen synes også å fungere tilfredsstillende sett i relasjon til det formål den er ment å dekke.

Opprinnelig var det tenkt at beskrivelsen av det enkelte kjennemerke fra filebeskrivelsen skulle punches for å kunne framstille en kjennemerkeliste eller variabelkatalog. Når dette ikke har blitt gjort, skyldes det følgende forhold:

- Kjennemerkebeskrivelsen kan ikke betraktes som noen definisjon av kjennemerket.

- Navnene er ikke tilstrekkelig beskrivende for en som ikke kjenner rutinen.
- Mange av kjennemerkene er å betrakte som uinteressante mellomresultater.

Det er meget tvilsomt om det er mulig eller ønskelig å gjøre denne kjennemerkebeskrivelsen til en fullstendig definisjon av kjennemerket. Det formatet en har til rådighet for beskrivelsen vil f.eks. ikke strekke til for en slik definisjon.

I kjennemerkebeskrivelsen er det satt av plass til en referanse til kodeliste for kvalitative kjennemerker. Dette forutsetter en registrering og tildeling av identifikasjonsnr. for alle kodelister, noe som foreløpig ikke er gjennomført i særlig utstrekning.

En svakhet som synes å hefte ved denne delen av dokumentasjonen dersom den skal få en videre anvendelse, synes å være at det mangler en god referanse fra kjennemerkebeskrivelsen til en definisjon av kjennemerket. Det bør her kunne etableres en slik referanse til skjemanr. eventuelt også postnr. innen skjemaet, dersom vi velger å bruke spørreskjemaet som en definisjon av kjennemerket. Avledede kjennemerker må i så fall behandles på en annen måte. I dag er disse definert gjennom programbeskrivelser som er mindre systematisk oppstilt og vanskeligere tilgjengelig, men det er alltid en referanse fra filebeskrivelsen til programbeskrivelsen.

3.2. Skjema

De fleste individualdata hentes inn på spørreskjema. Disse dataene må sies å være dokumentert ved teksten på spørreskjemaet og rettledningen for utfylling av skjemaet.

Alle spørreskjema blir tildelt et nummer ved Trykningskontoret. Dette nummeret blir ikke byttet ut selv om det blir foretatt endringer på nye utgaver av skjemaet. Nummer som blir ledige ved at skjema går ut av bruk, blir tatt i bruk til nye skjema. Et skjema er derfor identifisert ved hjelp av nummer og periode (årgang).

Trykningskontoret holder et arkiv over alle skjema som er i bruk i Byrået. Arkivet er ordnet etter skjemanummer. For hvert skjema er det dessuten et kartotekkort med enkelte tekniske data. Disse kortene er ordnet etter kontor og skjemanavn.

Skjemaarkivet har vært lite brukt utenfor Trykningskontoret og er heller ikke organisert med tanke på oppslag og søking etter informasjon.

Arkivet kan muligens automatiseres og gjøres lettere tilgjengelig ved å overføre skjemaene til mikrofilm og legge opp en indeks for søking.

3.3. Tabeller

En tabell kan beskrives ved navn (som inneholder enhet og omfang), forspalte og hode.

Det eksisterer ingen sentral oversikt over alle tabeller i Byrået. Tabeller som er eller har vært arkivert på magnetbånd er imidlertid dokumentert som maskinlesbare data. Det er ingenting i vegen for å dokumentere alle tabeller på denne måten, også tabeller som lages manuelt, men det er tvilsomt om dette gir en lett lesbar dokumentasjon av tabellene.

Det er mulig en liste over alle tabellnavn ordnet etter statistikknr. med referanse til publikum hvor tabellen er trykt (eller fagkontor hvor den foreligger utrykt) og den filen som er brukt ved framstilling av tabellen kan dekke et behov hos statistikkbrukere. En slik liste trenger neppe inneholde absolutt alle tabeller som blir laget. Ved en intervjuundersøkelse kan det f.eks. bli framstilt flere hundre tabeller og en opplisting av alle disse vil trolig ha liten hensikt. Lista kan også inneholde tabeller som ennå ikke er ferdige, med planlagt ferdigdato, slik at det er mulig å ta ut ekstrakter over statistikk som vil komme f.eks. i neste halvår.

3.4. Publikasjoner

Vegviser i norsk statistikk gir en oversikt over statistikkområdene og hvilke publikasjoner som presenterer statistikkproduktene. Vegviseren refererer bare til publikasjonene. Disse refererer igjen til en viss grad til utrykte tabeller.

Det bør være mulig å bygge ut vegviseren med noe flere detaljer og gi den bedre ajourhold uten at dette betyr noe omlegging av systemet eller innføring av ny teknikk. Tilsvarende er det mulig å referere til flere utrykte tabeller i publikasjonen med pris for kopier. Dette trenger ikke nødvendigvis bare være tabeller som er kjørt ut, men kan også inkludere tabeller som er lette å produsere på standardprogram.

3.5. Kodelister

Kodelister er en del av datadokumentasjonen. For en del kjennemerker finnes det vedtatte standarder som blir brukt på tvers av statistikkområder. For de fleste kjennemerker er det ingen slike standarder. Byrået har ikke noe sentralt organ for registrering, identifisering og ajourhold av kodelister.

Systemkontoret startet for noen år siden en registrering og identifisering av kodelister og har hittil registrert ca. 20 lister.

Vår arkivering av data over tiden forutsetter at vi også tar vare på kodelistene for hver enkelt periode av de arkiverte data. Skal vi få en enkel referanse mellom kjennemerkebeskrivelsen i filebeskrivelsen og kodelistene, må vi ha en registrering av listene med tildeling av identifikasjonsnr. for hver liste. Svært korte kodelister er det neppe behov for å registrere. I dag blir disse gjengitt i sin helhet i kjennemerkebeskrivelsen i filebeskrivelse for programmering. En slik registrering av kodelister er allerede startet ved Systemkontoret, og det synes naturlig å bygge denne videre ut. Det vil da også være naturlig at Systemkontoret får til oppgave å arkivere et eksemplar av kodelistene med alle endringer som blir foretatt.

4. OPPSUMMERING AV FORSLAG

De tiltak som er omtalt under avsnitt 3 vil her bli konkretisert i forslag listet under det kontor hvor arbeidet synes å høre heime. Det er her først og fremst snakk om tiltak på kort sikt.

4.1. Fagkontorene

Fagkontorene skal:

- godkjenne Systemkontorets navnsetting av filer og kjennemerker,
- gå igjennom fileoversikter og merke data som skal slettes og data som aldri vil komme igjen i arkivet, i samarbeid med Systemkontoret foreslå utvidet lagring av individualdata og aggregerte data,
- rapportere nye kodelister og endringer i gamle til Systemkontoret (dersom det blir vedtatt at Systemkontoret skal arkivere slike lister),
- i større grad enn hittil referere til utrykte tabeller i publikasjonene.

4.2. Systemkontoret

Systemkontoret skal:

- legge fram forslag til et referansesystem mellom filebeskrivelsene for programmering og skjemaene, både på nivå file/skjema og kjennemerke/skjemapost,
- legge fram forslag til registrering, identifisering og arkivering av kodelister,
- samarbeide med Driftskontoret om utbedring av fileoversikten, gjøre den kortere og mer lesbar,

- samarbeide med fagkontorene om planer for utbygging av dataarkivet og fagkontorenes bruk av fileoversiktene,
- vurdere behovet for videre utvikling av en variabelkatalog, spesielt behovet for rask søking etter dokumentasjon.

4.3. Driftskontoret

Driftskontoret skal:

- utbedre fileoversikten i samarbeid med Systemkontoret,
- vurdere om referanse til båndarkivet kan sløyfes på de fileoversikter som leveres ut fra kontoret,

4.4. Trykningskontoret

Trykningskontoret skal:

- lage forslag om identifisering av hver post på skjemaet,
- lage forslag om en mer standardisert form for navnsetting av skjemaene,
- vurdere metoder for å gjøre skjemaarkivet lettere tilgjengelig, f.eks. ved bruk av mikrofilm.

4.5. Informasjonskontoret

Informasjonskontoret skal:

- vurdere en utbygging og muligheten for et kontinuerlig ajourhold av "vegviseren" og behovet for en liste over alle viktige tabeller,
- peke på svakheter i det systemet vi har i dag når det gjelder å finne fram til statistikk som etterspørres utenfra.

5. ORGANISERING AV ARBEIDET

Det arbeidet som er foreslått under avsnitt 4 bør koordineres av en styringsgruppe med en representant fra et fagkontor og en fra hvert av de øvrige kontorene som er engasjert. Byråets ledelse bør peke ut representanten fra fagkontorene og formannen i gruppen.

6. TIDSPLAN, KOSTNADER

Arbeidet bør settes i gang snarest og drives fram med det tempo ressursene tillater. Ved neste rullering av Byråets arbeidsprogram må styringsgruppen sørge for at prosjektet blir ført opp på programmet og i tidsplanen under de kontorer som deltar.

Statistisk Sentralbyrå

Dato: _____ Ark.nr. _____

Filebeskrivelse

Original ☐ eller kopi ☐ Type D

Original arkivert: _____

Oppdrag:

Stat.nr./Oppdragsgiver:

Filenavn:

Fileomfang:

Prosjekt nr.: Hovednummer:

Enhet	Rec.t	Felt nr.	Pos./Ord Fra	Pos./Ord Til	Databeskrivelse	Feltbeskrivelse	Kodeliste Nr.	Kodeliste Kode

Type S(ystem) | **Type P(rogrammering)**

1	2	3	4	5	6		8	9	10	11	12	13	14
Utgave-nr.	Opprinnelse-ident.	Dia-gram-ref.	Navn/Omfang (Omfang settes i parentes)	Sortering (pos.x pos.x .....)	Output type	Sort. beskr.	Arkiv-opplysn.	Progr.	Maskin-type	Label	Blokk-faktor	Bånd-tetth.	Bånd-kval.

Type D(rift) | Side .....

15	16		17	18	19	20	21	22	23	24	25	26	27	28	29
Linje nr.	Opprinnelse			Bånd-nr.	File ser.no.	Vol. seq. no.	Antall ruller	File seq. no.	Gen. no.	Version no.	Creat. date	Exp.date	Godkjent blanking		Dato blanket.
	Ident.		Linje nr.								år, dag nr.	år, dag nr.	Dato	Sign.	

Rubrikk: 6. M = Magnetbånd, K = Kort, L = Liste
8. D = Driftskontor, S = Systemkontor, F = Fagkontor, L = Ledelsen
11. S = Scope, B = Bios, D = Dos
13. 5 = 556 bpi, 8 = 800 bpi, 16 = 1 600 bpi
14. B = Beste kvalitet, S = Standard kvalitet

A 30-1

Statistisk Sentralbyrå

Dato: _____ Ark.nr. _____

Filebeskrivelse

Original ☐ eller kopi ☐ | Type S/P

Original arkivert: _____

Oppdrag:

Stat.nr./Oppdragsgiver:

Filenavn:

Fileomfang:

Prosjekt nr.: | Hovednummer:

	Enhet	Rec.t	Felt nr.	Pos./Ord Fra	Til	Databeskrivelse	Feltbeskrivelse	Kodeliste Nr.	Kode

Type S(ystem)								Type P(rogrammering)					
1	2	3	4	5	6	7	8	9	10	11	12	13	14
Utgave nr.	Opprinnelse-ident.	Dia-gram-ref.	Navn/Omfang (Omfang settes i parentes)	Sortering (pos...x pos...x ...)	Output-type	Gen. beskr.	Arkiv-opplysn.	Progr.	Maskin-type	Label	Blokk-faktor	Bånd-tetth.	Bånd-kval.

Rubrikk: 6. **M** = Magnetbånd, K = Kort, L = Liste
8. **D** = Driftskontor, S = Systemkontor, F = Fagkontor, L = Ledelsen
11. **S** = Scope, **B** = Bios, D = Dos.
13. 5 = 556 bpi, 8 = 800 bpi, 16 = 1600 bpi
14. B = Beste kvalitet, S = Standard kvalitet

EH/JMH/GH, 6/2-73

DOKUMENTASJON AV DATA I STATISTISK SENTRALBYRÅ

M e r k n a d e r til notat av Erik Aurbakken, EA/WA, 30/1-73

av Eivind Hoffmann og Jan M. Hoem

1. Notatet gir en oversikt over det nåværende system sett fra Produksjonsavdelingen. Et av de oppdrag som fagkontorene bør få - i tillegg til dem som framgår av avsnitt 4 i notatet, er å redegjøre for hvordan dokumentasjonen i dag skjer hos fagkontorene for å ta vare på følgende hensyn:

a. Holde seg selv ajour og orientert om de løpende statistikkprosjekter og tilknytting til andre - tidligere og parallelle - prosjekter (årganger).

b. Gi svar på forespørsler utenfra - både slike som klart refererer seg til et bestemt prosjekt og slike som muligens eller klart vil kreve kontakt med flere statistikker og/eller kontorer.

c. Sette nye saksbehandlere (eventuelt nye byråsjefer) inn i arbeidet med de enkleste statistikker ved kontoret.

Vi mener at man bør gå igjennom vårt nåværende "system" for å få klarhet i hvordan man i dag løser slike oppgaver. Det vil gi en nødvendig bakgrunn for å vurdere hvordan en "oppstramming" av systemet kan gi fordeler for de aktuelle brukergrupper - fagkontorene, systemarbeiderne og "publikum". Denne gjennomgangen bør konkretiseres ved hjelp av belysende eksempler - konstruerte eller faktisk forekommende.

For en arbeidsgruppe som arbeider med utviklingen av en variabelkatalog ved Statistiska Centralbyrån, har Hoffmann presentert følgende to spørsmål for at de skal vurdere hvordan (hvorvidt) deres system kan besvare dem. De kan tjene som eksempel på spørsmål som vil berøre flere av Byråets kontorer, fra én type brukere:

1) Utgangspunktet er en antakelse om at regionale ulikheter i fruktbarhet skyldes ulikheter i befolkningens alders-struktur og andre økologiske variable. For hvilke regiontyper (kommuner, fylke, handelsområder, kirkesogn e.l.) kan det gis aldersspesifikke fruktbarhetsrater for hvert år i en fem-års-periode? Hvilke økologiske variable finnes for de ulike region-typene i samme periode? (Der det finnes flere observasjoner av en variabel ønskes midtpunktet i perioden.)

2) Vi ønsker å analysere samvariasjonen mellom personers utdanning og deres inntekt ut fra data om individer. Vi ønsker å vite hvilke observasjonssett som kan være aktuelle - med observasjoner fra 1965 eller senere. Hvilke inntektsmål har man i de ulike observasjonssett og

hvordan er utdanningsvariablen målt? Hvilke andre variable er observert om personene i disse sett?

Hvordan vil Byrået i dag gå fram for å svare på slike spørsmål?

2. Vi kan tenke oss følgende konkretisering av avsnitt 5 og 6 i notatet:

Formannen i styringsgruppen er ansvarshavende for framdriften av prosjektet. Han har en heltids prosjektassistent (konsulent II/førstesekretær) som har det daglige praktiske arbeid med utforming av referansesystemer og oppfølging og koordinering av arbeidet på fagkontorene, Systemkontoret osv. Denne assistenten bør helst være en person med erfaring både fra systemarbeid og fra et fagkontor. I det minste bør han kjenne Systemkontorets dokumentasjonssystem godt og ha erfaring som bruker av Byrået's data. Disse kvalifikasjonskrav gjør at man nok ikke kan fylle stillingen ved nytilsetting. Ventetiden før vedkommende kan frigjøres fra nåværende oppgaver, kan f.eks. nyttes bl.a. til slikt grunnlagsarbeide som vi har foreslått under punkt 1 i dette notatet.

3. Vi regner med at det opplegg Aurbakken skisserer, representerer det første trinn i den praktiske gjennomføringen av variabelkatalogprosjektet. Det etterfølgende trinn i prosjektet må diskuteres og planlegges i lys av de erfaringer som det første fremstøt gir. Slik planlegging bør antakelig foregå når en står ved slutten av det arbeid Aurbakken legger opp til.