

Samkörning av mikrodata inom och mellan Statistikansvariga myndigheter

- 1 Sammanfattning och förslag
- 2 Problembeskrivning och läget idag
 - 2.1 Behov och efterfrågan
 - 2.2 Tillgängligheten
 - 2.3 Vad säger lagen och varför
 - 2.4 Utlämnandepolicy i Sverige, Norden och övriga länder
 - 2.5 Tekniska och metodologiska lösningar och nya möjligheter
- 3 Förslag till förbättringar
 - 3.1 Fortsatt samarbete kring utlämnande frågor
 - 3.2 Produktionsteknisk samordning
 - 3.3 Förbättrad tillgänglighet och innehållssamordning

Samkörning av mikrodata inom och mellan Statistikansvariga myndigheter

1 Sammanfattning och förslag

En ökad efterfrågan på samkörning av olika mikrodata i kombination med den ökande användningen av Internet för informationsutbyte innebär stora utmaningar för de statistikansvariga myndigheterna. Hur skall man möta användarnas förväntningar på ökad tillgänglighet till mikrodata och samkörning av mikrodata, samtidigt som man bibehåller uppgiftslämnarnas och allmänhetens förtroende för att uppgifter om enskilda inte blir tillgängliga för obehöriga eller missbrukas. Det kan behövas ökat samarbete när det gäller i) rutiner för skadeprövning och information rörande utlämnande av mikrodata, ii) produktionstekniskt samarbete samt iii) spridningen av information om statistiken och dess innehåll via Internet.

Samordning av utlämnanderutiner

Den arbetsgrupp, som Rådet för den officiella statistiken har tillsatt med representanter från olika myndigheter, bör fortsätta sitt arbete med frågor som rör samkörning och utlämnande av mikrodata. En specifik fråga är hur uppgiftslämnarna och allmänheten skall informeras om hur statistikuppgifter skyddas mot missbruk. Det också viktigt att ta del av den ansats för utlämnande av data via Internet som implementerats vid Danmarks statistik, och som SCB f.n. utvärderar.

Produktionsteknisk samordning

De förslag, som Lars Olsson lägger fram i sin utredning "Samverkan på det produktionstekniska området mellan de statistikansvariga myndigheterna" (2002-11-20), är viktiga även för att förbättra användarnas tillgänglighet till samkörda mikrodata. Inom ramen för det organiserade erfarenhetsutbyte, som Olsson föreslår, skulle SCB kunna ta ett särskilt ansvar för att följa och sprida information om den internationella utvecklingen. Det gäller, förutom tekniken för lagring och utbyte av data, även modeller för beskrivning av statistiken, klassifikationsdatabaser mm., dvs. modeller för metadata. Av särskilt intresse för samkörning och utlämnande av mikrodata är också s.k. statistiska brandväggar och andra metoder för att skydda data mot obehörig användning

Förbättrad tillgänglighet

Avgörande för att en användare skall kunna efterfråga och nyttja samkörda mikrodata är att han eller hon känner till vilka material som finns och vad de innehåller.

Vi föreslår att Rådet för den officiella statistiken uppdrar åt arbetsgruppen för webbpublicering att utreda hur man skulle kunna skapa bättre överblick över vilka data som finns och kan användas för mikrodata-baserade analyser.

Den ur samkörningssynpunkt ideala situationen med fullständig samordning av statistikens innehåll är varken möjlig eller alltid önskvärd. För dem som skall använda samkörda material är det dock nödvändigt att materialen är dokumenterade på sådant sätt att det klart framgår i vilken utsträckning materialen skiljer sig åt. I *Beskrivningen av statistiken* som skall finnas för all officiell statistik, tillgänglig via Internet, bör alla sådana uppgifter ingå och kontinuerligt uppdateras.

2 Problembeskrivning och läget idag

2.1 Behov och efterfrågan

De tekniska möjligheterna att göra statistiska data tillgängliga för samkörningar och vidarebearbetningar har ökat dramatiskt under senare år, inte minst genom Internet-utvecklingen. Därigenom har nyttan av samhällets investeringar i officiell statistik ökat kraftigt.

Under de senaste decennierna har forskare allt ivrigare efterfrågat samkörningar av statistiska individdata för att kunna genomföra olika analyser. SCB mötte i början denna efterfrågan genom att tillhandahålla forskararbetsplatser inom SCB:s lokaler eller genom att tillhandahålla data aggregerade på mycket låg nivå. SCB och andra statistikansvariga myndigheter har därutöver under senare år - efter skadeprövning i varje enskilt fall - i ökad utsträckning utlämnat avidentifierade individdata för olika statistik- och forskningsändamål.

Under 1990-talet ökade efterfrågan på samkörning av mikrodata dramatiskt, både när det gäller kvantitet och innehållsmässigt. Forskningen har blivit alltmer tvärvetenskaplig och utredningar alltmer sektorsövergripande. Ökad datorkapacitet och sofistikerade programvaror har möjliggjort analyser av allt större datamängder på mikronivå.

Olika producenter av mikrodata

För sektorsövergripande analyser krävs ofta statistiska data producerade inom olika organisationer och lagrade på olika sätt. Detta ställer särskilda krav på samordning och dokumentation; inte minst gäller detta då material samkörs på mikronivå och forskaren kanske önskar härleda nya variabler selekteringar.

Longitudinella analyser

För att få djupare kunskap om dynamiken och samspelet mellan olika processer, t.ex. familjebildning och yrkeskarriär, behövs ofta multivariata longitudinella analyser. Detta kan i sin tur kräva samkörningar av olika årgångar av statistikmaterial. För tidigare insamlade data kan det också finnas behov av egisteruppföljningar på individnivå.

Företagsdata på mikronivå

Förutom samkörning av individdata har det även under det senaste åren tillkommit en ökande efterfrågan på att få tillgång till mikrodata avseende företag/arbetsställen. Av särskilt intresse vid studier av företagens demografi är tillgång till samkörda mikrodata med kopplingar mellan individer och företag/arbetsställen. Denna typ av mikrodata efterfrågas inte minst av utländska forskare. Utlämnandet av företagsuppgifter på mikronivå kräver dock särskild eftertanke och utlämnade till annat land är omgärdat av särskilda restriktioner.

Externa forskardatabaser

Ett flertal forskningsinstitut önskar bygga upp egna, uppdateringsbara mikrodatabaser för olika, mer eller mindre specificerade ändamål. Önskemålen kan till viss del tillgodoses genom s.k. nyckeldatabaser, som beskrivs nedan. Ju fler sådana databaser som skapas utanför de statistikansvariga myndigheterna, desto svårare blir det att hålla reda på var olika mikromaterial finns och hur de används. Trots att det inte finns något känt fall på missbruk av mikrodata som utlämnats för forskning, så ökar risken eller åtminstone farhågorna för missbruk ju mer data sprids. Det kan därför finnas anledning att försöka möta forskarnas behov av att bearbeta samkörda mikrodata på annat sätt än genom att lämna ut hela databaser.

2.2 Tillgängligheten

Statistikansvariga myndigheter har inte helt lyckats tillgodose användarnas olika önskemål när det gäller samkörning av mikrodata. Kritiken från forskare och utredare har främst gällt bristande information och dokumentation men även krångliga administrativa rutiner. Dessutom innebär naturligtvis skadeprövningen, att forskarna inte kan få alla sina önskemål uppfyllda vad gäller variabelmängd och detaljeringsgrad.

Bristande överblickbarhet

Ett första krav för att data skall vara tillgängliga är att användaren vet vilka datamaterial som finns. Decentraliseringen av statistikansvaret på olika SAM och inom SCB på olika program medför att en statistikanvändare har svårt att få en samlad bild över helheten.

För att anknyta till en aktuell politisk fråga kan vi exemplifiera med de ökande sjukskrivningskostnaderna i Sverige. Det finns en mängd statistiska data för att belysa och analysera denna fråga. Ett grundmaterial är RFV:s statistik över utbetalda sjukersättningar. Ett kompletterande material är RFV:s statistik över de korta sjukskrivningarna (med arbetsgivarersättning) som produceras av SCB. Genom att samköra dessa data med Yrkesregistret (och Dödsorsaksregistret) kan man studera yrkessjuklighet (och yrkesdödlighet). Registret över totalbefolkningen (RTB) ger underlag för att studera den demografiska utvecklingen. Genom SCB:s Arbetskraftsundersökningar (AKU) och AV:s Arbetsmiljöundersökningar kan man analysera samband mellan sjukskrivningar och arbetsförhållanden.

IP/Led, Ingrid Lyberg
Staben, Birgitta Pettersson
VL, Bo Sundgren

Undersökningen om hushållens ekonomi (HEK) samkört med AMS:s Händeldatabas ger underlag för att studera individernas och hushållens försörjningssituation; t.ex. arbetsinkomster kontra olika former av bidrag. Undersökningarna om levnadsförhållanden (ULF) med data från 1975 ger underlag för att studera utvecklingen av faktisk och upplevd ohälsa. Integrationsregistret LOUISE med uppgifter om individernas sysselsättning (härledd via kontrolluppgifter och inkomster), utbildning mm ger underlag för att studera regionala skillnader, skillnader mellan invandrare och infödda svenskar etc.

På liknande sätt finns det en mängd material för att beskriva och analysera en annan aktuell politisk fråga; invandrarnas situation i Sverige.

I dagsläget är det mycket svårt för en forskare eller utredare att få en överblick över de statistiska material som kan vara relevanta för aktuell frågeställning. En tendens är att man koncentrerar sig på det material man råkar känna till bäst.

Ofullständig dokumentation

Nästa förutsättning för att data skall kunna användas för samkörning är att materialen är dokumenterade på ett tillfredsställande sätt och så långt som möjligt innehållsmässigt samordnade. Materialen måste naturligtvis innehålla en kopplingsvariabel att basera samkörningen på (t.ex. personnummer, organisationsnummer för företag o.dyl.). Goda innehållsbeskrivningar är nödvändiga för att användaren skall kunna bedöma om det överhuvudtaget är meningsfullt att samköra olika datamaterial, och för att man skall kunna uttolka resultaten av samkörningen på ett korrekt sätt.

Idealt – ur samkörningssynpunkt - är datamaterialen fullständigt samordnade när det gäller objekt, variabeldefinitioner, referenstidpunkter mm.. Eftersom de olika materialen är framställda för olika ändamål är detta sällan uppfyllt. Om detta inte framgår av dokumentationen kan användaren få till synes oförklarligt motstridande resultat i sina analyser. Exempelvis definieras deltidssjukskrivna som ”i arbete” i AKU men ingår som sjukskrivna i RFV:s statistik. P.g.a. av detta har skillnaderna mellan RFV:s och AKU:s uppgifter ökat över tiden.

Externa användare, som önskat samköra datamaterial på mikronivå, har ibland klagat på de administrativa rutinerna kring själva skadeprövningen och myndigheternas olika debiteringsprinciper. För att förbättra för användarna i dessa hänseenden har en särskild arbetsgrupp bildats inom Rådet för den officiella statistiken under ledning av SCB:s chefsjurist.

2.2 Vad säger lagen och varför?

När forskare m.fl. vill ha ut mikrodata från flera myndigheter måste en prövning av utlämnandet i enlighet med sekretesslagens regler göras av varje

IP/Led, Ingrid Lyberg
Staben, Birgitta Pettersson
VL, Bo Sundgren

berörd myndighet. Vid utlämnande och sambearbetning av individdata måste också personuppgiftslagens regler beaktas. Några särskilda bestämmelser som reglerar sambearbetning av data om juridiska personer finns inte.

Sekretsslagen och statistiksekretess

Den starka sekretess som gäller i en myndighets statistikverksamhet har tillkommit efter en avvägning mellan intresset att statistiken skall komma till användning för ökad kunskap om samhället å ena sidan och sekretessintresset å den andra. Det starka sekretesskyddet är viktigt för allmänhetens förtroende för SCB och andra statistikansvariga myndigheter och är också en förutsättning för statistikverksamheten. Ett minskat förtroende kan påverka viljan att lämna uppgifter negativt och försämra uppgifternas kvalitet.

Enligt 9 kap. 4 § sekretesslagen (1980:100) gäller sekretess i sådan särskild verksamhet hos myndighet som avser framställning av statistik för uppgift som avser enskilds personliga eller ekonomiska förhållanden och som kan hänföras till den enskilde. Huvudregeln är alltså att uppgifter i statistikverksamhet är hemliga och inte får lämnas ut. Undantag från huvudregeln har dock gjorts för uppgift som behövs för forsknings- eller statistikändamål och uppgift, som inte genom namn, annan identitetsbeteckning eller därmed jämförbart förhållande är direkt hänförlig till den enskilde. Uppgifter får i dessa fall lämnas ut, om det står klart att uppgiften kan röjas utan att den som uppgiften rör eller honom närstående lider skada eller men. Med enskild avses såväl fysiska som juridiska personer.

Inför ett utlämnande enligt något av undantagen måste en bedömning alltid göras av risken för skada eller men för den som uppgifterna rör eller någon närstående. Det skall stå klart att ett utlämnande kan ske utan risk. Med skada avses ekonomisk skada och med men avses integritetskränkningar av olika slag. En skada kan t.ex. uppstå om en uppgift om ett företags ekonomi lämnas till ett konkurrerande företag. En integritetskränkning kan t.ex. avse att någon blir utsatt för andras missaktning om hans eller hennes personliga förhållanden blir kända. Vid bedömningen av risken för men har både den enskildes upplevelser och de gängse värderingarna i samhället betydelse.

Sekretessen gäller också i förhållandet mellan olika verksamhetsgrenar inom samma myndighet när de är att betrakta som självständiga i förhållande till varandra. Vad som utgör en verksamhetsgren får avgöras från fall till fall.

Personuppgiftslagen

Statistiksekretessen innebär inte någon skillnad mellan fysiska och juridiska personer. Vid utlämnande och/eller sambearbetning av uppgifter om fysiska personer måste dock även personuppgiftslagens regler beaktas.

IP/Led, Ingrid Lyberg
Staben, Birgitta Pettersson
VL, Bo Sundgren

Personuppgiftslagen (1998:204) innehåller vissa grundläggande bestämmelser om behandlingen av personuppgifter; bl.a. skall behandlingen vara laglig och ske på ett korrekt sätt och enligt god sed. Uppgifter som behandlas måste vara adekvata och relevanta. Uppgifterna får behandlas bara om det är nödvändigt för ändamålet med behandlingen. Den registeransvarige får inte behandla fler uppgifter eller bevara dem längre än som är nödvändigt med hänsyn till ändamålet med behandlingen.

Uppgifter får vidare bara samlas in för ”särskilda, uttryckligt angivna och berättigade” ändamål och inte behandlas för något ändamål som är oförenligt med det för vilket uppgifterna samlades in. Enligt personuppgiftslagen är dock behandling av personuppgifter för bl.a. statistiska och vetenskapliga ändamål särskilt gynnad. Senare behandling för sådana ändamål anses inte oförenlig med de ursprungliga ändamål, t.ex. administrativa, för vilka uppgifterna samlades in.

Känsliga uppgifter och samtycke

För behandling av känsliga personuppgifter, d.v.s. uppgifter angående ras eller etniskt ursprung, politiska åsikter, religiös eller filosofisk övertygelse, fackligt medlemskap samt hälsa eller sexualliv, gäller särskilt stränga regler. Behandling av känsliga personuppgifter för forskningsändamål är tillåten efter ett uttryckligt samtycke från den registrerade och även utan sådant samtycke, om behandlingen är nödvändig och behandlingen har godkänts av en forskningsetisk kommitté eller anmälts i förväg till Datainspektionen för s.k. förhandskontroll.

I förordningen (2001:100) om den officiella statistiken finns särskilda regler som anger i vilken utsträckning som känsliga uppgifter får behandlas för framställning av officiell statistik inom respektive statistikområde. Känsliga uppgifter, som får behandlas för framställning av officiell statistik, får även användas för annan statistik och forskning. Enligt personuppgiftslagen får känsliga personuppgifter alltid behandlas med stöd av ett uttryckligt samtycke från den registrerade.

Utlämnande och sekretessprövning

Innan uppgifter lämnas ut måste en sekretessprövning göras. Det ankommer på den myndighet som förvarar uppgifterna att pröva om förutsättningarna för att lämna ut uppgifter är uppfyllda.

När det gäller utlämnande av uppgifter mellan statistikansvariga myndigheter för framställning av officiell statistik eller annan statistik, föreligger normalt inte några hinder ur sekretessynpunkt. Alla statistikansvariga myndigheter torde ha en särskild verksamhet för framställning av statistik. Statistikansvariga myndigheter har vidare rätt att från annan statlig myndighet få tillgång till de data som behövs för den officiella statistik som myndigheten ansvarar för.

När det gäller utlämnande till en myndighet för forskning så innebär gällande regler att statistiksekretessen i regel följer med till den mottagande myndigheten, se 13 kap. 3 § sekretesslagen. Vid utlämnande för utredningsändamål måste särskilt utredas vilken sekretess uppgifterna kommer att omfattas av. Lämnas uppgifter ut till enskilda organisationer och institutioner kan utlämnandet ske med ett särskilt förbehåll om sekretess. Omfattar utlämnandet/samkörningen personuppgifter måste man även kontrollera, att behandlingen är förenlig med personuppgiftslagen.

Särskilda regler för utlämnade till utlandet

Vid utlämnandet av uppgifter till utlandet måste också beaktas att det är förbjudet att lämna ut personuppgifter till tredje land, d.v.s. land utanför EU, som inte har en skyddsnivå som motsvarar vad som gäller inom EU. USA har t.ex. inte någon sådan skyddsnivå. Är det aktuellt att lämna ut uppgifter till tredje land bör man först kontrollera med Datainspektionen vad som gäller för det aktuella landet.

Longitudinella databaser - nyckeldatabaser

Enligt lagen (2001:99) om den officiella statistiken, som trädde ikraft i april 2001, är det möjligt att för forskning och statistik lämna ut data med löpnummer, som via en nyckel hos en statistikansvarig myndighet kan kopplas till personnummer/organisationsnummer. Sådana uppgifter får lämnas ut, om mottagaren har ett särskilt behov av att senare kunna komplettera materialet med t.ex. nya årgångar. Det är endast en statistikansvarig myndighet som har denna möjlighet att lämna ut s.k. nyckeldatabaser. När flera statistikansvariga myndigheter är inblandade i ett utlämnandeärende får man beroende på omständigheterna i ärendet avgöra vilken myndighet som skall spara nyckeln. För närvarande är det framför allt SCB som har bevarat nyckeln vid utlämnanden som berört flera myndigheter, men även Socialstyrelsen har i några fall bevarat en nyckel. Normalt torde det inte innebära några problem att komma överens om vilken myndighet som skall bevara nyckeln. I praktiken torde det bli den myndighet vars uppgifter skall uppdateras.

2.4 Utlämnandepolicy i Sverige, Norden och övriga länder

En nordisk arbetsgrupp har under 2002 arbetat med att utreda frågor som rör utlämnande av mikrodata. Arbetsgruppen har jämfört lagstiftningen i bl.a. de nordiska länderna. Man har också studerat praxis hos de statistikansvariga myndigheterna. Arbetsgruppens rapport visar att lagstiftningen i de nordiska länderna inte skiljer sig åt i någon större utsträckning. Avidentifierade mikrodata lämnas ut till bl.a. forskare för specificerade ändamål. De övriga nordiska länderna har dock till skillnad mot Sverige möjlighet att förena även utlämnande till andra myndigheter med villkor, t.ex. om vilka personer

IP/Led, Ingrid Lyberg
Staben, Birgitta Pettersson
VL, Bo Sundgren

som får använda data, hur länge data får bevaras, m.m. I Sverige är detta möjligt endast vid utlämnande till en enskild organisation

När det gäller metoderna för utlämnande så skiljer sig detta åt i de nordiska länderna. Sverige och Norge och till viss del Finland, lämnar ut data på CD eller motsvarande till den som begärt uppgifterna. I Danmark har statistikbyrån tidigare inte lämnat ut data utanför myndigheten, utan forskare som vill ta del av mikrodata har fått komma till statistikbyrån, där särskilda forskarum finns. Sedan 2001 har Danmarks statistik emellertid utvecklat ett system, som innebär att forskare kan få tillgång till data via Internet. Finland är restriktivt när det gäller utlämnande av data, framför allt företagsdata.

Frågor som gäller utlämnande av mikrodata till forskare diskuteras även inom EU och i andra länder.

Danska modellen

Som ovan nämnts har Danmarks Statistik infört ett system som innebär att forskare och andra kan få tillgång till vissa avidentifierade mikrodata via Internet. Det är dock inte möjligt att få tillgång till särskilt känsliga data, t.ex. uppgifter om brott och i fråga om företag vissa verksamhetsdata, via Internet. Forskare som vill ha tillgång till känsliga data är hänvisade till forskararbetsplaster inom Danmarks statistik.

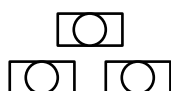
Internetåtkomst till data sker över en krypterad linje, säkrad för obehörigt intrång. Användarna kan inte överföra mikrodata till sin egen dator, utan bearbetningarna sker i servern på Danmarks statistik. Resultaten från bearbetningarna skickas till användarna med e-post som registreras och sedan kontrolleras stickprovsmässigt. Denna kontroll anses tillräcklig för att garantera att forskare inte tar ut mer detaljerade uppgifter än vad som är tillåtet.

Tillgång till data medges för specifika projekt. Det är 32 forsknings- och analysverksamheter som fått medgivande att använda Internet. Till november 2002 är åtkomst medgivits till 88 projekt. Utländska forskningsinstitutioner har inte fått tillstånd .

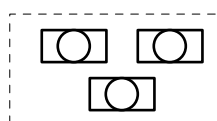
2.5 Tekniska och metodologiska lösningar och nya möjligheter

Modeller för datalagring

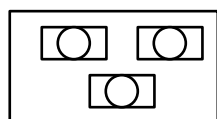
Samkörning av statistiska datamaterial kan tekniskt sett ske på flera sätt:



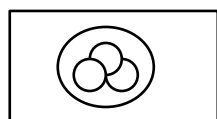
Materialen är fysiskt åtskilda och ej samordnade



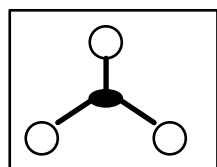
Materialen är fysiskt åtskilda men tekniskt samordnade



Materialen är fysiskt samlade i en gemensam databas



Integrationsregister



Datalager (data warehouse)

1. Datamaterialen ligger fysiskt åtskilda inom en och samma myndighet eller hos olika myndigheter. Vid varje samkörning krävs mer eller mindre omfattande "handpåläggning".

2. Datamaterialen är fysiskt åtskilda men förberedda för samkörningar; de är enhetligt beskrivna och lagrade, och det finns färdiga rutiner för samkörningarna.

3. Datamaterialen ligger i en och samma fysiska databas och är därmed ur teknisk synpunkt enhetligt beskrivna och lagrade. Databasen hanteras av någon databashanterare av standardtyp (relationsdatabas med SQL-gränssnitt) och samkörningar kan därmed göras på ett enkelt och flexibelt sätt. Ett exempel på denna ansats är den planerade kommundatabasen.

4. Datamaterialen har i vissa delar integrerats genom att man har framställt ett s.k. integrationsregister för vanligt förekommande samkörningsbehov. Integrationen kan kräva ett omfattande manuellt kvalitetsarbete för att eliminera inkonsistenser i registret. Ett exempel på en integrationsdatabas är det longitudinella registret LOUISE med individuppgifter hämtade från sysselsättningsregistret, utbildningsregistret, inkomstregistret m.m..

5. Datamaterialen ligger i ett s.k. datalager (data warehouse) som är organiserat enligt vissa standardiserade principer, och som är lätt tillgängligt för användarna med hjälp av särskilda standardprogramvaror: databashanterare och s.k. OLAP-verktyg¹. I ett datalager finns vanligen både mikrodata och makrodata. Frekvent efterfrågade aggregeringar kan därigenom erhållas mycket snabbt, samtidigt som mikrodata ger full flexibilitet i uttagen. I Sverige har Riksförsäkringsverket (RFV) utvecklat ett statistiskt datalager (STORE) enligt dessa principer, baserat på programvaror från ORACLE Corporation.

¹ OLAP = On-Line Analytical Processing.

IP/Led, Ingrid Lyberg
Staben, Birgitta Pettersson
VL, Bo Sundgren

Alla varianter av samkörningar kräver tillgång till databeskrivningar (metadata), såväl tekniska metadata som innehållsorienterade. Standardiserade tekniska metadata kan utnyttjas av standardprogramvaror och möjliggör en mer eller mindre långtgående automatisering av samkörningarna. I alternativen 2-5 har man investerat olika långt i samordning av materialen vilket möjliggör större eller mindre flexibilitet för automatiska uttag.

Kommersiella programvaror som databashanterare och OLAP-verktyg har numera ett bra stöd för tekniska metadata. Däremot saknar de i stort sett färdiga lösningar för att hantera de innehållsorienterade metadata, som är så viktiga i statistiksammanhang. Statistikbyråer, internationella organisationer, universitet m.fl. bedriver dock ett omfattande utvecklings- och standardiseringsarbete inom området statistiska metadata, bl.a. i form av ett antal EU-projekt: MetaNet², METAWARE³, COSMOS⁴, m.fl.

Säkerhetslösningar

En annan utmaning består i att hantera de säkerhets- och sekretessproblem som i och för sig alltid har funnits i samband med statistikproduktion, men som givetvis blir mera komplicerade i och med den ökade tillgången på lättillgänglig statistik och annan information.

Sedan flera årtionden tillbaka har forskare intresserat sig för den s.k. röjandeproblematiken, d.v.s. möjligheterna att ur anonyma statistiska data (såväl mikrodata som makrodata) härleda känslig information om enskilda individer och företag. Resultaten av all denna forskning kan synas nedslående. Sammanfattningsvis innebär de, att man egentligen aldrig kan vara helt säker på att inte någon – med tillgång till lämplig bakgrundsinformation och goda tekniska resurser – kan åstadkomma sådana röjanden. Detta skulle i sin tur kunna innebära allvarliga skador för såväl berörda individer och företag som för de ansvariga för den officiella statistiken, vilka är starkt beroende av uppgiftslämnarnas förtroende.

Finns det då något sätt att möta dessa hot? Det finns, men det räcker inte med att använda sig av enbart teknik och metoder. Man måste kombinera denna typ av åtgärder med lagstiftning och administrativa rutiner. I Sverige är det ju t.ex. förbjudet att försöka röja uppgifter om enskilda genom att bearbeta och kombinera officiell statistik. Till förbudsregeln är kopplad en straffbestämmelse.

² MetaNet är ett s.k. "network of excellence", där ledande experter på statistiska metadata försöker beskriva den samlade kunskapen på området, jämför olika teorier och praktiska ansatser samt försöker komma fram till vissa rekommendationer.

³ METAWARE är ett projekt där man undersöker hur s.k. datalager (data warehouses) skulle kunna förse med den metadatahantering som behövs i officiell statistikproduktion.

⁴ COSMOS är ett projekt som samordnar andra metadataprojekt.

IP/Led, Ingrid Lyberg
Staben, Birgitta Pettersson
VL, Bo Sundgren

När det gäller den rent tekniska säkerheten bör statistikansvariga följa den allmänna utvecklingen på området och anamma de standardslösningar som successivt tas fram. Det handlar bl.a. om kryptering, brandväggar, certifikat för identifiering av behöriga användare m.m.⁵ Helt nyligen har Statskontoret presenterat ett förslag, som är av stort intresse i detta sammanhang: "Säker kommunikation mellan svenska myndigheter och med EUs institutioner från 2003" (PM 2002-12-11). Förslaget går i första hand ut på att bygga upp ett svenskt myndighetsnät för säker kommunikation med motsvarande nät på EU-nivån, TESTA. Mer intressant ur statistikproduktionssynpunkt är att nätet även skulle kunna användas för säker och effektiv överföring av stora datamängder mellan svenska myndigheter.⁶

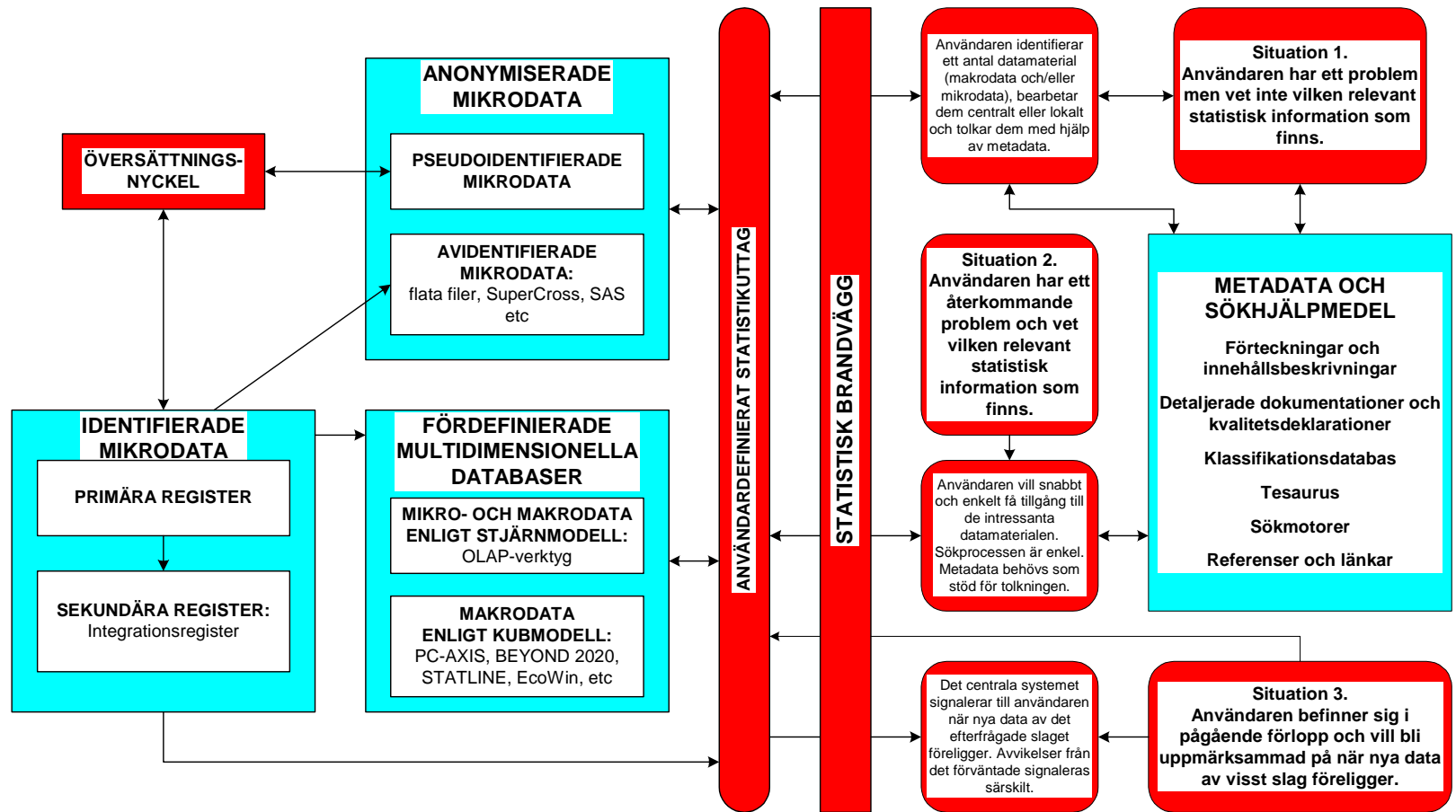
Utöver den rent tekniska säkerheten har statistiken, som ovan framhållits, ett behov av att säkerställa att statistiska data enbart används för statistiska ändamål. I detta sammanhang har begreppet "statistisk brandvägg" lanserats. Begreppets innebörd åskådliggörs i *Figur 1*. Endast statistik tillåts passera genom brandväggen, samtidigt som användarna har stor frihet och flexibilitet när det gäller att från sin egen arbetsplats styra bearbetningar av de olika datamaterial, som fysiskt ligger lagrade innanför brandväggen.

Den ovan beskrivna lösningen vid Danmarks Statistik utgör en mycket intressant implementering av denna modell för att tillgängliggöra mikrodata för forskare m.fl.

Utöver de metoder för tillgängliggörande av mikrodata som berörts här förekommer i utlandet, främst i USA, s.k. public files, avidentifierade och sekretessgranskade mikrodata som tillgängliggörs helt fritt. I ett litet land som Sverige kan det vara svårt att få fram tillräckligt säkra public files, som samtidigt inte är så grova, vad gäller upplösningen, att data blir ointressanta för forskare och andra kvalificerade användare.

⁵ Denna utveckling sker i huvudsak internationellt och på kommersiell basis. Standardiseringsorganen har också en viss roll. I Sverige har man bl.a. inom statsförvaltningen tagit vissa gemensamma initiativ såsom utvecklingen av ett generellt spridnings- och hämtningssystem (SHS) och framtagandet av s.k. medborgarcertifikat i samarbete med bankerna.

⁶ Det föreslagna svenska myndighetsnätet är, liksom EU:s TESTA, helt fristående från Internet. Detta är möjligen en fördel ur säkerhetssynpunkt, men det innebär tyvärr också många nackdelar i form av kostnader för dubblerad kompetens, utrustning, programvaror m.m. Alla användare av nätet får en ny brevlåda att hålla reda på vid sidan av Internet-baserad e-mail. Det nya nätet bör därför inte i första hand användas i sammanhang, där väl fungerande Internet-lösningar redan finns, utan inriktas på situationer där nuvarande lösningar inte är tillräckligt säkra eller effektiva, t.ex. överföring av stora mängder mikrodata mellan administrativa myndigheter och statistikansvariga myndigheter samt mellan statistikansvariga myndigheter.



Figur 1. Datalager i kombination med statistisk brandvägg och metadata för olika typer av användningar.

3 Förslag till förbättringar

3.1 Fortsatt samarbete kring utlämnandeåfrågor

När forskare vill ha tillgång till data från flera myndigheter är det väsentligt att berörda myndigheter samverkar. För att berörda myndigheter skall kunna pröva utlämnandet av de uppgifter myndigheten ansvarar för, måste ansökan finnas hos samtliga berörda myndigheter. En forskare bör dock inte behöva bli slussad mellan myndigheterna, utan *en* myndighet bör kunna vara ansvarig för att samordna utlämnandet. En dialog mellan myndigheterna bör vidare föras så att liknande bedömningar görs i utlämnandeåfrågan. Var samkörningen skall äga rum bör kunna lösas från fall till fall beroende på vad som är mest lämpligt så länge bearbetningen sker i enlighet med reglerna i personuppgiftslagen eller annan registerförfattning.

Rådet för den officiella statistiken har tillsatt en arbetsgrupp med representanter från Arbetsmiljöverket, Brottsförebyggande rådet, Centrala studiestödsnämnden, Finansinspektionen, Fiskeriverket, Institutet för tillväxtpolitiska studier, Socialstyrelsen, Riksförsäkringsverket och SCB. Gruppen diskuterar bl.a. hur man skall väga användarbehoven och den nytta, som forskning kan ge, mot kraven på sekretess och integritetsskydd. Vidare diskuteras frågor om samordning av utlämnandeårenden, som berör flera myndigheter, och rutinerna för hur olika årenden bör hanteras. Denna grupp bör fortsätta sitt arbete med frågor som rör samkörning och utlämnande av mikrodata. En specifik fråga är hur uppgiftslämnare och allmänheten skall informeras.

Utlämnande via Internet

Inom SCB har tillsatts en arbetsgrupp som skall utreda möjligheterna att i enlighet med den danska modellen lämna ut mikrodata via Internet. Den danska modellen har den fördelen att myndigheten fortfarande har kontroll över de mikrodata, som görs tillgängliga, samtidigt som forskarna kan sitta på sina arbetsplatser och arbeta med mikrodata. Vid ett utlämnande på de sätt som görs i Sverige och Norge får myndigheterna svårt att kontrollera hur data används och att eventuella förbehåll följs. Det finns t.ex. risk för att uppgifterna sprids till andra användare. Metoden innebär dock större frihet för forskarna.

3.2 Produktionsteknisk samordning

Lars Olssons utredningsrapport pekar på ett stort intresse bland de statistikansvariga myndigheterna för ett ökat erfarenhetsutbyte och samarbete på det produktionstekniska området. Bland de områden som SAM är särskilt intresserade av, nämner Olsson statistiska databaser och datalager (data warehouses) samt metadataåfrågor. Detta är områden som också är av största betydelse för att man på ett effektivt och användarvänligt sätt skall kunna tillgodose de ökande behoven av samkörningar av mikrodata.

Vårt förslag ligger helt i linje med Lars Olssons och innebär att Rådet för den officiella statistiken tar initiativ till ett organiserat samarbete i de nämnda frågorna. Samarbetet skulle till en början bestå i ett erfarenhetsutbyte mellan myndigheterna och en gemensam kunskapsuppbyggnad, gärna med hjälp av internationellt framstående experter. Detta samarbete skulle senare kunna operationaliseras till mera fokuserade samarbetsprojekt på utvalda områden.

Såsom berördes i avsnitt 2.5 ovan finns ett aktivt internationellt samarbete kring statistiska databaser och datalager samt därmed sammanhängande metadatafrågor. Samarbetet bedrivs bl.a. i form av EU-projekt och andra, mer informella samarbetsgrupper. SCB skulle kunna ta ansvar för att följa och informera vidare om detta samarbete och om de resultat som hittills uppnåtts. Lars Olsson pekar bl.a. på att Eurostat lagt fast en standard för utbyte av statistiska data och metadata. Standarden heter GESMES⁷ och är en av FN formellt etablerad EDIFACT-standard. F.n. pågår ett arbete med att överföra GESMES till XML-format.

3.3 Förbättrad tillgänglighet och innehållssamordning

Det största problemet för externa användare, som har behov av samkörda mikrodata, torde vara svårigheterna att få reda på vilka datamaterial som finns, vad de innehåller och vilka möjligheter och begränsningar som föreligger när det gäller att koppla ihop materialen. Ett mål bör vara att användare via Internet på ett enkelt sätt skall kunna få en överblick över vilka material som finns. Vidare skall användaren via Internet få tillgång till sådan dokumentation att han eller hon kan bedöma om materialen är lämpliga för den tänkta analysen och lämpliga att samköra på mikronivå.

När det gäller att förbättra överblicken över vilka material som kan tillgängliggöras på mikronivå föreslås att Rådet för den officiella statistiken uppdrar åt arbetsgruppen för webbpublicering att utreda denna fråga.

När det gäller datamaterialens dokumentation bör utgångspunkten vara de *Beskrivningar av statistiken* som skall finnas tillgänglig via Internet för all officiell statistik. Även annan statistik som kan vara lämplig att återanvända på mikronivå bör dokumenteras enligt denna mall.

Den ideala situation att alla statistikregister har samma objektsavgränsningar, variabeldefinitioner och referensidpunkter varken kan eller bör eftersträvas. Detta förhindrar inte att många material ändå kan sambearbetas såvåda användaren får adekvat information om de olikheter som finns. Den dokumentation som åtföljer officiell statistik bör innehålla all information som kan vara relevant för en användare som funderar på att samutnyttja statistik och mikrodata från olika källor. *Beskrivningarna av statistiken* är för närvarande ofta ofullkomliga i detta avseende. För datamaterial som skall

⁷ GESMES = GEneric Statistical MESsage.

användas för forskning m.m., behöver de översiktliga *Beskrivningarna av statistiken* kompletteras med mera detaljerade beskrivningar av mätinstrument och processer som använts för datainsamlingen.

Statistikansvariga kan också aktivt bidra till att göra statistiken mer jämförbar. Detta innebär inte nödvändigtvis att alla variabler skall vara definierade på samma sätt i all statistik – olika definitioner kan vara nödvändiga för att tillgodo se olika behov – men det skall som nämnts vara lätt att få reda på vari olikheterna består, och helst skall närbesläktade begrepp bestå av ”byggstenar” som är enhetligt definierade.

När det gäller samordningen av innehållet för olika statistikmaterial kan man sträcka sig olika långt när det gäller ambitionsnivån; från endast standardisering av ett begränsat antal kopplingvariabler till fullständig indataharmonisering.

Ju längre man sträcker sig i innehållssamordningen, desto mer blir det en fråga om att kompromissa mellan olika användarbehov, vilket ligger utanför ramen för detta diskussionsunderlag.

Ambitionsnivåer när det gäller innehållsamordning av statistik

<i>Ambitionsnivå</i>	<i>Exempel</i>
Standardiserade metadata	<p>Inom området dokumentation och kvalitetsdeklarationer har Eurostat kommit med vissa riktlinjer, och inom den ekonomiska statistiken är IMF mycket aktivt vad gäller metadata för tidsserier. Den s.k. Neuchâtel-gruppen har föreslagit en enhetlig metadatamodell för klassifikationer, och ett antal statistikbyråer håller nu på att implementera klassifikationsdatabaser baserade på denna modell.</p> <p>Inom SOS-systemet kan <i>Beskrivning av statistiken</i> ses som en standard för metadata som översiktligt beskriver statistikens innehåll och kvalitet.</p>
Standardiserade kopplingvariabler	<p>Här gäller hur t.ex. personnummer, län-kommun-koder, organisationsnummer för företag mm. skall vara representerade i datalager. Inom SCB finns riktlinjer för detta, vilka möjligen också kan vara till gagn för andra. Denna fråga kan diskuteras inom ramen för det produktionstekniska samarbete, som föreslås i Lars Olssons utredning.</p>
Standardiserade klassifikationer	<p>Det finns redan standarder för många klassifikationer, t.ex. näringsgren (SNI), utbildning (SUN), socioekonomisk gruppering (SEI), yrke (SSYK). SCB har ett övergripande ansvar för klassificeringar som används inom den officiella statistiken och publicerar dessa standarder i Meddelande i samordningsfrågor och på Internet. Det kan diskuteras om det behövs fler "officiella" klassifieringsstandarder, och hur dessa skall spridas på Internet.</p>
Standardiserade variabeldefinitioner	<p>Genom att systematiskt samla variabeldefinitioner i en variabeldatabas får man en överblick över eventuella brister i samordningen av variabler mellan olika statistikprodukter. Samordning av variabler är ett mödosamt arbete och bör därför i första hand inriktas på särskilt viktiga variabler, t.ex. kopplingsvariabler, d.v.s. identitetsbeteckningar och koder som används för att koppla samman olika datamaterial med varandra.</p>
Indataharmonisering (standardisering av enkätvariabler)	<p>I sina harmoniseringssträvanden försöker Eurostat ofta ta fram förpliktigande riktlinjer för inte bara hur variabler <i>definieras</i> utan även för hur uppgifterna skall <i>inhämtas</i>. Det kan gälla urvalsförfarande, insamlingsförfarande, svarsalternativ och frågeordning m.m. Detta är inte lika aktuellt inom det svenska statistiksystemet, där en stor del av statistiken baseras på administrativa material, som statistikansvariga ändå inte har stora möjligheter att påverka, vare sig vad gäller definitioner eller mätmetoder.</p> <p>En viss form av indataharmonisering på detta område kan sägas förekomma inom SOS-systemet genom det samarbete som sker inom statistisk metod och mätteknik.</p>