FI 53

muistio

Tilastokeskus

Tekijä	Päiväys	N:o
Reino Hjerppe	9.10.1977	46

TIETOKANTAMENETELMISTÄ

YHTEISKUNTATILASTOJEN

TUOTANNOSSA

1277031415—12/443

Esipuhe

Seuraavassa pyritään luomaan katsaus eräisiin tietokanta-menetelmiä koskeviin kehittämisnäkymiin yhteiskuntatilastojen kannalta. Tätä asiaahan on käsitelty kansainvälisellä tasolla viime aikoina useissakin yhteyksissä. Seuraavat näkemykset perustuvat paljolti niihin raportteihin, joita tältä suunnalta on ilmestynyt. Mainittavimpina ovat ns. Bratislavan paperi "Computing in the National Statistical Services beyond 1980", CES:n atk-työryhmän raportti "Report on the Use and Future Needs for Data Base Management in National Statistical Services" 26.1.1977 sekä tutustuminen Ruotsissa tapahtuvaan tietokanta-menetelmien kehittämistyöhön keväällä 1977. Keskustelut erityisesti Tuula Nurmisen kanssa ovat olleet myös varsin hyödyllisiä tätä raporttia ajatellen.

SISÄLLYSLUETTELO

1. Johdanto

Tietokoneilla ja tietojenkäsittelyteknologian kehittymisellä arvellaan olevan erittäin suuria vaikutuksia yhteiskuntatilaston tuotantotapaan tulevaisuudessa.

Keskeisenä käsitteenä ja tavoitteena tietojenkäsittelyn kehittämisessä tulevaisuutta silmällä pitäen on useissa yhteyksissä mainittu kansallisen tietokannan käsite. Kansallinen tietokanta ideaalitapauksessa sisältää ja pitää saatavilla sen tieton, jota yhteiskunta pitää välttämättömänä erilaisia yhteiskunnallisten toimintojen ohjausta suorittaessaan. [1)]

Tällainen tietokannan käsite on looginen eikä fyysinen käsite. Käytännössä kansallinen tietokanta voi koostua useista eri osatietokannoista, joiden välillä vallitsee vain looginen yhteys. Osatietokannat voivat fyysisesti sijaita eri paikoissa tai eri alueilla.

Loogisesti yhdistävänä linkkinä kansallisessa tietokannassa on keskitetysti ylläpidettävät tietoaineistohakemistot ja -sanakirjat sekä aineiston standardisointi yhteensopivaksi kansallisen tietovarannon osaksi.

---

1) Tällainen käsite on keskeisenä lähtökohtana raportissa: Computing in the National Statistical Services beyond 1980.

2. Tietokantojen historiasta lyhyesti

Tietokantateknologian kehittyminen on kiinteässä yhteydessä atk:n kehittämiseen niin hardwaren kuin softwarenkin osalta. Hardwaren osalta suorasaantimuisteja voidaan pitää eräänä tietokantateknologian perusedellytyksenä.

Softwaren osalta puhutaan varsin yleisesti tietokantojen hallintajärjestelmistä (seuraavassa lyhennetty THJ:ksi). Tässä suhteessä eräänä perustavaa laatua olevana työnä voidaan pitää Codasyl komitean Data Base Task Groupin työtä vuodelta 1971, johon varsin usein viitataan tietokannoista puhuttaessa. Tässä yhteydessä ko. työryhmä teki merkittävää työtä käsitteistöjen ja tiedonhallintajärjestelmien standardisoimiseksi.

Coddin kehittämä relationaalinen malli tietokantoihin lupaa nyttemmin huomattavia parannuksia tietokantojen hallintamenetelmiin.

Tällä hetkellä on olemassa useita kaupallisia TJH:ä. Tämän lisäksi eräät johtavat tilastolliset keskusvirastot maailmassa ovat kehittämässä omia nimenomaisesti tilastolliseen käytäntöön soveltuvia tietokantamenetelmiä. Näihin palataan esityksessä myöhemmin.

3. Infologia ja datalogia

Näkyvissä on merkkejä erityisen tietokantateorian kehittymisestä. Eräänä peruslähtökohtana tässä teoriassa on informaation (tiedon) sekä tietoaineiston (datan) käsitteiden erittely. Tämän peruskäsitteistön mukaan tietoaineisto (data) on informaation materiaalinen eli aineellinen esitysmuoto, kun taas itse informaatiokäsite liitetään inhimilliseen tietoon ja tietämykseen ja siis tajuntaan (knowledge).

Tämän erottelun jälkeen voidaankin tietojärjestelmien teoriassa ja tietokantajärjestelmien suunnittelussa puhua infologisista eli käyttäjäorientoituneista ja datalogisista eli tietokoneorientoituneista suunnitteluongelmista.

Infologiassa siis tutkitaan miten käyttäjä muodostaa käsitteitä pyrkiessään saamaan tietoa asiasta, johon hän haluaa ratkaisun. Infologia liittyy siis läheisesti yleiseen käsitteenmuodostusprosessien tutkimukseen inhimillisessä ajattelussa.

Tavallinen informaatiojärjestelmän käyttäjä ei ole sen sijaan useinkaan kiinnostunut itse tietojenkäsittelystä. Tietokoneen käyttö kuitenkin edellyttää, että tietojenkäsittelyn tekninen puoli suunnitellaan ja rakennetaan suurella huolella ja täsmällisyydellä. Tietoaineiston käsittelyyn liittyviä ongelmia voidaankin näin ollen kutsua datalogisiksi ongelmiksi.

---

1) Ks. esim. Bo Sundgren: The Theory of Data Bases, Petrocelli/Charter, New York, 1975.

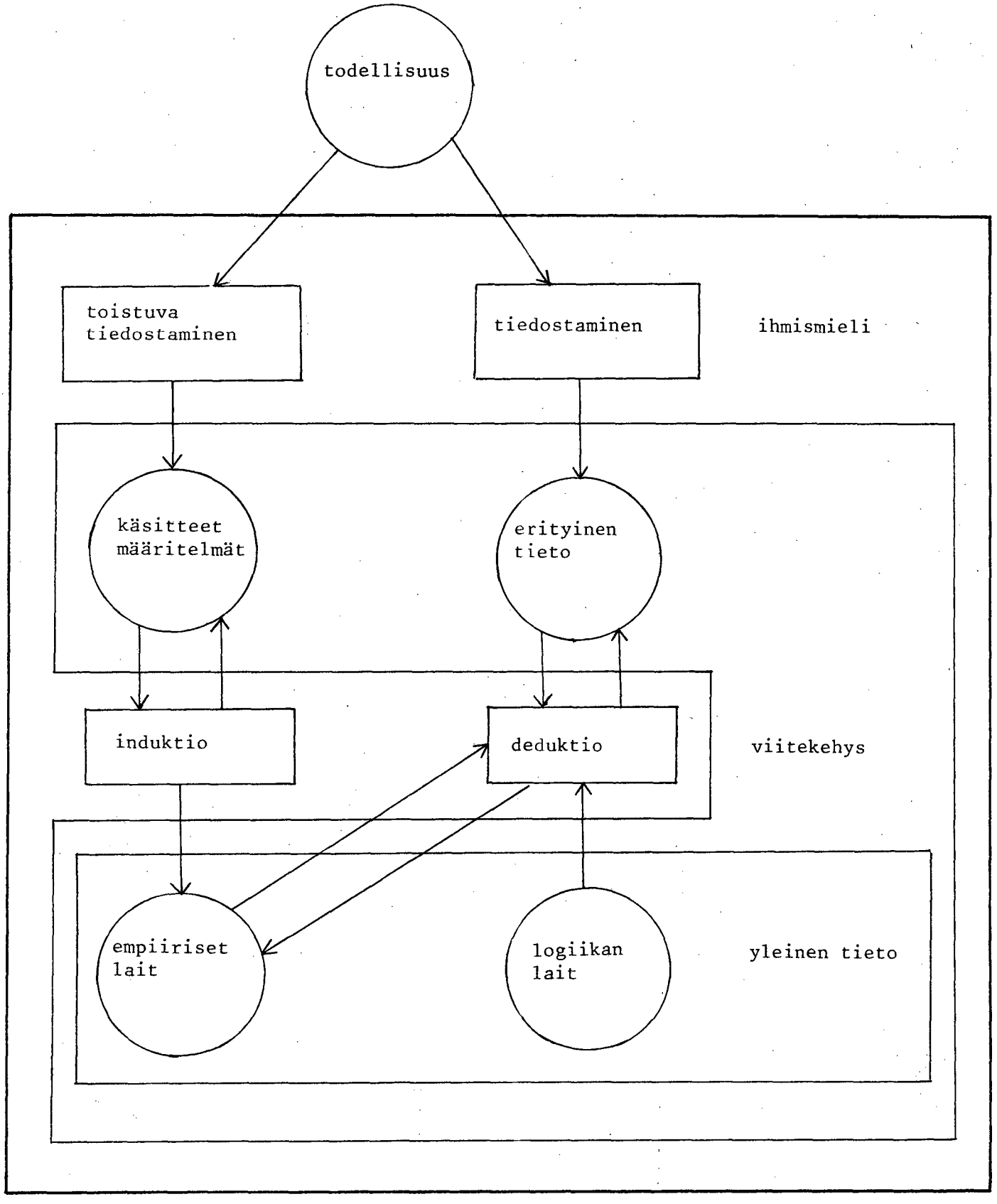
Bo Sundgrenin mukaan infologisia ongelmia voidaan hahmottaa seuraavalla tavalla (ks. oheinen kaavio).

1. Postuloidaan, että hankkiakseen ylipäätänsä tietoa todellisuudesta ihminen muodostaa käsitteitä. Tämä käsitteenmuodostus tapahtuu toistuvan tiedostamisen kautta. Tämä prosessi on jatkuvasti käynnissä koko ihmisen elinajan. Uusia käsitteitä muodostetaan koko ajan, vanhoja muokataan tai unohdetaan.

2. Käsitteenmuodostusta käyttäen ihminen voi muuttaa aistihavaintonsa erityiseksi todellisuutta koskevaksi tiedoksi, joka liittyy erityisiin tapahtumiin ja tilanteisiin.

3. Erityistietonsa varastosta ihminen voi johtaa yleisen tiedon empiirisiä lakeja. Erityistieto saattaa antaa aihetta uusien käsitteiden muodostamiselle, jotka saatetaan määritellä jo olemassaolevien käsitteiden avulla. Johdetut käsitteet saattavat vaikuttaa puolestaan takaisin tiedostamisprosessiin.

4. Päättelyprosessissa ihminen käyttää mm. logiikan lakeja, joiden oletetaan olevan perittyjä ja identtisiä kaikille ihmisille kaikkina ajankohtina. Yhdistämällä logiikan lakeja, empiiristen lakien ja erityistiedon kanssa ihminen voi tehdä johtopäätöksiä uudesta erityisestä tiedosta ja uusista empiirisistä laeista.



Kaavio: Ihmismielen infologinen malli
Bo Sundgren, mt. s. 6.

Infologinen lähestymistapa korostaa eroja

- reaalimaailman ilmiöiden
- reaalimaailman ilmiöitä koskevan informaation ja
- reaalimaailman ilmiöitä koskevan tietoaineistoesityksen (datan) kesken.

Infologiamalliin perustuvassa tietokantateoriassa reaalimaailman ilmiöitä, joista ollaan kiinnostuneita kutsutaan objektijärjestelmäksi eli siis kohdejärjestelmäksi. Teoriaan liittyy runsaasti käsitteitä, joita ei ole tässä yhteydessä mahdollista tarkemmin eritellä.

Perusteiltaan infologista lähestymistapaa tietokantajärjestelmiin lienee pidettävä hyödyllisenä. Traditionaalisiin automaattisen tietojenkäsittelyn teorioihin verrattuna siinä korostuu informaation käyttäjän asema. Häntä pyritään lähestymään selvittämällä, miten ihminen tosiasiassa on tottunut toimimaan pyrkiessään hankkimaan tietoa todellisuudesta. Tämän seikan tutkiminen on erittäin ratkaisevaa informaatiojärjestelmän käyttökelpoisuuden kannalta. Aivan liian usein traditionaalit lähestymistavat ovat keskittyneet datalogisiin ongelmiin ja pyrkineet pikemminkin mukauttamaan käyttäjää kuin mukautumaan käyttäjän tarpeisiin. Käyttäjän mukauttaminen tietojärjestelmän vaatimuksiin on kuitenkin useimmissa tapauksissa tuomittu epäonnistumaan. Nykyisin tämä ongelma lienee varsin hyvin tiedostettu, mutta se ei vielä aina näy konkreettisessa tietojärjestelmän rakentamistyössä. Eräänä syynä tähän voi olla puutteet nimenomaan infologisen puolen tutkimisessa.

4. Tietokannan määrittely

Tietokannan määrittely on tähän saakka ollut varsin epämääräisellä perustalla. On jopa lähdetty siitä, että tietokanta on mikä tahansa korttipakka, jota voidaan käsitellä tietokoneella. Tällainen määrittely ei kuitenkaan ole hedelmällinen.

Tietokannoista voidaan puhua myös ilman tietokonetta. Voidaan ehkä esittää, että sumerilaisten v. 2500 e. Kr. muodostama kirjasto oli maailman ensimmäinen tietokanta. Myös tällöin on kysymys hedelmättömän laveasta määritelmästä. Niinpä tietokantakäsitteet nykyisin liitetään nimenomaisesti automaattiseen tietojenkäsittelyyn tietokoneella.

CES/WP. 9/150 eli Euroopan tilastotieteilijöiden atk-työryhmä on tammikuussa 1977 esittänyt tietokantojen seuraavaa määritelmää:

Tietokanta on kokoelma tai varasto, jossa tietoainesto on

- loogisessa yhteydessä keskenään
- sillä on yhtenäinen määrittely ja kuvaus
- se on strukturoitu määrätyllä tavalla.

Tietokanta on myös reaalimaailman malli ja sellaisenaan sitä tarvitaan useissa erilaisissa käytöissä ja sovellutuksissa.

Tietokannan hallintajärjestelmällä tarkoitetaan software järjestelmää, joka saattaa koostua useista osista, joita voidaan kutakin erikseen käyttää määrättyjen toimintojen suorittamiseen tietokantaa käsiteltäessä. Tietokannan hallintajärjestelmä voidaan siten kuvata järjestelmäksi, joka hallitsee (manage) tietokantaa. Se kontrolloi tietoaineiston tallennusta, hakua ja päivitystä ja toimii välittäjänä tietokannan ja sovellutusohjelman välillä. Tietokannan hallintajärjestelmä on myös vastuussa aineiston varmistuksesta ja integriteetistä [1] ja monista muista tehtävistä, jotka liittyvät tietoaineiston hallintaan ja kotrollointiin.

---

1) Varmistuksella (security) tarkoitetaan sitä, että aineistoa ei käytetä luvattomiin tarkoituksiin.

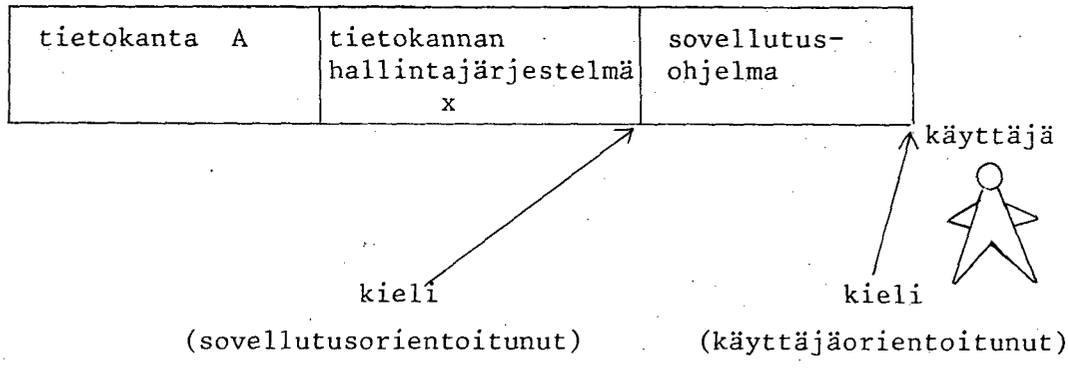
Integriteetillä (integrity) tarkoitetaan sitä, että aineiston tuhoutuminen vahingon kautta torjutaan.

Aineiston päällekkäisyydellä (redundance) tarkoitetaan sitä, että samaa tietoa säilytetään useissa paikoissa.

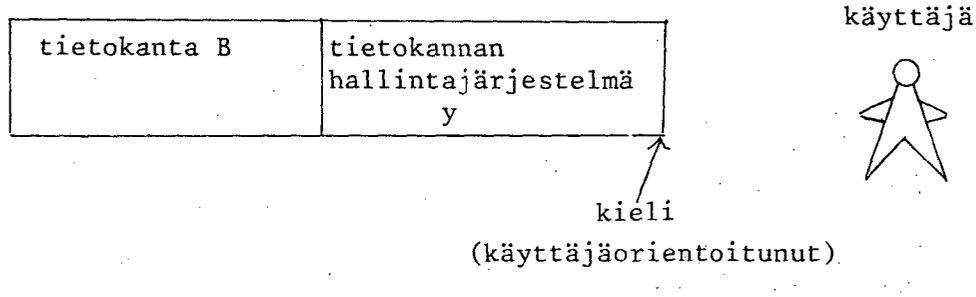
Tietokannan käsittelijä on ohjelmajärjestelmä joka tekee mahdolliseksi tallentaa, ottaa esille, poistaa ja muuttaa tietokannan sisältöä. Tietokannan käsittelijän ja käyttäjän välistä yhdyssidettä kutsutaan tietokantakieleksi.

Tietokantaorientoitunut järjestelmä eli tietokantajärjestelmä koostuu tietokannasta, tietokannan hallintajärjestelmästä ja sovellutusohjelmasta. Joskus voi sovellutusohjelma olla yhteenrakennettu tietokannan hallintajärjestelmän kanssa.

Esim. 1: Sovellutusohjelma ei ole rakennettu sisään tietokannan hallintajärjestelmään [1]



Esim. 2: Sovellutusohjelma on rakennettu yhteen tietokannan käsittelijän kanssa [2]



1) Tässä käytetty hyväksi SCB:n paperia "Databaser och databashanterare vid SCB, L. Olsson etal. 12.5.1977. Esimerkkeinä Ruotsista RAPID, SIMDBM

2) Esimerkkeinä Ruotsista RSDB, ARKDABA

Edellä mainittujen yleisten tietokannan ominaisuuksien lisäksi tilastollisilla tietokantajärjestelmillä on seuraavia ominaisuuksia:

- tilastollisen THJ:n tulee tukea uusien tilastollisten mallien rakentamista ja helpottaa sellaista tietoaineiston analysointia, jota ei mahdollisesti ole tiedetty tai määritelty silloin, kun tietokanta on alunperin luotu

- Tilastollinen tietokanta voi olla hyvin suuri ja suuria aineistomääriä on saatava usein normaaleihin operaatioihin kuten esim. aggregointiin

- Tilastoaineistojen erityisominaisuus on niiden jatkuva kumuloituminen ajassa, jolloin ne muodostavat perustan aikasarjoille ja indikaattoreille. Vaikka useimmiten tarvitaan tuoreimpia aineistoja, vanhemman aineiston tulee olla saatavilla eikä sitä voida hylätä.

- Tilastollinen tietokanta on useiden asiantuntijoiden yhteistyön tulos. Mukana voivat olla esim. aihepiirin asiantuntija, tilastoanalyysispesialisti ja tietokonespesialisti. Heidän kaikkien tulee ymmärtää tietokannan rakenne ja sisältö.

Tilastollisen tietokannan tulee sisältää myös ns. metatietokanta eli informaatiota informaatiosta. Metatietokannan tulee olla tilastollisen tietokannan integroitu osa erityisesti tapauksissa, joissa ulkopuoliset käyttäjät voivat ottaa yhteyden tietokantaan tietojenkäsittelypäätteen avulla.

5. Tietokantojen tarve tilastotyössä

Tutkittaessa tietokantojen soveltuvuutta tilastotuotantoon on selvitettävä mm. seuraavat peruskysymykset:

- ovatko tietokantamenetelmät relevantteja tilastojen tuotannossa?

- voidaanko yhteiskuntatilaston tuotannon tehokkuutta ja tarkoituksenmukaisuutta parantaa tietokantamenetelmillä?

Tietokantamenetelmiä tarvitaan toisaalta tilastollisten lopputuotteiden aikaansaamisessa, toisaalta tilastotoimen sisäisessä työskentelyssä.

Lopputuotteiden tuotannossa tietokantajärjestelmää käytetään antamaan tilaston käyttäjälle nopeampaa ja joustavampaa palvelua. Sen sijaan että pakotetaan käyttäjä pitkään etukäteen määrittelemään, mitä taulukointeja halutaan - mikä johtaa usein siihen, että tilataan enemmän kuin mitä itse asiassa tarvitaan - voi käyttäjä tietokantajärjestelmässä (mahdollisesti päätteen avulla) tehdä suoraan tietyn käyttötarpeen esiintyessä räätälintyönä sen taulun tai ne taulut, joita hän haluaa.

Sisäisessä työskentelyssä tilaston tuottaja voi tietokanta-järjestelmän avulla päästä helposti käsiksi tietoaineistoon. Täten voidaan helpottaa useita valmistusvaiheita esim. tarkistusta ja korjausta, joita tarvitaan pyrittäessä saavuttamaan tietty laatutaso aineistossa.

Käsitykset integroitujen tilastollisten informaatiojärjestelmien luonteesta ovat viime vuosina kehittyneet suhteellisen idealisteista näkemyksistä huomattavasti realistisemmiksi. Eräs tälläinen muutos ajattelutavoissa liittyy siihen kuinka laajasti esim. tietokantamenetelmiä voidaan soveltaa kansallisessa tilastojärjestelmässä. Pienissäkin maissa tilastojärjestelmä kokonaisuudessaan on varsin laaja ja monimutkainen kokonaisuus, minkä vuoksi koko järjestelmää kokonaisuudessaan ei voida kehittää tietokannaksi jakamatta sitä ensin osiin.

Ideana onkin nykyisin rakentaa useita erillisiä tietokanta-järjestelmiä sovellutuskohtaisesti. Näitä sovellutuksia yhdistää toisiinsä yhtenäinen käsitteistö ts. ne ovat loogisesti eivätkä välttämättä fyysisesti integroituja keskenään.

Toisaalta voidaan ajatella, että tällaisia osasysteemitietokantoja voidaan myöhemmin integroida keskenään myös fyysisesti sitä mukaa, kun tekninen tieto ja muu systeemin kehittelytyö sen sallivat. Näin esimerkiksi Ruotsissa on suunnitelmissa yhdistää tulevaisuudessa keskenään RSDB (aluetietokanta) ja TSDB (aikasarjatietokanta).

Edellisestä seuraa myös, että tietokantajärjestelmää ja yhtenäistettyjen tilastonaineistojen järjestelmän käsitteitä ei voida pitää identtisenä. Tähän palataan esityksessä myöhemmin.

Tilastollisista tietokantamenetelmistä ei ole käytettävissä mistään maasta toistaiseksi perusteellisia kustannus-hyötyanalyysejä.

Tietokantamenetelmien hyötyjen sanotaan kuitenkin liittyvän seuraaviin seikkoihin: [1)]

a) aineistojen integrointi (vrt. edellä sanottu)
b) keskitetty aineistojen kontrollointi
c) nopea ja yksinkertaistettu aineiston haku käyttöön
d) aineiston konsistenttisuus
e) aineiston redundanssin (päällekkäisyyden) vähentäminen

Tilastotarkoitusten kannalta tietokantamenetelmien soveltuvuutta voidaan tarkastella kolmelta näkökannalta:

- tilastollisen tietojenkäsittelyn luonteen mukaan
- tilastoaineiston luonteen mukaan
- tilastotoimen organisaation luonteen mukaan

---

1) Report on the Use and Furure Need for Data Base Management in National Statistical services, CES mt.

6. Kustannuksista

Tietokantateknologian kehittäminen ja soveltaminen luonnollisesti vaikuttaa tilastotoimen kustannuksiin. Kustannuksiin vaikuttavina tekijöinä ovat mm.:

- kustannukset jotka aiheutuvat erilaisten tietokantojen hallintajärjestelmien tutkimisesta
- THJ:n hankinnasta aiheutuvat kustannukset
- lisääntyvät laitteistokustannukset. Mahdollisesti tarvitaan lisää keskusmuistikapasiteettia, suorasaantimuisteja, päätteitä ja kommunikointivälineitä
- tietokannan luomisesta aiheutuvat kustannukset ts. kustannukset jotka johtuvat siitä, että olemassa olevat tiedostot muunnetaan sopiviksi valitun THJ:n kanssa
- henkilöstökustannukset, jotka sisältävät mm. koulutuskustannukset sekä aineistohallinnon ja koordinaattorin toimintojen ylläpitokustannukset
- aineiston integriteetin ja varmistusten aiheuttamat kustannukset.

Kustannuksia vähentävinä tekijöinä voivat puolestaan olla mm.:

- vähentyneet ohjelmistojen ylläpitokustannukset
- vähentyneet ohjelmistojen kehittämiskustannukset, jolloin niukkoja ohjelmointiresursseja voidaan säästää vaativimpiin ja kiireellisimpiin tehtäviin
- atk-osasto voi saavuttaa suuremman aineistojen käsittelyvauhdin sekä paremman kyvyn vastata tilastohenkilöiden taholta tulevaan kysyntään.

7. Yleisten tietokantamenetelmien soveltuvuudesa

Yleisiä tietokantamenetelmiä on sovellettu erityisesti organisaatiossa, joissa

- erilaisia aineistoja linkataan keskenään
- aineistoja päivitetään usein
- samaa aineistoa käytetään useissa sovellutuksissa
- sama aineisto useissa tiedostoissa on päällekkäistä
- systeemin kontrollointia varten tarvitaan tilinpitotietoa.

Tutkittaessa yleisten tietokantamenetelmien soveltuvuutta tilastotoimeen on selvitettävä, missä määrin em. ominaisuudet ovat luonteenomaisia tilastolliselle tietojenkäsittelylle.

Tilastollisessa tietojenkäsittelyssä on ilmeisesti yksittäisten tietojen välillä linkkejä esim. sama vastaaja esiintyy tiedostossa useina ajankohtina / eri tiedostoissa. Näiden aineistojen yhdistelyä saattaa estää kuitenkin tietyt yksityisyyden suojaan liittyvät menettelytavat.

Tilastonaineistoja ei useinkaan päivitetä, koska ne ovat yleensä tuloksia erillisistä tiedusteluista. Päivitystä tapahtuukin tavallisesti aineiston tarkistus- ja korjausvaiheissa. Tilastoaineistot tuotetaan usein taulumuotoon, eivätkä ne siten vaadi useinkaan monia sovellutusohjelmia.

Myöskään tilinpitotietojen liittäminen tilastolliseen tietojenkäsittelyprosessiin ei ole olennaista vaikkakin usein toivottavaa.

Niinpä näyttäisikin siltä, että yleisillä tietokantamenetelmillä on tilastotoimessa varsin vähän merkitystä. Tämä näkyy myös siinä, että kansalliset tilastovirastot eivät kovinkaan laajasti käytä yleisiä kaupallisia tietokannan hallintajärjestelmiä. Sen sijaan eräissä maissa suoritetaan kylläkin intensiivistä tutkimusta tietokantamenetelmien kehittämiseksi ja tällöin pyritään kehittämään nimenomaan tilastollisiin tarkoituksiin soveltuvia tietokantamenetelmiä.

8. Tilastollisen tietojenkäsittelyn luonteesta

Tilastollinen tietojenkäsittely voidaan jakaa

- tiedon prosessointiin, joka sisältää tuotantovaiheet raakatiedoista valmiiksi taulukoiksi
- tilastolliseen analysointiin, mikä on tietoaineiston matemaattista analysointia

Tiedon prosessointi ei koostu pelkästään keruusta, käsittelystä ja taulukoinnista. Monissa tapauksissa pidetään yllä tilastollisia rekistereitä, joita tarvitaan:

a) ylläpitämään luetteloa vastaajayksiköiden nimistä ja osoitteista, alueen, koon tai tyypin tms. mukaan luokiteltuna
b) otoskehikkoja varten
c) puuttuvien tietojen tarkistukseen laajoissa tiedusteluissa

Rekisterityössä on mm. seuraavia ominaispiirteitä:

- nimet, osoitteet ja luokitukset tulee päivittää mahdollisimman usein
- rekisteritiedot on voitava linkata vastaaviin aineistotietoihin (esim. teollisuustilaston toimipaikkarekisteri on voitava linkata teollisuustilaston aineiston kanssa)
- eritasoisia yksiköitä on voitava linkata keskenään (esim. konsernit, yritykset, toimipaikat, osoitteet)

Useissa tapauksissa esim. makrotaloudellisen tiedon käyttäjä haluaa yhdistellä tietoja tuotannosta, työvoimasta, tuonnista, ulkomaankaupasta, hinnoista jne. Tällaisia tietoja pidetään yleensä eri tiedostoissa, joten esiintyy tarvetta yhdistellä näitä tiedostoja. Tietokantamenetelmät ovat sopivia ratkaisuja tämän tyyppisiin ongelmiin. Tiedon analysoija haluaa edelleen saada aineistonsa varmistettuna, tuoreena, ajankohtaisena sekä useille analyysiohjelmille sopivassa muodossa ilman uudelleen muokkauksia.

Tietokantamenetelmät näyttäisivät näin ollen sopivilta tilastollisen tietojenkäsittelyn molempiin puoliskoihin ts. prosessointiin sekä analysointiin.

Tietokantamenetelmien soveltuvuus mikro- ja makroaineistojen käsittelyyn

Tilastolliset aineistot voidaan jakaa

- mikroaineistoihin, jotka ovat alkuperäisiä tietoja yksittäisistä kohteista
- makroaineistoihin, jotka ovat aggregoituja tietoja.

Mikroaineistot edustavat pääosaa kaikesta tilastollisista tiedoista. Niitä on yleensä varsin paljon ja niitä säilytetään tavallisesti magneettinauhoilla, jotka soveltuvat huonosti THJ-tekniikoihin. Mikroaineistojen käsittely ts. virheiden tarkistus, korjaus, aggregointi jne. eivät myöskään edellytä monimutkaista tietokantatekniikkaa. Perättäiskäsittely näyttää yleensä näissä toiminnoissa olevan riittävä. Myöskin yksityisyyden suojaaminen mikroaineistoissa voi tapahtua tietokantateknologiaa käyttämättä.

Tietokantateknologian tärkein merkitys mikroaineistojen käsittelyssä näyttääkin liittyvän seuraaviin seikkoihin:

- aineiston saatavuus ja uudelleen käyttö muihin kuin alkuperäisiin tarkoituksiin
- samojen aineistojen monikertaista keruuta ja tallennusta voidaan välttää
- mikroaineistojen prosessointiohjelmien moninkertaisuutta voidaan välttää.

Eräiden maiden kokemukset viittaavat siihen, että tietokantateknologialla voidaan alentaa mikroaineistojen prosessointikustannuksia (Kanada, DDR, Unkari, Ruotsi).

Makroaineistojen osalta on syntynyt epäilyjä kaupallisten tietokantamenetelmien soveltuvuudesta niiden käsittelyyn, koska kaupalliset THJ:t ovat ko. aineistoihin nähden turhan monimutkaisia rakenteeltaan. Makroaineistoihin näyttää soveltuvan parhaiten matriisimuodossa esitettävät taulut ja tällä hetkellä missään kaupallisessa THJ:ssä ei käytetä tällaista aineiston rakennetta.

Kuitenkin tilastojen analysoijat asettavat yhä suurempia vaatimuksia analysoitaville tiedoille sekä tietojen saannin nopeudelle. Näyttääkin siltä, että kaupallisten THJ:en sijaan on kehitettävä erityisesti tilastolliseen tietojenkäsittelyyn soveltuvia THJ:tä.

9. Yhteenveto edellisestä

Yhteenvetona edellisestä voidaan todeta:

1. Tietokannan ja THJ:n toteuttaminen on relevantti kaikissa tilastollisen tietojenkäsittelyn osatehtävissä, mutta tilastollisen tietojenkäsittelyn erityisluonne vaatii erityisesti tilastolliseen tietojenkäsittelyyn soveltuvien järjestelmien kehittämistä

2. Erityisiä syitä tietokantamenetelmien soveltamiseen voivat tilastotoimessa olla:

- tuetaan aineiston integrointia eri sovellutusalueiden kesken
- meta-aineistojen ja muiden kontrollitietojen avulla voidaan saada aikaan se, että tietoaineistot ovat itsemääritteleviä ja riippumattomasti käytettävissä sovellutusohjelmiin
- saavutetaan ohjelmien ja tietopyyntöjen riippumattomuus aineiston rakenteesta (voidaan puhua aineistoriippumattomuudesta)
- saavutetaan parempi käyttäjäorientoituneisuus kuin nykyjärjestelmässä
- voidaan vähentää ulkopuolisista lähteistä koottavien tietojen määrää sekä lisätä tilastollisten lopputuotteiden käyttökelpoisuutta
- voidaan hallita aineiston integriteetti (integrity, ts. se, että aineistoja ei tuhoudu vahingossa), varmistus (security, ts. se, että aineistoa ei käytetä luvattomiin tarkoituksiin) sekä aineiston päällekkäisyys (redudanssi ts. samoja aineistoja ei säilytetä tarpeettoman monissa paikoissa). Nämä ovat erityisiä ongelmia tilatoaineistojen suurten määrien vuoksi.

3. Tietokantakäsite on osoittautunut soveliaaksi sekä makroaineistojen ts. kansantalouden tilinpidon, tilastollisten aikasarjojen, makrotaloudellisen analyysin että myös mikroaineistojen ts. tilastollisten rekisterien yhteydessä. Mikroaineistojen yhteydessä tietokantakäsite voi olla merkittävä ns. Nordbottenin käsitteen "tietoaineistopääoman tuoton lisäämisessä".

4. Taloudellisen ja sosiaalisen suunnittelun yhteydessä lisääntyvää eksaktien menetelmien käytön lisäystä ei voida tyydyttävästi palvella perinteisin menetelmin.

6. Tietokantojen kehittäminen ei ole kuitenkaan mikään "sesam-aukene" taikasana tilastollisessa tietojenkäsittelyssä. Toisaalta tällä ratkaistaan ongelmia mutta toisaalta syntyy uusia esim. niitä jotka liittyvät hajautettujen tietoaineistojen organisointiin sekä asiasisällöllisen koordinoinnin tarpeeseen.

10. Erilaisia tiedonhallintajärjestelmiä

Edellä käsiteltiin tietokannan hallintajärjestelmien jakoa sen mukaan sisältyykö niihin sovellutusohjelmia vai ei. Sovellutusohjelmia sisältäviä tiedonhallintajärjestelmiä voidaan jakaa sen mukaan, minkä osan tilaston tuotantoprosessista sovellutusohjelma peittää. Edelleen jakoperusteena voi toimia ohjelmointijärjestelmän sisäinen tehokkuus, miten sopiva se on suurten materiaalien käsittelyyn.

Karkeasti ottaen voidaan sanoa, että mitä enemmän järjestelmä on orientoitunut käyttäjän suuntaan, sitä vähemmän joustavuutta

se sisältää muiden toimintojen suhteen. [1]

Ruotsissa on päädytty eräiden siellä kokeiltujen tietokannan hallintajärjestelmien osalta seuraavaan luokitteluun. [2]

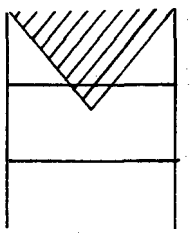
Suuressa määrin käyttäjäorientoitunut	RSDB ARKDABA GEMIC MKS
sovellutusohjelma-orientoituneet	SIMDBM RAPID

Jos tilastotuotanto jaetaan seuraaviin osa-alue osaprosesseihin:

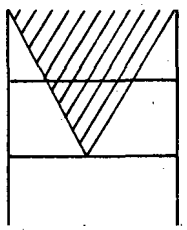
a) keruu (rekisteröinti, tarkastus, korjaus)
b) aggregointi (aggregaattien muodostaminen mikrodatasta)
c) tulostus (taulutulostusten muodostaminen aggregaateista)

niin em. tietokannan hallintajärjestelmiä voidaan luonnehtia sen mukaan, missä määrin toimintoja em. osaprosessien alueessa on rakennettu järjestelmään sisään:

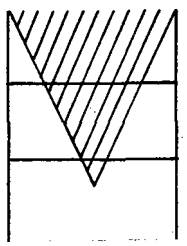
A) RSDB



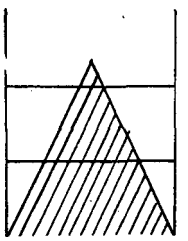
B) ARKDABA



C) GEMIC



D) MKS



1) Näistä järjestelmistä tarkemmin Hjerppe - Nurminen: Matkakertomus SCB:n metodiseminaarista 30.5.1977, Tilastokeskus.

2) L. Olsson et al: Databaser och databasharterare vid SCB, ma.

Tietokantajärjestelmien tehokkuudella ja suuriin aineistoihin soveltuvuudella on useita ulottuvuuksia, joista on vaikeata tehdä yhteenvetoa. On kysymys mm. mahdollisuuksista jakaa systeemi osiin, kirjoittaa tietokannan hallintajärjestelmän osia konelähei-sessä muodossa ja ajotehokkuudesta. Lisäksi tulee erilaisia ongelmia esiin riippuen siitä, työskenteleekö tietokannan hallin-tajärjestelmä aggregoidulla aineistolla (esim. RSDB) vai käsit-teleekö se mikroaineistoja (esim. ARKDABA).

Kaikkia näitä kysymyksiä ei ole voitu arvioida. Se mitä ruotsa-laiset ovat nyt nähneet mahdolliseksi sanoa yhteisesti on se että järjestelmiä, jotka on kirjoitettu SIMULAlle (GEMIC ja SIMDBM) tai APL:lle (MKS) ei tulisi käyttää kovin suurissa aineistoissa.

Eri tietokannan hallintajärjestelmien vertailu on vaikeata tehdä ja erityisen vaikeata se on tehdä lyhyesti. Arviot ovat sen vuoksi hyvin karkeita. Tämä johtuu mm. seuraavista syistä:

- tilastotuotannossa esiintyy erilaisia vaatimuksia ja tietokannan hallintajärjestelmät peittävät eri alueita
- tietokannan hallintajärjestelmät ovat erilaisissa määrin valmiita järjestelmiä
- tietokannan hallintajärjestelmiä ei ole vielä riittä-västi testattu ja vertailtu
- toistaiseksi ei ole riittävästi kokemusta siitä mitkä vaatimukset ovat tärkeimpiä tilastotuotannossa.

11. Eri maissa käytössä olevia järjestelmiä

Kaupallisista tiedonhallintajärjestelmisä voidaan mainita mm. TOTAL (USA, Kanada), ADABAS (Kanada), IMS ja SOCRATE (Ranskan INSEE), MARK IV (Unkari), CDC:n MARS III (Tshekkos-lovakia), Honeywellin IDS (Norjan yritysrekisteri), DMS 1100 (Yhdistyneet Kuningaskunnat).

Kaupallisten ohjelmien käyttöön vaikuttavia seikkoja ovat olleet mm:

- määrättyihin tilastollisiin sovellutuksiin on liian kallista kehittää omaa järjestelmää, koska yleensä suunnitteluresursseista on puutetta
- kaupallisia sovellutuksia voidaan käyttää väliaikais-ratkaisuina sekä henkilökunnan perehdyttämiseen tieto-kantaratkaisuihin
- kaupallisia THJ:ä käytetään usein myös ei-tilastollisiin tarkoituksiin (ts. hallinnollisiin, joihin niistä useat on nimenomaan suunniteltukin)

Kaupallisten sovellutusten puutteina voidaan pitää mm:

- tietokannan rakenne määräytyy kaupallisen THJ systeemin mukaan
- aineistorakenne kaupallisissa THJ-systeemeissä on tarpeettoman monimutkainen tilastollisiin sovellutuk-siin. Tulevaisuudessa saatetaan tosin tilastollisissa sovellutuksissa tarvita myöskin monimutkaisia raken-teita ei tosin välttämättä sellaisia, joita nyt on tarjolla.

- tilastotyössä käsitellään suuria tietomääriä, jolloin niiden käsittely kaupallisilla järjestelmillä on tehotonta
- koska saatavilla olevia sovellutuksia ei ole rakennettu tilastollisia tarkoituksia varten ne eivät sisällä kaikkia välttämättömiä toimintoja.

Tilastollisten keskusvirastojen kehittämiä järjestelmiä

Kanada kehitti vuosina 1968-1970 kaksi erityisjärjestelmää CANSIM ja STATPAK, jotka ovat osoittautuneet menestyksellisiksi ja joita edelleen käytetään.

Vuosina 1974-1975 Kanadassa siirryttiin ns. Coddin realtionaalisen lähestymistavan käyttöön: tämän tuloksena on ollut kaksi järjestelmää

- RAMP (relational access method primitives)
- RAPID (relational access processor for integrated data bases)

Tshekkoslovakiassa kehitetään ISIS järjestelmää, jonka olemassaolo meilläkin tunnetaan varsin hyvin.

Ranskan INSEE:llä on useita software pakkauksia mm. LEDA, BIBLOS ja AGROS-C, joista viimeksimainittu on kansantalouden tilinpidon ja aikasarjojen hallintajärjestelmä.

DDR:ssä on DSS-järjestelmä (Data Storage of Statistics). Tämä koostuu useista toisistaan verraten riippumattomista aliprojekteista.

Ruotsissa kehitteillä olevista järjestelmistä on ollut puhetta jo edellä.

12. Esimerkki tietokannan konstruoimisesta:
Ruotsin RSDB järjestelmä

Ruotsissa on vuodesta 1973 alkaen kehitetty erityistä aluetilastollista tietokantaa ja sen hallintajärjestelmää (RSDB). Koeversio tästä valmistui syksyllä 1976. Koeversion tarkoituksena on saada kokemuksia siitä, miten tulevaa kehittämistyötä tulee suunnata sekä tietokannan käsittelytapojen että tietosisällön osalta.

RSDB on tietokanta ts. tietoaineistovarasto, josta voidaan tuottaa erilaisia tilastollisia yhteenvetoja nopeasti käyttämällä tietokantaan sidottua ohjelmointijärjestelmää (databashanteraren). Käyttäjä voi itse tilata yhteenvetoja ja myöhemmin myös tietynlaisia tilastollisia analyysejä ja mallien muokkauksia keskustelemalla suoraan järjestelmän kanssa tietokoneterminaalia käyttäen, johon hän spesifioi haluamansa tiedot. Tulokset tulevat suoraan joko käyttäjän terminaaliin tai niitä voidaan myös tulostaa Örebrossa sijaitsevan tietokoneen kirjoittimella.

Kehittämistyön taustalla ovat ne vaikeudet, jotka liittyvät siihen, miten perinteisellä tilastojen tuotantotekniikalla voidaan tyydyttää sellainen alueellisesti jaotellun tilastotiedon tarve, jota tehostettu suunnittelu alue- ja kuntatasolla

edellyttää. Samanaikaisesti halvemmat valmistuskustannukset, suorasaantimuistit sekä lisääntynyt taito rakentaa tietokantoja ovat luoneet uusia teknis-taloudellisia edellytyksiä tilastotuotannolle. Tätä uutta tekniikkaa käyttämällä on RSDB:ssä mahdollista saavuttaa olennaisesti suurempi käyttäjäystävällisyys kuin perinteisessä tuotannossa. Näin menetellen toivotaan, että tilastollista informaatiota voidaan paremmin integroida suunnitteluun ja täten rationalisoida myös suunnitteluprosessia.

13. RSDB:n koeversio

Edellä esitettyä taustaa vasten on tapahtunut vuodesta 1973 lähtien kehittämistyötä. Syksyllä 1976 valmistui RSDB:n koeversio. Tämän version ja siihen liittyvän kokeilutoiminnan tarkoituksena on saada kokemuksia, jotta voitaisiin suunnata jatkotyötä sekä tietojenkäsittelyfunktioiden että tietokannan sisällön suhteen.

Aluetilastojen koordinointia ja yhteistyötä koskevan selvityksen (SSRS-selvityksen raportti 1975) ja sen RSDB valiokunnan työryhmän suositukset ovat olleet suuntaa antavia koeversioissa. Tämän mukaisesti tallennetaan ensisijaisesti väestö- ja tulotietoja järjestelmään sekä priorisoidaan yksinkertaisia taulukointitoimintoja. Lisäksi aineisto suuntautuu tuoreisiin tietoihin ts. historiallisia aikasarjoja ei ole toistaiseksi priorisoitu kovinkaan korkealle. Muut mahdolliset valmistus- ja analysointitoiminnot ovat toistaiseksi saaneet jäädä sikseen.

RSDB:n tapahtuva tulostus perustuu ensi sijassa hienojakoisiin tilastotauluihin, jotka on tallennettu järjestelmään (ns. matriisit). Matriisit on tallennettu suorasaantimuistille ns. makrokannaksi. Sieltä viedään tiedot tietokoneen primäärimuistille, summaukset ja yhteenvedot tehdään käyttäjän osoittamalla tavalla tauluiksi, jotka ovat luettavissa käyttäjän terminaalista. Makrokannassa oli marraskuussa 1976 tietoja väestöstä ja väestönmuutoksista vuosille 1974-1975, tulotiedot vuodelta 1974 sekä tietoja jotka valaisevat tulo- ja työllisyyskehitystä vuosian 1970-1974. Tämän lisäksi löytyy ns. katsausosa, jossa on noin 200 erilaista muuttujaa jokaista kuntaa, lääniä ja koko valtakuntaa kohden. Suuri osa tämän yleiskatsauksen muuttujista on suhdelukuja, joiden avulla on helppo verrata eri kuntia keskenään ja kuntia koko valtakunnan keskilukujen kanssa. Alin aluejako makrokannassa on kunta. Tämän lisäksi samat tiedot löytyvät läänistä ja valtakunnasta.

Väestötietojen osalta on olemassa voimakkaita vaatimuksia saada tietoja pitkälle menevällä aluejaottelulla. Samoin esiintyy tarpeita voida seurata väestönmuutoskehitystä ajassa kuntien osa-alueiden perustalta. Nämä vaatimukset asettavat erityisiä vaatimuksia tuotantojärjestelmälle, eikä niitä voida tyydyttää varastoitavassa makrokannassa, koska varastovolyymi tällöin kasvaisi kohtuuttoman suureksi. Jotta tämä vaatimus voitaisiin tyydyttää on RSDB:n koeversiossa tietokannan hallintajärjestelmä, jonka avulla voidaan valmistaa makromatriiseja yksilötietokannasta (mikrokanta).

Kun tilataan sellaisia tauluja, joissa tarvitaan mikrokantaa tekee mikrotietokannan hallintajärjestelmä ensin yksilöaineiston aggregoinnin hienojakoisiksi tauluiksi (makromatriiseiksi). Nämä toimitetaan sitten tietokoneen primäärimuistille, minkä jälkeen edelleen valmistus voidaan tehdä samalla tavalla kuin makrokantaan sisältyvällekin aineistolle.

Käyttäjä ei tule koskaan suoraan kosketukseen mikrokannan kanssa. Vain etukäteen annettuja tilastollisia tulostuksia voidaan tehdä mikrokannasta ja varastoida makromatriiseiksi, joista käyttäjä sitten voi spesifioida tauluja. Tilastollisten tulostusten rajoitukset on muodostettu siten, että ne estävät ns. mikrotietojen esiintymisen tauluissa. Käyttäjän näkökulmasta mikrokanta vastaa erilaisia tilastoaineistoja makrokannassa.

Mikrokanta, jonka sisältö vastaa SSRS:n suorituksia pyrkii käsittämään kaikki yksiköt valtakunnassa vuodesta 1974 lähtien.

Nykyisin on olemassa versio, joka käsittää tiedot koko valtakunnasta vuodelta 1974. Parhaillaan on menossa työ, jonka avulla tehdään vuosina 1974-1976 tapahtuneet muutokset. Tämän työn laskettiin valmistuvan tammikuussa 1977. Datainspektion on tarkastanut RSDB:n mikrokannan, koska se sisältää henkilörekisterin.

Vuosina 1976/1977 tapahtuva kehittämistoiminta suuntautuu koeversion trimmaamiseen (testaukseen ja pienehköihin muuttamisiin) tietokannan sisällön laajentamiseen sekä tiettyjen yksinkertaisten toimintojen edelleenkehittelyyn (suhdeluku- ja prosenttilaskut, taulujen sijasta histogrammien tekemien jne).

Kehitystoiminnan lisäksi on käynnissä kokeilutoimintaa, jonka tarkoituksena on saada käyttäjien suoria reaktioita, millainen RSDB:n sisällön tulisi olla ja mitä funktioita RSDB:n tulisi sisältää. Koetoiminnassa testataan niitä järjestelmän osia, jotka ovat valmiina. Tämän lisäksi tehdään pienessä mittakaavassa selvityksiä lisäsisällön ja lisäfunktioiden osalta.

Tähän mennessä kokeilutoimintaa on harjoitettu pääasiallisesti Älvsborgin läänin läänhinhallituksessa, jossa on havaittu, että RSDB voi poistaa tiettyjä puutteita nykyisessä tilastotuotannossa. Olennaisia näkökohtia on lisäksi saatu järjestelmän edelleen kehittämiseksi. On tehty myös yrityksiä laskea saatavia työnsäästöjä. Tämä on kuitenkin vaikeata, koska järjestelmää ensi sijassa käytetään korkeamman ambitiotason saavuttamiseen suunnittelussa. RSDB:n käytöstä aiheutuvien kustannusten laskemisessa on myös tiettyjä vaikeuksia. Alustavat laskelmat viittaavat siihen, että käyttäjä, joka käyttää tunnin järjestelmää ja tuottaa sinä aikana kuusi taulua aiheuttaa n. 200 kruunun kustannuksen. Tämän lisäksi tulevat vielä terminaali- ja tiedon siirrosta aiheutuvat kustannukset.

Kokeilutoimintaa on laajennettu syksyllä 1976 jolloin kaikille lääninhallituksille ja joillekin kunnille tuli mahdolliseksi käyttää RSDB:tä. Tilastollinen päätoimisto arvioi, että vuoden 1977 aikana voidaan kaikille läänhinhallituksille antaa käyttöön terminaali pysyvää järjestelmän käyttöä ajatellen. Olettaen, että järjestelmän kuormitus ei kasva odottamattoman nopeasti ja että SCB:n tietokonekapasiteettia voidaan lisätä pyydetyssä

laajuudessa voidaan järjestelmään liittää kiinnostuneita kuntia ja muita kiinnostuneita osapuolia vähitellen vuosina 1977/1978.

Koeversiossa on useita ominaisuuksia, jotka tekevät siitä olennaisesti uudenlaisen menetelmän valmistaa tilastotietoja.

Tällaisia ominaisuuksia ovat:

- käyttäjä, jolla ei ole erityistä atk-koulutusta, voi keskustella järjestelmän kanssa terminaalia käyttäen
- järjestelmään on varastoitu "sisältöluettelo", joka osittain auttaa käyttäjää keskustelussa ja osittain hyödyttää tietokannan ohjelmajärjestelmää (tietokannan hallintajärjestelmää) etsittäessä niitä tietoja, joita muokataan. Kun RSDB:n sisältöä lisätään, päivitetään sisältöluetteloa samanaikaisesti, kun tietoja lisätään tietokantaan. Sisältöluettelo auttaa siinä, että koeversiossa lisäaineistoa voidaan helposti liittää tietokantaan (ns. aineistoriippumattomuus).
- Tulostusten spesifiointi ja itse tulostus tapahtuu päätteiden avulla (joko kirjoituskonepäätteitä tai kuvapäätteita käyttäen), jotka on kytketty televerkon kautta RSDB:hen. Käyttäjän niin halutessa voidaan taulu (mikäli se on päätteelle epäsopivassa muodossa tai erittäin laaja sisällöltään) kirjoittaa SCB:n rivikirjoittimella paperille ja lähettää käyttäjälle postitse.
- Järjestelmään sisältyy esteitä, jotka estävät mm. yksilötietojen muokkaamisen käyttäjän taholta. Ts. käyttäjä voi saada ulos vain aggregoituja tietoja järjestelmästä.
- RSDB:n hallintajärjestelmä sisältää ohjelmia, jotka valmistavat eri tavoin tallennettuja aineistoja. Hallintajärjestelmässä on osa, joka käsittelee makrokantaa ja osa joka käsittelee mikrokantaa.

14. Muutamia näkökohtia tietokantamenetelmien kehittämisesta Tilastokeskuksessa

Tietokantamenetelmien todellista merkitystä ei voida vielä kovinkaan tarkaan arvioida. Näiden menetelmien kehittämiseen liittyy paljon luonteeltaan pitkän tähtäyksen kehittämistoimintaa. Kuitenkin tietokantamenetelmiä voitaneen jo nyt pitää eräänä merkittävänä vaihtoehtona traditionaaliselle tilastojen tuotannolle. Erityisesti erilaisten hardware- laitteiden kehitys ja suhteellinen halpeneminen viittaa siihen, että tietokantamenetelmillä voi tulevaisuudessa olla erittäin suuri merkitys myös yhteiskuntatilastojen tuotannossa.
Kehitystyön turvaamiseksi tarvitaan kuitenkin välttämättä erityisiä resursseja, jotka on kohdistettu yksinomaan ao. menetelmien kehittämiseen ja ulkopuolella tapahtuvan kehitystoiminnan seurantaan. Tällaisten resurssien tulisi olla myös pitkälle menevästi irroitettu juoksevasta tilastotuotantovastuusta.

Atk-teknisen puolen korostuessa tietokantamenetelmissä tulee huomattava osa kehittämisresursseista sijaita tietojenkäsittely-

osastolla nykyisen organisaatiorakenteen puitteissa tietojenkäsittelyosaston vastatessa sekä atk-teknisistä käyttötehtävistä että atk-menetelmien kehittämisestä.

On kuitenkin välttämätöntä, että ao. menetelmien kehittämistyöhön osallistuu myös tilastojen ja tilastojärjestelmien suunnittelusta ja kehittämisestä vastaavia henkilöitä.

> "Tilastotoimen johdon tulee hyväksyä ajatus, että tekninen kehitys ja edistys johtaa myös muutoksiin käyttäytymisessä ja että tilastotoimen traditio on irrelevantti, ellei se valmista nykyisiin ja tuleviin toimenpiteisiin. Tarrautuminen traditioon menneiden sankaritekojen vuoksi on epäilyttävä asenne. Näin ollen tilastomiesten tulee paneutua atk-projektiensa johtamiseen, jossa erityispiirteenä on moni-spasiaalisuus. Heidän tulee valmistautua toimimaan yhteistyössä ei vain omalla erikoisalallaan vaan myös tietokonemiesten kanssa, jotka useissa tapauksissa perinteisesti ovat olleet, piikki heidän lihassaan"".
> T.F. Hughes

Tietokantamenetelmät voidaan nähdä eräänä osana yhtenäistettyjen tilastoaineistojen järjestelmän kehittämisessä. Jos YTJ nähdäänkin eräänlaisena koko tilastotoimintaa koskevana periaatemallina ei tietokantamenetelmiä voida kuitenkaan toteuttaa samanaikaisesti koko tilastotointa koskevana projektina. Tietokantamenetelmien kehittämiseksi on pyrittävä löytämään sellaiset tilastotoimen osa-alueet, jotka sovellutusten kannalta näyttävät lupaavimmilta. Voidaan tietenkin ajatella, että erillisinä kehittämisalueina aluksi esiintyneet kehittämiskohteet voidaan myöhemmin integroida keskenään suuremman kokonaisuuden kattavaksi osajärjestelmäksi. Näin on esimerkiksi tapahtumassa Ruotsissa, jossa RSDB ja TSDB (aikasarjojen tietopankki) ovat olleet toistaiseksi toisistaan riippumattomia kehittämiskohteita, mutta aivan viime aikoina on ryhdytty suunnittelemaan toimenpiteitä, joilla nämä kaksi järjestelmää voidaan integroida keskenään yhdeksi järjestelmäksi eli siis järjestelmäksi, jolla on yhtenäinen hallintajärjestelmä.

Jos ajatellaan tilastokeskuksen kannalta soveliaita kehittämiskohteita eräinä tällaisinä voitaisiin pitää esimerkiksi:

- kausitasoitettujen aikasarjojen järjestelmää
- kehitteillä olevaa kansantalouden tilinpitojärjestelmän tietoaineistoa.

Kausitasoitettujen aikasarjojen osalta ollaan ehkä pisimmällä tietokantamenetelmiä ajatellen. Lisäksi tämä tiedosto on tärkeä mm. suhdanneanalyysin kannalta. Aikasarjatiedosto on luonteeltaan varsin sopiva kohde tietokantamenetelmien kehittämiseen.

Kansantalouden tilinpitojärjestelmä on sopiva kehittämiskohteena sen vuoksi, että järjestelmä on parhaillaan uusittavana ja kysymys on ehkä eräästä tilastojen käytön kannalta keskeisimmästä hankkeesta. Toisaalta tässä yhteydessä tulisi voida turvata se, että itse SKT-uudistushanke ei rupea viivästymään mahdollisista tietokantahankkeista.

Voitaisiinpa mennä niinkin pitkälle, että ajateltaisiin kehitystyötä yhdessä esimerkiksi VM:n kansantalousosaston kanssa siten, että pyrittäisiin rakentamaan terminaaliyhteys kansantalousosastolle. Tämän yhteyden kokeilu ja kehittäminen loisi erään konkreettisen pohjan hankkeelle. Aineiston tärkeimpänä käyttäjänä kansantalousosastolta saataisiin myös arvokkaita palautteita ylipäätänsä tietokantamenetelmien soveltuvuudesta tällaisessa yhteydessä.

Kolmantena mahdollisena kohteena voitaisiin ajatella aluetilastollista tietokantaa Ruotsin RSDB:n tavoin. Tässä suhteessa TK:n valmius ei kuitenkaan liene niin suuri kehittämistyöhön kuin kahdessa em. hankkeessa. Toisaalta tässä yhteydessä voitaisiin ainakin seurata tiiviisti kehitystyötä Ruotsissa ja jokatapauksessahan aluetilastot tulevat lähitulevaisuudessa kehittämiskohteeksi TK:ssa. Näin voitaisiin pyrkiä ainakin aluksi siihen, että myös tällä alalla valmisteltaisiin tietokantaprojektia.

Sisäisen koordinoinnin alueella on luonnollisesti useita mahdollisia sovellutuskohteita. Tällaisia ovat mm.

- kehitteillä olevat koordinointitiedostot (sisältötiedostot, käsite-määritelmä-luokitustiedostot jne)
- kustannuslaskentatiedosto
- henkilöstötiedosto tms.

Sovellutuskohteita on luonnollisesti löydettävissä useita.

Kehittämistyön laajuus on kuitenkin olennaisesti riippuvainen käytettävissä olevista resursseista sekä odotettavissa olevista hyödyistä. Vaikka TK:lla ilmeisesti onkin jo vuosikymmenen aikana kertynyt huomattava määrä kokemusta automaattisesta tietojenkäsittelystä lienee tilanne kuitenkin tietokantamenetelmien osalta niin, että asiaa on ryhdyttävä opiskelemaan alusta lähtien. Näin ollen ei voida olettaa, että juoksevan tilastotuotannon kannalta saataisiin kovinkaan paljon käyttöön sopivaa irti tällä vuosikymmenellä. Toisaalta kysymys on tässä asiassa pitkän tähtäimen investoinnista, jonka tuotto esim. 1980-luvun puolivälissä voi jo olla melkoinen. Riski tulisi ottaa.

LÄHDELUETTELO

(1) Codasyl Data Base Task Group, April 1971, Report.

(2) Computing in the National Statistical Services beyond 1980 Computing Research Centre, Bratislava, April 1975

(3) T.F. Hughes: The Impact of the Computer on the Statistical Service 1974

(4) Information on RSDB-systemet, Statistiska centralbyrån P/DBM, S/REK, 14.10.1976, Stockholm

(5) L. Olsson et al: Databaser och databashanterare vid SCB, Statistiska centralbyrån, 12.5.1977, Stockholm

(6) SCB-metodiseminaari, Tuula Nurminen - Reino Hjerppe, matkakertomus 30.5.1977, Tilastokeskus

(7) Sundgren, Bo: The Theory of Data Bases, Petrocelli/Charter, New York, 1975

(8) Report on the Use and Future Need for Data Base Management in National Statistical Services, CES/WP.P/150, United Nations, 26.1.1977