

# Ett verksgemensamt datalager vid SCB?

Bo Sundgren



R&D Report  
Statistics Sweden  
Research - Methods - Development  
1994:1

## INLEDNING

### TILL

**R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.**

**Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.**

### **Föregångare:**

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

### **Efterföljare:**

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

# Ett verksgemensamt datalager vid SCB?

Bo Sundgren



R&D Report  
Statistics Sweden  
Research - Methods - Development  
1994:1

Från trycket  
Producent  
Ansvarig utgivare  
Förfrågningar

Mars 1994  
Statistiska centralbyrån, utvecklingsavdelningen  
Lars Lyberg  
Bo Sundgren, tel 08-783 41 48

© 1994, Statistiska centralbyrån  
ISSN 0283-8680

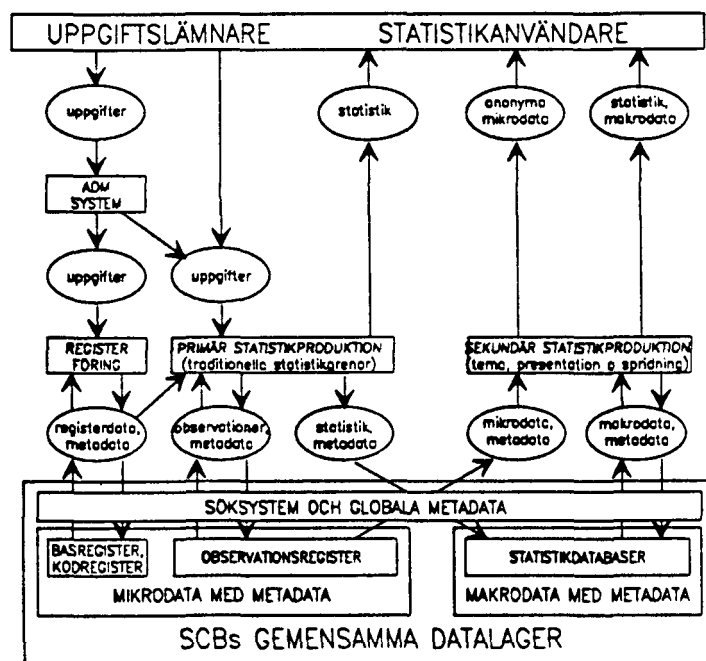
Printed  
Producer  
Address

March, 1994  
Statistics Sweden,  
Department of Research and Development  
S-115 83 Stockholm  
Lars Lyberg  
Bo Sundgren, telephone +46 08 783 41 48

Publisher  
Inquiries

© 1994, Statistics Sweden  
ISSN 0283-8680

### Ett verksgemensamt datalager vid SCB?



## Ett verksgemensamt datalager vid SCB?

*U-avdelningen har ombetts utreda olika aspekter på ett eventuellt inrättande av ett verksgemensamt datalager vid SCB. Utredningen skulle bl a jämföra för- och nackdelarna med ett sådant datalager jämfört med en situation "där de centrala registren liksom nu ligger inom ämnesavdelningsorganisationen". (Jan Eklöf, 1993-02-11)*

**Utredningsuppdraget reser omedelbart två huvudfrågor:**

- 1• *Vad är det för slags register/databaser, som ett eventuellt verksgemensamt datalager skulle innehålla?* (Denna fråga är närbesläktad med frågan om vad som i statistikpropositionen menas med databaser, som eventuellt bör anslagsfinansieras.)
- 2• *Vad menas egentligen med ett verksgemensamt eller centralt datalager?* Fysiskt centralt? Organisatoriskt centralt? Innehållsmässigt samordnat?

Jag börjar med en begreppsdiskussion kring dessa frågor.

### 1 Tänkbara typer av register/databaser i ett centralt datalager

#### 1.1 Basregister

En "ortodox" definition av begreppet "register" är ungefär följande:

Ett **register** är en (mer eller mindre officiell eller auktoriserad) förteckning av de objekt av ett visst slag, som existerar vid en viss tidpunkt.

Typiska SCB-exempel på register, som passar in på denna definition, är registret över totalbefolkningen (RTB) och centrala företags- och arbetsställeregistret (CFAR).

Ett register kan ha administrativa eller statistiska syften - eller båda. Ur statistisk synpunkt är den grundläggande funktionen för ett register (fortfarande enligt ett "ortodoxt" synsätt) att fungera som **ram** för statistiska undersökningar: registret *identifierar* och *lokaliserar* på ett eller annat sätt de objekt, som skall undersökas; vid totalundersökningar skall samtliga objekt av ett visst slag undersökas, vid urvalsundersökningar bara en delmängd av dem.

Denna typ av register kallas i fortsättningen **basregister**. Förutom identifierande och lokalisering innehåller ett basregister ofta ytterligare en del grundläggande information om objekten i registret, s k **basdata**. Ett

basregister kan därigenom i sig utgöra grunden för viss statistikproduktion, baserad enbart på variablerna i basregistret.

## 1.2 Observationsregister

Basregistrets huvudfunktion i statistikproduktionssammanhang är som sagt att "peka ut" och lokalisera de objekt för vilka uppgifter på ett eller annat sätt skall inhämtas genom en eller flera statistiska undersökningar. De uppgifter, som man samlar in genom en statistisk undersökning, organiseras i sin tur ofta efter viss bearbetning (granskning och rättning, kodning, framställning av härledda variabler m m) i en annan typ av register, ett s k (slutligt) **observationsregister**. Det slutliga observationsregistret är i själva verket tillsammans med åtföljande dokumentation (metadata) en av de viktigaste slutprodukterna från en statistisk undersökning.

Det slutliga observationsregistret från en statistisk undersökning används på åtminstone tre olika sätt:

*För det första* ger det slutliga observationsregistret mikrodataunderlaget till den aggregerings- och estimationsprocess, som resulterar i statistik (makrodata) från undersökningen. Statistiken publiceras i sin tur och ställs till statistik användarnas förfogande för vidare analyser, t ex via SM och/eller i elektronisk form, via spridningsdatabaser, disketter, CDROM, etc. Denna typ av användning av ett slutligt observationsregister kan vi kalla **primär användning** eller **primär aggregering och estimation**.

*För det andra* händer det inte sällan, att delar av det slutliga observationsregistret från en statistisk undersökning vidareanvänds av andra undersökningar. Vidareanvändning av det slag, som här avses, är i allmänhet planerad och ligger vanligen tidsmässigt i någorlunda nära anslutning till primär användningen. Ibland rör det sig om vidareanvändning (för jämförelser över tiden) inom olika tidsversioner av en och samma undersökningsserie. Denna typ av användning av ett slutligt observationsregister kan vi kalla **planerad vidareanvändning** eller **planerad sekundär användning**.

*För det tredje* sparas (arkiveras) det slutliga observationsregistret från en statistisk undersökning för mer eller mindre okänd och oplanerad framtida återanvändning av forskare, utredare m fl. Framtiden kan i detta sammanhang ligga relativt nära primär användningstillfället eller mycket långt borta, "om tiotals eller hundratals år". Denna typ av användning av ett slutligt observationsregister kan vi kalla **oplanerad återanvändning** eller **oplanerad sekundär användning efter arkivering**.

### 1.3 Kodregister

Med **kodregister** avses en förteckning av de värden, koder och namn på dessa, som ingår i en värdemängd, klassifikation, indelning, nomenklatur eller liknande.

De värdemängder, klassifikationer etc, som dokumenteras genom ett kodregister, har vanligen status av **standard** på någon nivå: internationell standard, svensk standard, SCB-standard. Rent tekniskt finns det dock givetvis inga hinder mot att använda kodregister för att dokumentera även "ostandardiserade" värdemängder etc.

I begreppet "kodregister" ingår även **transformationsregler** eller **översättningsnycklar** mellan olika koder. Till kodregistren räknas också de **koordinatregister**, som man använder inom geografisk informationsbehandling.

### 1.4 Databaser: mikro/makro/meta

En **databas** är minnesfunktionen i ett informationssystem. En mera utförlig definition av databasbegreppet lyder:

"En **databas** är en samling data, vars innehåll fortlöpande reflekterar tillstånd och förändringar i ett avgränsat stycke verklighet, **objektsystemet**. Databasen skall vara lättillgänglig för olika slags förutsedda och oförutsedda **frågor** kring objektsystemet, dess tillstånd och utveckling, frågor som är av intresse för olika grupper av användare inom olika användnings- eller **tillämpningsområden**."

En kortversion av definitionen:

"En **databas** är en *välorganiserad* samling data. Man skall på ett *enkelt och flexibelt* sätt kunna *bearbeta, förändra och utvidga innehållet* i en databas. Det skall också vara enkelt att göra olika slags *planerade och oplanerade uttag*."

Uppenbarligen kan **register** av alla de typer, som behandlats ovan, dvs *basregister, kodregister* och *observationsregister*, mycket väl vara så utformade att de uppfyller databaskriterierna enligt de just uppställda definitionerna. Egenskapen att "fortlöpande reflektera tillstånd och förändringar i objektsystemet" (jfr den längre av databasdefinitionerna ovan) är dock mest uppenbar vad gäller basregister.

Vanligen kräver man, att en databas skall vara **självbeskrivande** i betydelsen att den som en integrerad del av sig själv skall innehålla en **metadatabas**,



som beskriver databasen från såväl innehållsmässiga som tekniska utgångspunkter.

I statistiska sammanhang brukar man skilja mellan mikro- och makrodatabaser. En **mikrodatabas** är en databas, som innehåller information om enskilda objekt. En **makrodatabas** är en databas, som innehåller sammanfattande information om grupper av objekt. Sammanfattningen eller **aggregeringen** av mikrodata till makrodata görs vanligen med hjälp av en **statistisk operator** av typen *frekvensräkning*, *summering*, *medelvärdesberäkning*, *variansberäkning*, etc. Eftersom man bara i sällsynta undantagsfall har tillgång till en fullständig och garanterat korrekt uppsättning av de mikrodata, som skulle krävas, för att man skulle kunna utföra aggregeringen som en exakt beräkning, utgör aggregeringen vanligen en del av ett mer eller mindre sofistikerat **skattningsförfarande**.

Mikro/makro-distinktionen måste alltid göras relativt en tänkt eller verklig aggregeringsprocess. Makrodata är utdata från en aggregering, medan mikrodata är indata till en aggregering.

Enligt principen "*one man's ceiling is another man's floor*" kan samma data samtidigt vara både mikro- och makrodata (relativt olika aggregeringsprocesser). Ett företags redovisning kan sålunda ses som resultatet av en företagsintern aggregeringsprocess, som utgår från information om enskilda affärstransaktioner. SCB kan i sin tur utnyttja redovisningsuppgifter från enskilda företag som mikrodata, vilka aggregeras till makrodata om grupper av företag (t ex företagen i Sverige fördelade efter branschtillhörighet).

## 1.5 Statistiska databaser och statistikdatabaser

En **statistikdatabas** är en databas som innehåller statistik, dvs en makrodatabas (inklusive tillhörande metadata).

En **statistisk databas** är en databas för statistiska bearbetningar, analyser och uttag. Den kan innehålla både mikro-, makro- och metadata, men presenterade utdata från databasen utgörs normalt av makrodata i en eller annan form (tabeller, grafik etc).

## 1.6 Presentations- och spridningsdatabaser

När en statistisk undersökning genomförts skall resultaten av undersökningen ställas till statistikankvändarnas förfogande. En statistisk undersökning resulterar vanligen i slutprodukter av två olika slag:

- **mikrodata** med tillhörande förklaringar och kvalitetsdeklarationer (metadata);

- **makrodata** med tillhörande förklaringar och kvalitetsdeklarationer (metadata).

Den första typen av slutprodukter är vad som enligt definitionerna ovan kallas (slutliga) **observationsregister**.

Den andra typen av slutprodukter redovisas i form av tabeller, grafer och text i traditionella publikationer eller med hjälp av elektroniska media och är vad som vanligen kallas - helt enkelt - **statistik**.

Ur statistikansvändarsynpunkt är det föga ändamålsenligt, att varje statistisk undersökning ställer sina slutprodukter till förfogande på sitt eget sätt, utan hänsyn till hur slutprodukter från andra undersökningar presenteras och sprids. Flertalet statistikansvändare är inte bara intresserade av en enda undersökning, utan vill utnyttja data från olika undersökningar i kombination med varandra, och även i kombination med egna data och med data från andra källor än SCB. SCB måste utforma sina slutprodukter och ställa dem till förfogande med hänsyn till användarnas krav på överblickbarhet, åtkomlighet, jämförbarhet och kombinerbarhet.

Innehållsmässig och teknisk standardisering av data och metadata är viktiga instrument, som kan räcka ganska långt för att uppnå eftersträvad tillgänglighet hos slutprodukterna. För att uppnå en ännu högre grad av tillgänglighet och service gentemot användarna kan man organisera särskilt efterfrågade mikro-, makro- och metadata i särskilda **presentations-** eller **spridningsdatabaser**.

## 1.7 Metadatabaser och söksystem

Alla typer av databaser måste innehålla metadata, som beskriver innehållet i databasen ur såväl innehållsmässig som statistisk och teknisk synpunkt. Varje databas måste alltså innehålla ett delsystem, som man kan kalla **metadatabas**.

Det förekommer emellertid också metadatabaser, som är mera fristående från de databaser och datamaterial, som de beskriver. Sådana metadatabaser är ofta s k **passiva metadatabaser**. En användare kan ha glädje av en passiv metadatabas för att med hjälp av ett mer eller mindre sofistikerat **söksystem** identifiera och lokalisera potentiellt relevanta datamaterial. För att verkligen få tillgång till datamaterialen måste användaren dock vända sig till något annat system eller till en manuellt administrerad funktion.

En **aktiv metadatabas** är å andra sidan ett system, som först gör det möjligt för en användare att med hjälp av ett söksystem identifiera och lokalisera potentiellt relevanta datamaterial, och som därefter också erbjuder användare-

ren att "trycka på knappen" för att faktiskt ta ut eller bearbeta datamaterialet ifråga.

En metadatabas, som ger en överblick över data från många olika produktionssystem, kanske från organisationens samtliga system, sägs vara en **global metadatabas**. En metadatabas, vars beskrivningar är begränsade till data från enstaka produktionssystem, kallas analogt en **lokal metadatabas**. Lokala metadatabaser är ofta *hårt integrerade* med de produktionssystem, som de hör ihop med, medan globala metadatabaser är mera *löst integrerade* eller rentav *fristående*.

## 2 Tolkningar av begreppet "centralt datalager"

Begreppet "centralt datalager" kan tolkas på många olika sätt, t ex:

- Datalagret är **fysiskt centraliserat** till ett arkiv, en datacentral, en dator, ett programsystem, etc.
- Datalagret är **organisatoriskt centraliserat**, dvs en organisatorisk enhet har ansvaret för datalagret som sådant och för viktiga delar av verksamheten kring datalagret (uppdateringar, tekniskt underhåll, bearbetningar och uttag, vissa innehållsfrågor m m).
- Datalagret är **innehållsmässigt och tekniskt samordnat** i betydelsen att data är lagrade och beskrivna på ett enhetligt sätt, standarder (SCB-standarder, svenska och internationella standarder) efterlevs, etc.

Under hela stordatorepoken har SCB:s datalager av naturliga skäl varit starkt fysiskt centraliserat.

Viss organisatorisk centralisering av SCB:s datalager har existerat tidvis. Under en period fanns sålunda en registercentral (RC), sedermera vidareutvecklad till en databascentral (DBC), med ansvar i första hand för det som ovan kallades basregister (RTB, CFAR). Erfarenheterna från denna epok är inte särskilt uppmuntrande för tanken på ett fysiskt och organisatoriskt centraliserat datalager. A-avdelningen övervägde aldrig ens att släppa ifrån sig lantbruksregistret (LBR) i något avseende. F-avdelningen var ständigt frustrerad av att sakna såväl kontrollen som kompetensen beträffande företagsregistret, vilket föranledde många och destruktiva revirstrider mellan F och DBC. Endast på I-sidan tycks samarbetet ha fungerat ganska bra beträffande RTB.

Under senare år har utvecklingen, vården och driften av vissa verksgemensamma presentationsdatabaser skötts av enheten för statistiska databaser (SDB), i huvudsak med hjälp av egenutvecklade programvaror som AXIS och PCAXIS.

Den tekniska samordningen av SCBs datalager har varit relativt god, sedan dåvarande systemenheten utvecklade och GD fastställde en enhetlig systemutvecklingsmodell, SCBs systemeringsmodell, vilken bl a främjade enhetlig datalagring i s k **flata filer** och ett vidsträckt utnyttjande av **standardprogram**, såväl inköpta som egenutvecklade.

SCBs systemeringsmodell innehöll vissa instrument, främst beskrivnings- och analysmetoder - s k **infologisk modellering** - som i sig främjar en viss grad av **inhållsmässig samordning**. Dessa instrument har förbättrats i det föreslagna nya **dokumentationssystemet SCBDOK**, främst genom att relevanta statistiska begrepp och metoder har inarbetats i modellen.

Totalt sett har dock den inhållsmässiga samordningen av SCBs datalager varit tämligen svag under senare år. Den verksgemensamma funktionen för inhållsmässig samordning (STAND) har successivt rustats ner, samtidigt som ansvaret för samordningen i stor utsträckning lagts ut på ämnesavdelningarna.

Såsom närmare kommer att diskuteras i det följande, kan man förvänta sig följande utvecklingstendenser under de närmaste åren:

- en tämligen **snabb och långtgående fysisk decentralisering** av SCB:s produktionssystem, inklusive stora delar av datalagret, i takt med att stor-datordriften ersätts med billigare och lättillgängligare lokal datorkraft;
- en **förstärkt inhållsmässig och teknisk samordning** med fokus på **EG-samordning** och på **standardiserade gränssnitt och utbytesformat** (såväl formella standarder som kommersiellt etablerade *de facto*-standarder);
- en ungefärligen **oförändrad grad av organisatorisk decentralisering** vad gäller ansvaret för de olika komponenterna i SCB:s datalager; ett **mera välartikulerat centralt ansvar** kommer dock att krävas för vissa inhållsmässiga och tekniska standarder (jfr föregående punkt) och för vissa arkiv, presentationsdatabaser och söksystem.

### 3 SCB:s gemensamma datalager - en vision

*I ett framtida idealtillstånd är SCB:s samlade data- och metadatatillgångar enkelt, billigt och säkert tillgängliga för alla interna och externa användare, i den utsträckning de är behöriga att ta del av dem.*

### 3.1 Innehåll och funktionella egenskaper

SCB:s gemensamma datalager skall i första hand innehålla följande typer av datamaterial:

- **basregister** (ramar och datakällor);
- **slutliga observationsregister** från de statistiska undersökningarna;
- **presentations- och spridningsdatabaser**;
- **metadatabaser, inklusive kodregister.**

*Definition. Med SCBs gemensamma datalager skall vi fortsättningsvis avse datamaterial, som enligt definitionerna ovan utgör (delar av) basregister, slutliga observationsregister, presentations- och spridningsdatabaser, kodregister eller andra metadatabaser, och som samtidigt uppfyller följande villkor:*

1. *Datamaterialet har framställts av SCB.*
2. *Framställningen av datamaterialet har direkt eller indirekt finansierats genom anslag till SCB eller annan myndighet.*
3. *Interna eller externa användare är intresserade av att ha tillgång till datamaterialet.*

Även data/metadata vid andra myndigheter och (t ex internationella) organ kan efter överenskommelse med berörda myndigheter/organ på lämpligt sätt associeras med SCBs gemensamma datalager.

*Data, som inte faller inom någon av dessa kategorier, torde i huvudsak kunna betraktas som tillfälliga arbetsmaterial, som kan (och bör) förstöras, sedan de använts färdigt inom respektive undersöknings produktionsprocess.*

För närvarande torde SCBs datalager innehålla en hel del onödig, oönskad, och därmed skadlig **redundans** (dubbellagring). Under hela datorepoken har datalagret visserligen varit extremt fysiskt centraliserat, men ingen instans har haft någon som helst samlad överblick över datalagrets innehåll. Om man hade en sådan överblick, skulle det bli en möjlighet att ekonomisera med minnesutrymme på ett mera rationellt sätt än genom de "brandkårsutryckningar" i krissituationer, som hittills varit tämligen regelbundet återkommande. Datalagret skulle dessutom bli lättare att söka i och mera lättadministrerat i största allmänhet.

*Enkla överslagsberäkningar tyder på att SCBs datalager, om det gjordes mera välorganiserat och väldokumenterat, skulle kunna reduceras till något tämligen överblickbart och hanterbart med hjälp av billiga tekniska lösningar, som redan idag finns tillgängliga eller väntar "runt hörnet".*

### 3.2           Precisering av grundläggande tillgänglighetskrav

För att SCBs gemensamma datalager skall vara enkelt, billigt och säkert tillgängligt för användarna, måste följande krav - som vi i fortsättningen skall kalla de **grundläggande tillgänglighetskraven** - vara uppfyllda:

1. Det är **enkelt och billigt för användaren att hitta de data och metadata**, som han eller hon söker.
2. Det är **enkelt och billigt att selektera/kopiera data och metadata** från datalagret till en form (dator/programvarumiljö/lagringsformat), som passar användaren, och som uppfyller eventuella sekretessrestriktioner.
3. Selekterade/kopierade data åtföljs av **dokumentation, som gör det möjligt för användaren att tolka, analysera och jämföra data** på ett *lika säkert och ansvarsfullt sätt* som de kunde, som genomförde de undersökningar (motsv), som data ytterst härrör från.

**Anm.** Det tredje tillgänglighetskravet implicerar bl a följande. Om selekterade/kopierade data härrör ur flera olika undersökningar (motsv) innehåller dokumentationen information, som gör det möjligt för användaren att bedöma jämförbarheten hos data, som användaren önskar jämföra på ett eller annat sätt.

*För direkt eller indirekt anslagsfinansierade statistikprodukter bör det alltid anses ingå i uppdraget från statsmakterna att uppfylla de tre grundläggande tillgänglighetskraven ovan. Misslyckas den statistikansvarige med detta, har denne inte fullgjort det uppdrag, som anslaget givits för.*

De grundläggande tillgänglighetskraven har högsta prioritet, vilket innebär, att eventuellt planerad publicering, analysverksamhet m m utöver vad som direkt följer av nämnda krav, får stå tillbaka, om anslagsmedlen inte räcker till för allt.

Observera, att vi i denna diskussion tills vidare inte tagit ställning till hur SCBs gemensamma datalager skall vara organiserat. Genom de grundläggande tillgänglighetskraven har vi visserligen fastlagt vissa funktionella karakteristika hos det gemensamma datalagret, dvs vi har angivit, hur det skall te sig gentemot användaren, men vi har exempelvis inte tagit ställning till i vilken utsträckning det gemensamma datalagret skall vara fysiskt och/eller organisatoriskt centraliserat. Dessa frågor skall vi istället återkomma till i avsnitt 4.

### 3.3 Produktionssystem och statistikprogram

Vi skall nu analysera det önskvärda samspelet mellan ett SCB-gemensamt datalager och de olika **produktionssystem**, som "föder" datalagret med data och metadata. Statistikproduktionssystemen vid SCB - med ungefärligen samma innebörd även kallade (statistiska) **undersökningar**, **statistikgrenar** eller (statistik)**produkter** - grupperas sedan några år tillbaka i **statistikprogram**. Statistikprogrammen definieras på basis av samband mellan produkterna, som väsentligen är av produktionsmässig natur, men som också i någon mån tar hänsyn till produkternas användningar.

Statistikprogrammen väntas spela en kanske ännu viktigare roll än hittills efter den nu påbörjade organisationsöversynen. I bakgrundsmaterialet till organisationsöversynen talas om *två typer av statistikprogram*:

- **produktionsprogram**, definierade som "traditionella statistikprogram med egen produktion";
- **temaprogram**, definierade som "program som arbetar med sammanställningar tvärs över organisationen".

Genom att skärpa de citerade definitionerna något och lägga till ytterligare en programkategori får vi en lämplig indelning för att diskutera **data- och metadataflödena i ett framtida nationellt och internationellt statistiskt system**, och för att precisera rollen hos ett eventuellt gemensamt datalager vid SCB i ett sådant system.

Vi skall sålunda fortsättningsvis tala om *tre typer av statistikprogram*:

- *registerprogram*;
- *traditionella statistikproduktionsprogram*;
- *förädlingsprogram*;

som skulle kunna definieras enligt den följande framställningen.

#### **Registerprogram**

Ett **registerprogram** är ett program som syftar till att bygga upp, förvalta och vidareutveckla ett eller flera (datainsamlings- och produktionsmässigt nära relaterade) **basregister** och/eller **kodregister** vid SCB.

**Basregistren** baseras vanligen på **externa administrativa register**, som eventuellt kompletteras genom särskild datainsamling i SCB:s regi. Ett basregister kan ha **två huvudfunktioner**, som ibland kombineras:

- att utgöra **ram** (vid totalundersökningar) respektive **urvalsram** (vid urvalsundersökningar) för andra statistiska undersökningar;
- att utgöra en **datakälla** för andra statistiska undersökningar.

**Kodregistren** byggs vanligen upp och förvaltas genom ett klassificerings- och standardiseringsarbete vid SCB, inte sällan i samverkan med andra myndigheter, och ofta på basis av ett motsvarande arbete på internationell nivå.

Registerprogram (avseende såväl basregister som kodregister) utgör **stödprogram** för de båda andra typerna av statistikprogram (traditionella statistikproduktionsprogram och förädlingsprogram; se nedan).

#### *Traditionella statistikproduktionsprogram*

Ett **traditionellt statistikproduktionsprogram** omfattar en eller flera (produktionsmässigt nära relaterade) **statistiska undersökningar**, som helt eller delvis baseras på datainsamling i SCB:s egen regi, och som resulterar i ett eller flera **slutliga observationsregister** (som arkiveras) och viss **publicerad statistik** (i traditionell och/eller elektronisk form).

#### *Förädlingsprogram*

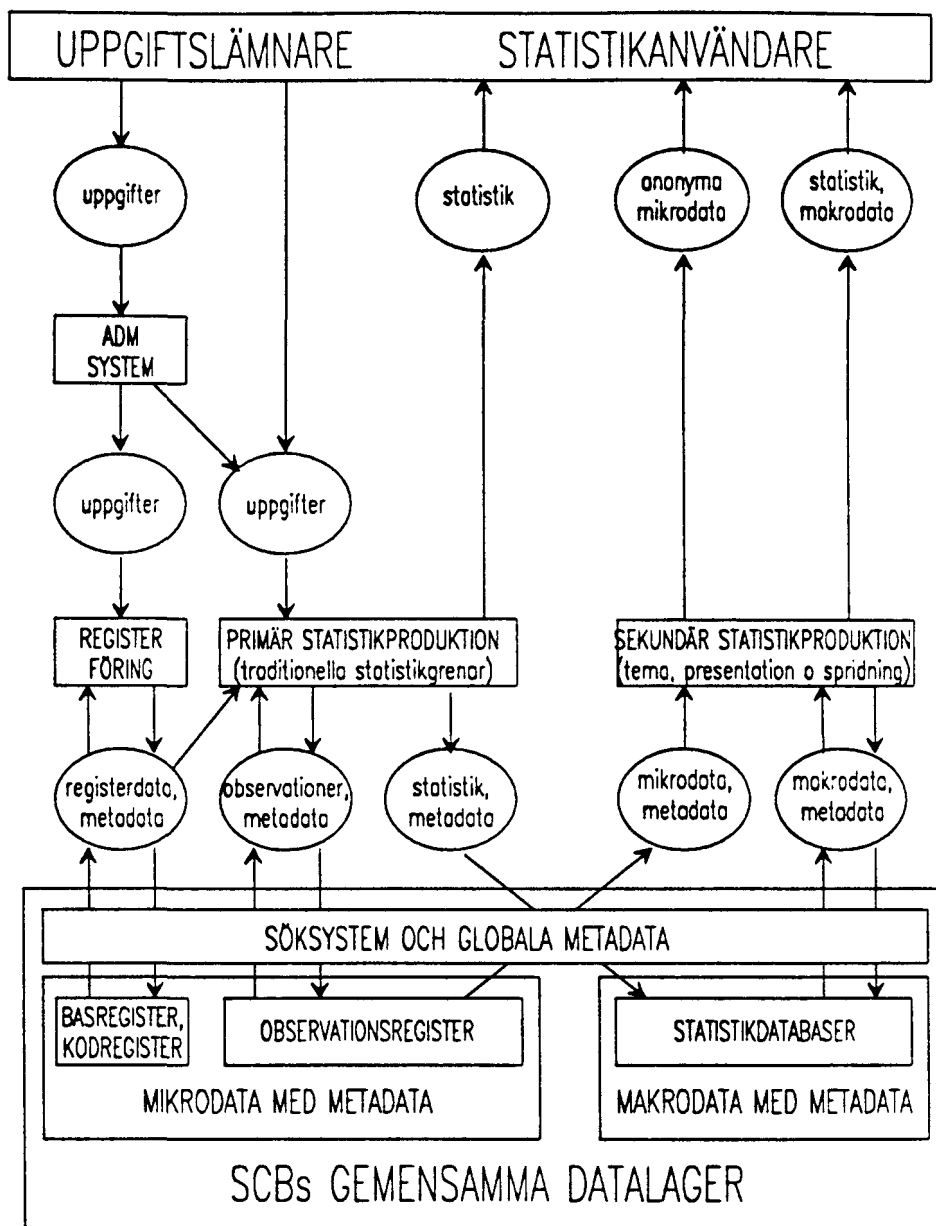
Ett **förädlingsprogram** förädlar mikrodata och/eller markrodata (statistik) från ett eller flera andra statistikprogram (basregisterprogram, traditionella statistikproduktionsprogram och/eller förädlingsprogram), t ex genom att vidarebearbeta, analysera och/eller kombinera data/statistik från de andra programmen.

**Presentations- och spridningsdatabaser** kan i sig själva utgöra förädlingsprogram. De kan dessutom utgöra stödprogram för andra förädlingsprogram, som ytterligare förädlar innehållet i databaserna för specifika användarkategoriers räkning.

### **3.4 Samspelet mellan statistikprogram och datalager**

De tre typerna av statistikprogram skulle samspela med ett tänkt SCB-gemensamt datalager enligt vissa typiska mönster, som närmare beskrivs i det följande. Se även **figur 1**.





**Figur 1.** Samspelet mellan uppgiftslämnare, statistikanvändare och ett tänkt SCB-gemensamt datalager - via tre olika typer av statistikprogram: registerprogram, primära statistikproduktionsprogram (traditionella statistikgrenar) och sekundära statistik-produktionsprogram (temaprogram, presentations- och spridningsdatabaser och andra typer av förädlingsprogram). Observera att det gemensamma datalagret kan realiserats i många olika former - från hårt fysiskt och organisatoriskt centraliserade till starkt decentraliserade lösningar, där datalagret hålls samman med hjälp av standardiserade gränssnitt och databytesformat; se vidare texten.

### *Samspelet mellan ett registerprogram och ett gemensamt datalager*

Vissa versioner av de **basregister** respektive **kodregister** som skapas och förvaltas i ett registerprogram skulle i sig själva (eller i form av kopior/utdrag) utgöra en del av ett gemensamt datalager, varifrån de kan vidareutnyttjas omedelbart och/eller vid senare tillfällen av andra statistikprogram (och även av samma registerprogram, som producerat dem).

De registerversioner, som är aktuella för att ingå i ett gemensamt datalager, har i princip samma karaktär som de slutliga observationsregistren från ett traditionellt statistikproduktionsprogram. Basregister och kodregister har dock ofta egenskapen att de **uppdateras fortlöpande**, i takt med att de registrerade storheterna "föds" och "dör" eller förändras i andra register-relevanta avseenden. För sådana register ställs man bl a inför problemet, huruvida man i det gemensamma datalagret skall inkludera

- *enbart* **händelserapporterande information** (transaktioner som informerar om registerobjekts födelse/död/förändring); eller
- *både* händelserapporterande information enligt ovan *och* **samlad tillståndsinformation** avseende samtliga existerande registerobjekts tillstånd vid valda tidpunkter (t ex årsvis, kvartalsvis, månadsvis).

Observera att det alltid är nödvändigt att lagra den händelserapporterande informationen, om man inte medvetet vill förstöra information. Om fullständig händelserapporterande information lagras, är det å andra sidan inte logiskt nödvändigt att också lagra tillståndsinformation - denna kan alltid härledas för en godtycklig tidpunkt - men det kan ändå ur tillgänglighets-synpunkt vara önskvärt att göra det. Det kan nämligen vara både dyrt och tidsödande att göra härledningar av tillståndsinformation ur händelseinformation varje gång tillståndsinformation efterfrågas.

### *Samspelet mellan ett traditionellt statistikproduktionsprogram och ett gemensamt datalager*

Ett traditionellt statistikproduktionsprogram skulle utnyttja flera typer av datamaterial från ett gemensamt datalager:

- **basregister** (som ram, urvalsram och/eller datakälla);
- **kodregister** (för fördefinierade värdemängder, indelningar, benämningar etc);
- **slutliga observationregister** från andra (eller samma) statistikproduktionsprogram.

Ett traditionellt statistikproduktionsprogram producerar i sin tur två typer av data/metadata-produkter av intresse för ett gemensamt datalager:

- **slutliga observationsregister** med tillhörande dokumentation;
- **statistik** med tillhörande förklaringar.

Två eller tre versioner av ett slutligt observationsregister är av särskilt intresse för det gemensamma datalagret:

1. **det fullständiga observationsregistret**, som innehåller data/metadata motsvarande hela informationspotentialen från den ursprungliga undersökningen;
2. **det avidentifierade observationsregistret**, där man avlägsnat all information, som skulle göra det möjligt att omedelbart (dvs utan särskilda bearbetningar med sikte på bakvägsidentifiering etc) identifiera enskilda objekt i registret;
3. **det obegränsat spridningsbara observationsregistret**, det som i många länder går under beteckningen "*public file*" ett datamaterial, som är så anonymiserat och uttunnat vad gäller känslig information, att det utan fara för integritetsintrång och andra sekretessbrott kan spridas fritt till allmänheten, utan särskild prövning i varje enskilt utlämnandefall.

#### ***Samspelet mellan ett förädlingsprogram och ett gemensamt datalager***

I bakgrundsmaterialet till organisationsöversynen skisseras en utveckling med ett växande antal relativt små och tillfälliga temaprogram. Temaprogrammen skulle vara mycket flexibla och kunna bildas och läggas ner efter behov. Antalet anställda på ett temaprogram skulle vara litet, två eller fler, och programmet skulle kunna ha planeringsansvar men inte personalansvar. Temaprogrammen skulle ges en egen budget, som de själva skulle kunna fördela i organisationen för produktion och utvecklingsarbete.

Flexibla temaprogram av ovan skisserad typ torde bli ett allt viktigare medel i en offensiv användaranpassning av SCB:s verksamhet och en vidareutveckling av SCB:s roll som kunskapsföretag. Temaprogram av olika slag skulle kunna förekomma på olika håll i SCB:s organisation, t ex:

- ***Produktionslokala förädlingsprogram***

Med "produktionslokala förädlingsprogram" avses här temaprogram/förädlingsprogram med nära anknytning till ett antal traditionella statistikproduktionsprogram, som är nära relaterade till varandra ur produktions-synpunkt. Temaprogram av detta slag förutsätter, att det finns en viktig grupp användare vars statistikbehov i huvudsak kan tillgodoses med hjälp av

vidareförädlade data från de produktionsmässigt sammanhörande statistikprogrammen. Till denna kategori av förädlingsprogram kan man räkna exempelvis **nationalräkenskaperna**.

- ***Tvärprogram***

Med "tvärprogram" avses program med inriktning på att tillgodose en väl-identifierad, "tung" användargrups behov av förädlade produkter, som här-rör från olika statistikprogram, vilka produktionsmässigt delvis är föga be-släktade med varandra, och som därför handhas inom olika avdelningar vid SCB.

- ***Breda och översiktliga program för allmänhetens behov***

Ett mycket brett och översiktligt program skulle kunna ha syftet att tillgodose en icke närmare specificerad allmänhets översiktliga statistikbehov.

- ***Breda och djupa program för professionella statistikanvändare***

Ett både brett och djupt program kan behövas för att tillgodose behoven av såväl mikrodataber som statistik hos en tämligen väldefinierad kategori av mycket professionella statistikanvändare (forskare, utredare, etc) enligt själv-betjäningsprincipen - med möjligheter till kvalificerad konsult hjälp, om statistikanvändarna själva så önskar.

*Ett välorganiserat datalager torde vara en helt nödvändig förutsättning för att SCB med önskvärd snabbhet och till rimliga kostnader skall kunna tillgodose vissa typer av statistikanvändares ibland tämligen oplanerade och plötsligt aktualiserade statistikbehov.*

- ***Presentations- och spridningsdatabaser***

Presentations- och spridningsdatabaser har flera viktiga roller i detta sammanhang:

*För det första* utgör de **ingångar till SCBs gemensamma datalager**. En väl-utformad presentations- och spridningsdatabas ger en god överblick över till-gängliga datamaterial inom ett visst område. Den kan genom sin inbyggda metadatadel innehålla sökvägar och referenser även till sådana datamaterial, som inte finns lagrade i den betraktade databasen själv.

*För det andra* ger skräddarsydda presentations- och spridningsdatabaser en **extra hög tillgänglighet** för de behov, som den skräddarsyfts för. Den högre tillgängligheten har naturligtvis ett **pris**. När man utvecklar och förvaltar en presentations- och spridningsdatabas är det därför viktigt att fortlöpande analysera olika marknadssegments villighet och förmåga att betala för olika tillgänglighetshöjande funktioner. Vissa kunder kan t ex vara

beredda att betala ett mycket högt pris för den höga aktualitet hos informationen, som är möjlig att upprätthålla i *on line-databaser*, medan denna funktionalitet kan vara tämligen ointressant för andra kunder, som måhända snarare prioriterar låg kostnad och/eller standardiserade lagringsformat, som gör det enkelt för användaren att själv vidarebearbeta data och metadata, med hjälp av populära programvaror.

*För det tredje* kan presentations- och spridningsdatabaser i sin tur ge **underlag för andra användaranpassade produkter och tjänster** liksom för att rationalisera produktionen av reguljära publikationer m m inom ramen för de traditionella statistikproduktionsprogrammen. Exempel kan hämtas från F-avdelningens användning av SCBs statistiska databaser.

Presentations- och spridningsdatabaserna utgör uppenbarligen en viktig del av SCBs gemensamma datalager. Eftersom de normalt framställs och förnyas med hjälp av data och metadata från andra delar av det gemensamma datalagret, är de vanligen logiskt sett redundanta. Ur tillgänglighetssynpunkt kan det ändå vara fördelaktigt att hålla sig med vissa presentations- och spridningsdatabaser. Det är exempelvis tidsbesparande och kostnadseffektivt att lagra antalet invånare i Sverige i en makrodatabas, trots att denna uppgift skulle kunna härledas ur RTB, varje gång den efterfrågas.

Förädlingsprogrammen har uppenbarligen stor nytta av de mikrodata, makrodata och metadata, som redan finns i ett gemensamt datalager. Man kan omvänt fråga sig om förädlingsprogrammen i sin tur ger upphov till ny, användbar information. En möjlighet är att använda söksystem och presentations- och spridningsdatabaser för att systematiskt fånga upp information om statistikanvändarnas användning av olika delar av det gemensamma datalagret och deras synpunkter på de produkter och tjänster i sammanhanget, som de finner - eller *inte* finner.

## **4 Slutsatser och åtgärdsförslag**

### **4.1 Kommunikation via standardiserade gränssnitt**

För att kraven på tillgänglighet, samordning och flexibilitet skall kunna tillgodoses till så låg kostnad som möjligt, måste produktionssystemen, som föder SCBs gemensamma datalager med data och metadata utformas på lämpligt sätt. Fysisk och organisatorisk centralisering av berörda produktionsfunktioner är ett enkelt men tämligen trubbigt sätt att tillgodose kraven på enhetligt utformade data och metadata med hög tillgänglighet. Då fysisk och organisatorisk centralisering står i direkt motsatsförhållande till de önskemål om en "plattare" och mera decentraliserad organisation, som uttrycks i direktiven till organisationsutredningen, finns det ingen anledning att annat än möjligen punktvis och i synnerligen välmotiverade undantagsfall överväga åtgärder av detta slag.

*Enligt modern systemteori åstadkommer man enklast och billigast mångsidig användbarhet hos ett system och dess olika komponenter, genom att se till att de kan kommunicera med varandra via standardiserade gränssnitt. Man standardiserar sålunda i första hand kommunikationen och gränssnitten mellan systemen/systemkomponenterna - inte systemen och systemkomponenterna som sådana.*

Hur skulle då produktionssystemen vid SCB utformas, om man vill följa systemteorins principer för att tillgängliggöra relevanta data och metadata på ett användarvänligt och samtidigt kostnadseffektivt sätt? Hur skall data-lagring och dataflöden samordnas, samtidigt som man håller fast vid de förenklings- och decentraliseringsidéer, som skall vara vägledande för organisationsöversynen, och som kraftigt understödjes av den pågående och förväntade tekniska utvecklingen i riktning *från* tekniskt enhetliga men dyrbara och tämligen rigida system och i riktning *mot* billigare, flexiblere och mera användaranpassade lösningar i PC-miljö?

Detta är viktiga frågor för SCB. Några försök till svar kommer i de följande punkterna.

#### 4.2 Internationell standardisering av statistiskt datautbyte

Tillämpad på statistikproduktionen innebär den just nämnda systemteoretiska principen, att produktionssystemen hos de statistikansvariga förses med standardiserade gränssnitt, via vilka data och metadata kan utväxlas med andra system (t ex söksystem och presentationsdatabaser) enligt **standardiserade meddelandeformat**. På internationell nivå håller detta synsätt på att utvecklas och etableras inom projekt som

- FN/EG/OECD-projektet "*EDI and Statistics*" (Western European EDIFACT Board: Message Development Group 6 - Statistics); och
- EG-projektet "*Distributed Statistical Information Systems*".

Arbetet på en internationell standard för utbyte av statistiska data och metadata (GESMES: *Generic Statistical Message*) har kommit ganska långt. Från svensk sida deltar bl a Erik Malmberg och Lars Rauch från SCB aktivt i detta arbete. Eftersom EG-standarder förväntas få en alltmer styrande effekt på det svenska statistiska systemet, är det angeläget att SCB även fortsättningsvis bevakar och påverkar utvecklingsarbetet på detta område så aktivt som möjligt.

### 4.3 Ökad tillgänglighet hos datalagret

Statsmakterna och statistikanvändarna förväntar sig uppenbarligen att SCB skall förbättra och förbilliga tillgängligheten till SCBs datamaterial. I statistikpropositionen säger sålunda departementschefen:

*"Särskilda insatser kan också behövas för att öka statistikens tillgänglighet och underlätta användning och framtagning av statistik. I detta sammanhang bör forskningens speciella behov särskilt uppmärksammas. Möjligheten att ställa statistiska databaser till användarnas förfogande bör utvecklas inom ramen för de rättsliga förutsättningar som föreligger enligt bl a datalagen och sekretesslagen. SCB har i sin treårsrapport föreslagit en omfattande satsning på att öka tillgängligheten i dessa avseenden. Detta ligger således i linje med min bedömning."*

Den SCB-satsning, som departementschefen hänvisar till pågår f n inom ramen för *Tillgänglighetsprojektet*.

I avsnitt 3.2 ovan formulerades tre grundläggande tillgänglighetskrav, som måste vara uppfyllda för att SCBs datalager skall kunna anses vara enkelt, billigt och säkert tillgängligt för användarna:

1. Det är enkelt och billigt för användaren att hitta de data och metadata, som han eller hon söker.
2. Det är enkelt och billigt att selektera/kopiera data och metadata från datalagret till en form (dator/programvarumiljö/lagringsformat), som passar användaren, och som uppfyller eventuella sekretessrestriktioner.
3. Selekterade/kopierade data åtföljs av dokumentation, som gör det möjligt för användaren att tolka, analysera och jämföra data på ett lika säkert och ansvarsfullt sätt som de kunde, som genomförde de undersökningar (motsv), som data ytterst härrör från.

För att de grundläggande tillgänglighetskraven skall kunna uppfyllas, bör man bl a vidta följande åtgärder:

#### **Identifiera "tillgänglighetskandidater"**

Vi behöver mera i detalj klargöra vilka data det är som skall vara tillgängliga. Denna rapport har identifierat följande typer av datamaterial som "tillgänglighetskandidater":

- basregister
- observationsregister
- kodregister
- statistikdatabaser

Inom tillgänglighetsprojektet har Björn Magnusson påbörjat en snabb-inventering av vilka register och databaser inom de olika statistikprogrammen, som kan vara konkreta "tillgänglighetskandidater". Detta arbete bör snarast slutföras, så att vi får åtminstone en grov uppfattning om vilka datamaterial, som bör anses ingå i SCBs gemensamma datalager, och som i konsekvens därmed skall uppfylla vissa **standarder** vad gäller lagring, dokumentation och tillgänglighet.

De just nämnda standarderna för lagring, dokumentation och tillgänglighet bör preciseras:

#### ***Precisera standarder för lagring och utbyte av statistiska data***

Vad gäller lagring och kommunikation av mikrodata, makrodata och metadata bör SCB noga följa och aktivt delta i det internationella standardiseringsarbetet. Den internationellt etablerade och kommersiellt kraftigt understödda **SQL-standard** för datalagring och datautbyte i allmänhet är en lämplig utgångspunkt för standardiseringen. Det återstår dock en del arbete med att i detalj klargöra hur SQL-standard skall tillämpas i statistiska sammanhang, framför allt vad gäller makrodata och metadata. SCBs arbete med PCAXIS har därvidlag givit värdefulla erfarenheter (se vidare nedan).

#### ***Fastställ dokumentationsstandarder***

Vad gäller dokumentation bör **SCBDOK-förslaget** fastställas och en tidsplan för dess införande utarbetas. SCBDOK innehåller f n preciserade mallar för dokumentation av produktionssystem och observationsregister. Motsvarande mallar för statistikdokumentation bör utarbetas. Arbetet har påbörjats under överinseende av *Kvalitetsdirektionen*. Datorstöd för dokumentationsarbetet utvecklas f n inom ADB-enheten. Prototyper finns redan och en produktionsversion beräknas vara klar till budgetårsskiftet. Arbetet bör därefter fortsättas med sikte på att ytterligare effektivisera dokumentationsarbetet och metadatautbytet inom SCB.

#### ***Utveckla ett verksgemensamt söksystem för mikrodata och statistik***

Även om vi inte skall ha ett fysiskt och organisatoriskt centraliserat datalager, så är det angeläget att vi förbättrar överblicken över alla de data som skall anses ingå i vårt gemensamma datalager, och att vi utvecklar ett bra **söksystem**, baserat på välorganiserade **globala metadata**, som gör det möjligt för såväl externa som interna användare att snabbt identifiera och lokalisera de mikrodata och den statistik i olika former, som skulle kunna vara av intresse för en viss användare i en viss situation - oavsett var mikrodatamaterialen respektive statistiken har sin fysiska och organisatoriska hemvist.



Idag finns den teknik och de programvaror, som gör det möjligt att utveckla mycket kraftfulla söksystem av nämnda slag. Ett sådant söksystem är dock beroende av, att vi har dokumenterat och klassificerat våra data och vår statistik på ett ändamålsenligt sätt. Här har dokumentationen enligt SCBDOK återigen en nyckelroll, liksom den **statistiktesaurus** och de **ämnesområdesindelningar**, som är under utveckling inom S-avdelningen.

#### 4.4 Sänkta kostnader för externa användare

Kostnadsfrågan är av mycket stor betydelse för tillgängligheten. Ett vanligt klagomål från externa statistikanvändare är att vi är för dyra. I det följande lämnas ett antal förslag till kostnadssänkande åtgärder.

##### *"Down-sizing": flytta ner produktionssystem från stordatorn*

För att kunna sänka vår kostnadsnivå rejält måste vi lämna stordatormiljön i största möjliga utsträckning. Som det ser ut idag, är redan kostnaden för att bara *kopiera* ett datamaterial från stordator till PC prohibitiv för många användare. Det är inte rimligt att en användare i dagens tekniksituation skall behöva drabbas av sådana kostnader. Alla relevanta datamaterial skulle idag kunna göras tillgängliga i PC- eller UNIX-miljö till mycket låg kostnad.

Det bör ingå i SCBs "**tillgänglighetsstandard**", att alla datamaterial, som överhuvudtaget skall vara tillgängliga för externa användare, skall vara tillgängliga på sådant medium och i sådant lagringsformat, att det utan någon extra kostnad för användaren kan användas av vanligt förekommande programvaror i PC-miljö. En eventuellt erforderlig kopiering av data från stordator till PC skall i förekommande fall ha gjorts "en gång för alla" inom ramen för SCBs (anslagsfinansierade) produktionssystem.

Ett naturligt sätt för SCB att minimera kostnaderna för att överföra data mellan å ena sidan stordator och å andra sidan persondatorer och s k arbetsstationer (UNIX-datorer) är att *flytta ner produktionssystemen från stordatorn i största möjliga utsträckning*. Ett rimligt mål kunde vara, att nerflyttningen (och utvecklingen av ADB-stöd för nerflyttningen) bedrivs i sådan takt, att SCB fullt ut kan utnyttja den handlingsfrihet som uppkommer, när SCBs övergångsåtaganden gentemot DAFA/Mittsystem successivt avvecklas under den närmaste fyraårsperioden.

*Det föreslås att ADB-enheten i samråd med SCBs avdelningar snarast tar fram en plan för hur en sådan omläggning - s k "down-sizing" - av SCBs produktionssystem skulle kunna genomföras så snabbt och effektivt som möjligt.*

I samband med en nerflyttning av SCBs produktionssystem och datalager från stordatorn till billigare och lättillgängligare datorer och media, måste

man se till att inga stödsystem blir kvar i stordatormiljön och där förorsakar de nerflyttade produktionssystemen onödiga kostnader. Man bör bl a titta på SCBs arkiveringssystem och AXIS-databaserna.

### *SCBs arkiveringssystem*

Riksarkivet orienterar sig alltmer mot standardiserade, öppna datormiljöer (UNIX). I princip accepterar man varken datorfabrikantberoende lagringsformat (såsom IBMs s k packade data eller kassetterna i Mittsystems robot) eller PC-media som disketter eller CD-skivor, vilka ännu anses för ostandardiserade och/eller obeprövade. På lite sikt torde dock CD-skivor bli ett utmärkt arkivmedium.

SCB överför numera väsentligen bara personregister i datalagens mening till Riksarkivet. I övrigt sker den långsiktiga lagringen av våra data vid vårt eget arkiv. SCBs arkiv består nu av 8200 filer på 5000 kassetter eller magnetband, samtliga i stordatormiljö. Uppskattningsvis skulle dessa data rymmas på något hundratal CD-skivor med den idag tillgängliga tekniken. Varje CD-skiva rymmer 600 MB. Man kan numera med hjälp av en vanlig PC som utrustats med en skrivande CD-spelare (pris: \$ 8,000) fylla en CD-skiva med 600 MB data på ca en halvtimme.

*Det förefaller uppenbart, att åtminstone en kopia av SCBs arkiv borde finnas på CD-skivor, inklusive metadata. En sådan arkivdatabas skulle i själva verket kunna vara en enkel och lämplig form av ett fysiskt och organisatoriskt centralt datalager vid SCB.*

### *Statistikdatabaserna under AXIS*

Databashanteraren AXIS är helt bunden till stordatorn. Med hjälp av PCAXIS har man dock numera en koppling till PC-miljön. Om merparten av SCBs produktionssystem och datalager flyttas ner från stordatorn enligt förslagen ovan, förefaller det orimligt, att AXIS-databaserna skulle leva kvar i stordatormiljö: dels vore datorlagringen och stordatordriften av själva databaserna onödigt kostsam, dels skulle onödiga kostnader och andra problem för dataöverföringar till och från stordatormiljön. Sist men inte minst upplevs stordatorkommunikation av de flesta användare som föråldrad i jämförelse med moderna användargränssnitt i PC-miljö.

*Slutsatsen är att programvaran AXIS, som utformades i början av 1970-talet, bör ersättas av en modern databaslösning, baserad på kommersiell programvara med SQL-gränssnitt, client/server-teknik, Windows-baserat användargränssnitt etc, och som därigenom blir fullt ut harmonierad med de standarder, som kommer att gälla för SCBs produktionssystem i övrigt.*

**Programvaran PCAXIS** är även fortsättningsvis mycket användbar, bl a som "front-end"-verktyg och som gränssnitt mot olika kommersiella

programprodukter för vidare bearbetningar och analyser av statistiska data. På "back-end"-sidan måste PCAXIS dock förses med SQL-gränssnitt.

Viktigt i detta sammanhang är också, att "nya AXIS" för SCBs gemensamma presentations- och spridningsdatabaser utformas så, att data enkelt kan utväxlas med **internationella statistikdatabaser** i enlighet med de standarder, som nu tas fram (EDIFACT, GESMES etc; se ovan).

När det gäller statistikspridningen från databaserna har billiga, standardiserade lösningar lämpade för PC-miljön högsta prioritet för många statistik-användare. Disketter och CD-skivor är här lämpliga media. Även när användarna har så höga krav på aktualitet och snabbhet i dataöverföringen, att det motiverar direktuppkoppling via telenätet till databaserna, bör enkla, billiga lösningar prioriteras.

#### 4.5 Bättre samordning

Även behovet av ökade samordningsinsatser framhålls i statistik-propositionen. Sålunda säger departementschefen:

*"Behov av samordningsinsatser för att stärka statistikens kvalitet och underlätta uppgiftslämnandet kan förväntas öka."*

Trycket på SCB att förbättra samordningen kommer från flera håll:

- En mera decentraliserad statistikmodell leder till ökade samordnings-behov. (Jfr citatet ovan.)
- EG ställer stora krav på statistiksamordning redan i och med EES-avtalet.
- Statistikanvändarna ställer ökade krav på att statistiken är jämförbar över tiden, rumsligt/administrativt samt även i övrigt vad gäller indelningar, klassifikationer och definitioner.
- De övergripande kraven på rationalitet och effektivitet i de statistiska produktionssystemen ställer krav på teknisk samordning.

Samordningskraven har viktiga implikationer för ett SCB-gemensamt datalager:

#### *Innehållsmässig samordning*

Mikrodata och makrodata (statistik) i SCBs gemensamma datalager skall vara *innehållsmässigt* utformade och dokumenterade enligt nationella och internationella standarder. Det gäller **variabeldefinitioner, mätmetoder,**

**indelningar och klassifikationer m m.** Eventuella avvikelser skall vara tydligt markerade och motiverade.

### ***Lagringsteknisk samordning***

Data i SCBs gemensamma datalager skall också vara *lagringstekniskt* standardiserade, så att t ex viktiga variabler, som är gemensamma för många datamaterial, genomgående lagras på samma sätt.

### ***Kodregistren som samordningsinstrument***

Kodregistren skulle - tillsammans med lämpligt utformad programvara - kunna få en central roll i såväl den innehållsmässiga som den tekniska samordningen av SCBs datalager. Som ett första steg i denna riktning har U/STAND uppdragit åt U/ADB att skissera ett datorstöd för hantering av klassifikationer m m. Detta arbete bör fullföljas och vidareutvecklas.

### ***Översyn av SCBs register- och databasstruktur***

SCBs register- och databasstruktur bör ses över ur samordningssynpunkt, förslagsvis på basis av Jan Eklöfs rapport "*Samband mellan SCBs statistikprodukter m m - Genomgång av centrala produktionssamband och samutnyttjande av basuppgifter vid SCB*". Ett exempel på en tänkbar förändring vore en sammanslagning av Centrala företags- och arbetsställeregistret (CFAR) med Centrala lantbruksregistret (LBR).

## **4.6 Central databasadministration**

Författaren till denna rapport har inte lyckats finna några starka argument för att bygga upp ett fysiskt och organisatoriskt starkt centraliserat datalager vid SCB. Tvärtom tenderar den förväntade successiva avvecklingen av stordator-tekniken att leda till en fysisk decentralisering av datalagret, och det finns enligt utredarens mening goda skäl att bejaka denna utveckling.

*Ett fysiskt och organisatoriskt centralt datalager bör sålunda inte byggas upp vid SCB - undantagandes den i avsnitt 4.4 nämnda arkivdatabasen, som, om den lagras och tillgängliggörs på lämpligt sätt (t ex på CD-skivor), enkelt och billigt skulle tillgodose många viktiga behov.*

*Emellertid finns det på vissa strategiska punkter starka argument för en rejäl uppstramning av den centrala kontrollen. De strategiska punkterna har alla berörts ovan. Det gäller etablerandet och upprätthållandet av vissa innehållsmässiga standarder, data- och metadatautbytesformat, dokumentation, söksystem m m. Det synes vara väl motiverat att i SCBs nya organisation peka ut en instans med ett samlat ansvar för dessa frågor, en verksgemensam databasadministratörsfunktion.*

Som "**legal plattform**" för sin verksamhet skulle databasadministratören ha en uppsättning verksbeslut, som anger ett antal standarder och riktlinjer för datalagringen och datahanteringen vid SCB, som de statistikprogramansvariga har att följa i sin verksamhet.

Den legala plattformen torde emellertid inte i sig vara ett tillräckligt stöd för att databasadministratören skall kunna leva upp till sitt ansvar. Det behövs också **motivationshöjande instrument**.

Exempelvis skulle ett **ekonomiskt styrinstrument** kunna utformas på ungefär följande sätt. När en statistikprogramansvarig tilldelas en budget, innehålls vissa belopp till dess att vissa "plikter" enligt den ovan angivna "legala plattformen" fullgjorts. Det är databasadministratörens uppgift att följa upp och "boca av" (och därmed frigöra innehållna budgetmedel) allteftersom plikterna fullgörs. Om någon plikt inte fullgörs inom föreskriven tid, skulle databasadministratören kunna använda de innehållna medlen till att lägga ut "pliktuppdraget" till någon annan instans än det ansvariga statistikprogrammet, t ex någon lämplig stödfunktion inom SCB eller en extern konsult.

En annan viktig kategori av motivationshöjande instrument skulle vara **datorstödda verktyg**. En standard får av naturliga skäl mycket större genomslagskraft, om den stöds av ändamålsenliga verktyg, som gör arbetet lättare, om man följer standarden. **Datorstödet PCDOK** för dokumentationsstandarden SCBDOK är ett aktuellt exempel. En välskött **kodregister-databas** med bra gränssnitt mot applikationsprogram och användare skulle kunna vara ett annat. Ett tredje viktigt exempel skulle kunna vara det i avsnitt 4.3 föreslagna **verksgemensamma söksystemet**, baserat på moderna programvaror och användargränssnitt, välorganiserade SCB-globala metadata, en avancerad statistiktesaur och en moderniserad ämnesområdesindelning.

Databasadministratörsfunktionen skulle lämpligen vara en mycket liten funktion, någon enstaka person, men den skulle ha stort ansvar och stora befogenheter. Vad gäller problemlösande, utveckling av ADB-verktyg och genomförande av åtgärder skulle den repliera på andra verksfunktioner såsom funktionerna för innehållsmässig samordning och standarder, internationell samverkan, ADB m fl.

#### 4.7 Säkerhet och sekretess

De säkerhets- och sekretessfrågor, som aktualiseras av utvecklingen mot ökad tillgänglighet och ökad fysisk decentralisering av SCBs datalager utreds för närvarande inom *Tillgänglighetsprojektet* i samarbete med SCBs *Dataskyddsråd*. De har därför inte behandlats i denna rapport. Hittills har

dock ingenting framkommit som tyder på att vi skulle behöva "tumma på säkerheten" för att kunna öka tillgängligheten hos vårt gemensamma datalager - oavsett vilken grad av centralisering/decentralisering vi väljer inom ramen för den modell som skisserats i föreliggande rapport.

Allmänt sett är det rimligen i hanteringen av identifierade data på mikronivå, som vi har de största riskerna. En sådan hantering är ofrånkomlig i samband med datainsamlingen, men den bör därefter begränsas så långt som möjligt och ske i strängt reglerade former, vare sig den sker i stordator-, UNIX- eller PC-miljö. Relativt enkla, lättbegripliga säkerhetsåtgärder torde i alla typer av system vara de mest trovärdiga och de som det är lättast att få lojal uppslutning kring hos den berörda personalen.

**R & D Reports** är en för U/ADB och U/STM gemensam publikationsserie, som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna.

**R & D Reports Statistics Sweden** are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers.

Reports published 1993:

- |                  |                                                                                                                           |
|------------------|---------------------------------------------------------------------------------------------------------------------------|
| 1993:1<br>(grön) | Kvalitetsfonden 1991/92 - Projekt, handläggning och utvärdering<br>(Roland Friberg)                                       |
| 1993:2<br>(grön) | Översyn av Undersökningen av lastbilstransporter i Sverige (UVAV)<br>(Bengt Rosén, Mahnaz Zamani)                         |
| 1993:3<br>(grön) | Bortfallsbarometern nr 8 (Mats Bergdahl, Pär Brundell, Jan Hörngren,<br>Håkan Lindén, Peter Lundquist, Monica Rennermalm) |
| 1993:4<br>(gul)  | Guidelines on the Design and Implementation of Statistical Meta-<br>information Systems (Bo Sundgren)                     |
| 1993:5<br>(grön) | Kvalitetsrapporten 1993 - Utveckling av kvaliteten för SCBs<br>statistikproduktion (Jan Eklöf, Per Nilsson)               |
| 1993:6<br>(grön) | Metoder att få konsistens mellan olika skattningar och samtidigt<br>öka precisionen (Sixten Lundström)                    |
| 1993:7<br>(grön) | Variance Estimation in the Swedish Consumer Price Index<br>(Jörgen Dalén, Esbjörn Ohlsson)                                |
| 1993:8<br>(grön) | Quantifying Errors in the Swedish Consumer Price Index<br>(Jörgen Dalén)                                                  |
| 1993:9<br>(gul)  | Ökad tillgänglighet för statistiska data - en studie av några gränssnitt<br>mot statistiska databaser (Mattias Lindberg)  |

Tidigare utgivna **R&D Reports** kan beställas genom Ingvar Andersson, SCB, U/LEDN, 115 81 STOCKHOLM (telefon 08-783 41 47, telefax 08-783 45 99).

**R&D Reports** listed above as well as issues from 1988-92 can - in case they are still in stock - be ordered from Statistics Sweden, att. Ingvar Andersson U/LEDN, S-115 81 STOCKHOLM, SWEDEN (telephone nr +46 08 783 41 47, telefax nr +46 08 783 45 99).