*SILVA FENNICA 1985, vol. 19 n:o 1: 61–65*

# HAVAINTOJEN KÄSITTELY JA AINEISTON MUODOSTUS METSÄNTUTKIMUKSEN TIEDONHALLINNAN NÄKÖKULMASTA

ERKKI KAILA

*Summary*

*OBSERVATION MANIPULATION AND DATA MATRIX DERIVATION FROM THE VIEWPOINT OF DATA MANAGEMENT*

Saapunut toimitukselle 19. 11. 1984

Artikkelissa tarkastellaan tiedonhallinnan ja tietokantojen muodostuksen merkitystä luonnontieteellisessä tutkimuksessa. Tutkimusprosessista osoitetaan vaihe, joka nimetään tiedonhallinnaksi ja joka kattaa tutkimuksessa suoritettavat toimenpiteet havaintojen merkitsemisestä havaintomatriisin esittämiseen.

Sen jälkeen esitellään tietokantoihin liittyvää käsitteistöä, erityisesti tietokannan rakentamista ja käyttöä. Tietokannan muodostamiselle esitetään kaksi perussyytä.

Artikkelin toisessa osassa tarkastellaan TUTKA-ohjelmistoa tutkimuksen tiedonhallinan apuvälineenä. Huomiota kiinnitetään viiteen rakentamiskriteeriin, joita on noudatettu ohjelmiston kehittämisen yhteydessä. Sen jälkeen syvennytään TUTKA-ohjelmistolla tapahtuvaan tietokannan muodostukseen ja käyttöön. Tietokannan muodostamisessa erotetaan kolme eri vaihetta: tietokannan suunnittelu ja kuvaaminen, tietokannan rakentaminen tietokoneelle ja tietokannan lataaminen.

Lopuksi esitetään joitakin näkökohtia metsäntutkimuksen tiedonhallintaan. Metsäntutkimuslaitoksen tutkimustoiminnasta erotetaan kolme päälinjaa: inventointitutkimus, koetoiminta ja tilaustutkimustoiminta, joiden tiedonhallinnan muodot ovat toisistaan poikkeavat.

## JOHDANTO

Metsäntutkimuslaitoksen Rovaniemen tutkimusasemalla tehdään tietokantojen soveltuvuutta metsäntutkimuksen ja metsätalouden käyttöön selvittävää tutkimustyötä. Sen yhteydessä on kehitetty erityisesti luonnontieteelliseen tutkimustoimintaan sopiva tietokannan hallintaohjelmisto TUTKA (Kaila & Taipale 1984). Tässä esityksessä tarkastellaan TUTKAn suunnitteluperiaatteita ja käyttöä. Tarkastelu rajataan *tutkimusprosessin tiedonhallinnaksi* nimetyn osavaiheen ilmiökenttään.

# TIEDONHALLINNAN MERKITYS TUTKIMUKSESSA

## Tiedonhallinta

Tutkimuksen tiedonhallinnalla tarkoitetaan tässä tapahtumasarjaa, joka alkaa tieteellisten havaintojen merkitsemisestä, jatkuu havaintokokonaisuuksien käsittelytoimina ja päättyy tiettyä analyysitoimintoa varten muodostetun havaintomatriisin esittämiseen.

Havaintojen merkitseminen voi tapahtua esim. tarkoitusta varten laaditulle lomakkeelle mittaajan toimesta tai automaattisella mittalaitteella, joka tulostaa antureista tulevat signaalit joko piirturille tai tietokoneelle tulkittavaan muotoon magneettiselle muistilaitteelle. Havaintokokonaisuuksien käsittelyssä havainnot talletetaan tarkoituksenmukaisessa muodossa myöhempää käyttöä varten.

Havaintokokonaisuus voidaan siirtää, muotoilla uudelleen tai tarvittaessa tuhota. Käsittelyn jälkeen kokonaisuudesta voidaan poimia tietyin ehdoin rajattuja osia ja syöttää ne esimerkiksi laskentaohjelmiin.

Tiedonhallintaan liittyvät perustoiminnot ovat yksinkertaisia ja sellaisina ohjelmoitavia. Kaikissa tietokonejärjestelmissä on oma ohjelmistonsa, jolla nämä toiminnot voidaan hoitaa. Suppeitten, esimerkiksi yhtä koetta palvelevien aineistojen osalta nämä välineet ovat täysin riittävät.

## Tietokanta

Laajat, useita erilaisia havaintotyyppejä käsittelevät näytteenotto- tai koejärjestelyt saattavat tarvita tiedonhallintaan kehittyneempää välinettä, tietokannan hallintajärjestelmää. Tietokannan hallintajärjestelmällä voidaan muodostaa eri havaintotyyppejä vastaavista tiedostoista integroitu kokonaisuus, tietokanta. Järjestelmä kokoaa havainnot haluttuun loogiseen järjestykseen, jota se sitten ylläpitää. Tietokantaa muodostettaessa liitetään siihen kuuluviin tiedostoihin kuvaus, jossa määritellään tiedostoista muodostuva rakenne sekä tiedostojen tietosisältö. Kuvausta nimitetään tietokannan kaavioksi. Kaavio sisältää tietokannan elementtejä koskevan nimistön. Siten se muodostaa tutkimusaineiston ylläpidettävän dokumentin. Muutokset tietokannassa edellyttävät muutosta kaaviossa, joten dokumentti on aina ajan tasalla.

Tietokantaperiaatteen omaksuminen tutkimuksen tiedonhallinnassa vaikuttaa tutkimusprosessin kaikkiin vaiheisiin. Tietokannan suunnittelu ja osa sen toteuttamista sisältyy tutkimuksen suunnitteluun. Tietokannan kaavion muodostaminen ja sitä koskevan selvityksen laatiminen merkitsevät samalla tutkimuksen tietopohjan kokoamista. Huolellisesti muodostettu tietokanta auttaa näkemään tutkimukseen liittyvän ilmiökentän kokonaisuutena. Eheä tietokanta takaa tehtyjen havaintojen käyttökelpoisuuden tutkimukseen liittyvissä päätöksentekotilanteissa. Tutkittavaan ilmiökenttään kuuluvat havainnot pysyvät koossa ja niiden yksikäsitteisyys säilyy.

## Tietokannan rakentaminen ja käsittely

Tietokannan rakentaminen muodostaa erillisen, yhtäjaksoisen vaiheen tutkimusprosessissa. Vaihe edellyttää automaattiseen tietojenkäsittelyyn ja erityisesti tietokannan rakentamiseen liittyvää osaamista. Yhteen tutkimushankkeeseen liittyvän tietokannan rakentaminen vaatii osaavalta henkilöltä keskimäärin yhden työkuukauden työpanoksen.

Tietokannan käsittely edellyttää myös sovellutukseen liittyvää ohjelmointityötä sekä tietokannan käsittelyohjelmiston tuntemista. Lisäksi kehittyneitä talletusrakenteita käyttävät ohjelmat kuluttavat tietokoneresursseja, keskusyksikköaikaa ja levytilaa huomattavasti enemmän kuin tavanomaiset, peräkkäistiedostojen käsittelyyn rakennetut ohjelmat.

Tietokannan muodostamiseen ja käyttöön liittyy siis kustannuksia. Vaikka periaatteessa kaikista tutkimusaineistoista voidaan muodostaa tietokanta, siihen ei yksinkertaisissa tapauksissa kannata ryhtyä. Tietokannan muodostamiselle on olemassa kaksi selvää perussyytä: käyttäjä haluaa 1) *helpottaa aineiston muodostamista suureksi kasvaneesta tietomassasta ja* 2) *dokumentoida suuren aineiston pitkäaikaista käyttöä varten.* Molemmissa tapauksissa oletetaan, että aineistoa halutaan käyttää useaan, mahdollisesti muuttuvaan tarkoitukseen.



# TUTKA-OHJELMISTO, TUTKIMUKSEN TIEDONHALLINNAN APUVÄLINE

## TUTKAn rakentamiskriteerit

TUTKA on erityisesti edellä esitettyjä näkökohtia huomioon ottava tietokannan hallintaohjelmisto. Sen suunnittelu aloitettiin INKA-koealoja koskevan tietojenkäsittelyn kehittämisen yhteydessä. Myöhemmin yhtä tarkoitusta varten kokoonpannun ohjelmiston suunnitteluperiaatteet yleistettiin muut tutkimushankkeet kattaviksi. Samassa yhteydessä luotiin perusteet ajattelutavalle, jota tässä yhteydessä on nimitetty tutkimuksen tiedonhallinnaksi.

TUTKAn suunnittelussa on kiinnitetty huomiota seuraaviin tutkimustoiminnan kannalta tärkeisiin näkökohtiin:

1) Tietokannan tieto- ja tiedostorakenteet on sovitettu mahdollisimman hyvin tutkimuksen tietojenkäsittelyn nykykäytäntöön.
2) Tietokannan pääsisältö muodostuu numeerisista havaintovektoreista. Havaintovektoreita kuvaava käsitteistö sekä tietokantaa koskeva rakenteellinen tieto voidaan esittää selväkielisesti merkkitietona.
3) Tietokantaan kohdistettavan kyselyn tarkoituksena on eristää tietokannan tietosisällöstä havaintomatriiseja tutkimuksessa suoritettavia jatkotoimenpiteitä varten.
4) Tietokannan käyttäjä hyötyy mahdollisten standardien käytöstä. Järjestelmä tukee esimerkiksi luokituksissa käytettävien tulkintojen standardointia ja kirjastointia.
5) Ohjelmisto ei aseta ehtoja tutkimuksen suorittamistavalle havaintojen merkitsemistä edeltävien ja havaintomatriisin muodostamista edeltävien ja havaintomatriisin muodostamista seuraavien toimenpiteiden osalta.

TUTKAN ensimmäinen versio on käyttökelpoisessa, ei kuitenkaan lopullisessa muodossa. Ohjelmiston osat ovat TUTKA UTILITY (apuohjelmisto) sekä TUTKA DML (datan käsittelyyn tarkoitettu aliohjelmakirjasto). Ohjelmiston rakentamista seuranneet tutkijat ovat esittäneet runsaasti hyviä tietokannan tietosisältöön ja ohjelmiston toimintaan kohdistuvia ehdotuksia, joista vasta osa on ehditty toteuttaa. Ensimmäinen versio täydentyy ainakin tietokannan datatiedostojen editointi- ja pakkausohjelmilla. Lisäksi olemassa olevia havaintomatriisin käsittely- ja tulostusohjelmia tullaan integroimaan TUTKAn apuohjelmistoon.

## Tietokannan muodostaminen ja käyttö TUTKAlla

TUTKAlla tapahtuva tietokannan käsittely voidaan jakaa kahteen erilliseen kokonaisuuteen: tietokannan muodostaminen ja tietokannan käyttö. Tietokannan muodostaminen on kolmivaiheinen prosessi, jonka osat ovat 1) tietokannan suunnittelu ja kuvaaminen, 2) tietokannan rakentaminen teitokoneelle ja 3) tietokannan lataaminen. Tutkimusaineistosta muodostettu tietokanta on yleensä suhteellisen suppea ja käsitteistöltään rajattu, joten erillistä ns. tiedon hoitajan roolia ei tarvita. Tietokannan muodostaminen ja käyttö on ajateltu tapahtuviksi pääasiassa hyödyn saajan itsensä eli aineistoista vastaavan tutkijan toimesta. Molemmat toiminnot edellyttävät tietokannan tietosisällön hyvää tuntemista.

Tietokannan muodostamisen ensimmäinen vaihe, tietokannan suunnittelu ja kuvaus, edellyttää tutkittavan ilmiökentän analysointia. Aluksi tutkittava kokonaisuus pyritään osittamaan hierarkkisiksi, mutta erillisiksi tutmuskohteiksi. Kohteet ja niiden väliset suhteet määräävät tietokannan rakenteen. Seuravaksi jokainen kohde (objekti) kuvataan joukolla muuttujia (attribuutteja). Muuttujat ovat joko mittalukuja tai luokiteltavia suureita. Mittaluvun tulkitsemiseksi riittää mittayksikön ilmoittaminen, mutta luokiteltavan muuttujan osalta täytyy luokkatunnuksia vastaavan luokituksen (ns. termiluettelon) olla käytettävissä. TUTKAlla toteutettavan tietokannan suunnitteluun on laadittu kolme lomaketta, joiden käyttö yksinkertaistaa tietokannan suunnittelua ja kaavion laadintaa.

Toisessa vaiheessa lomakkeille laaditusta kuvauksesta muodostetaan tietokannan fyysinen kaavio. Käytännössä se tapahtuu ajamal-

la kyselyohjelma, joka kyselee käyttäjältä lomakkeille kerätyt tiedot. Ohjelma synnyttää kaikki tietokannan käsittelyssä tarvittavat tiedostot datatiedostoja lukuunottamatta. Osaa tiedostoista voi tarkistaa ja korjata kaavion ylläpito-ohjelmilla. Jos luokiteltavien muuttujien luokituksia on valmiiksi käytettävissä, ne voidaan sisällyttää tässä vaiheessa tietokannan kaavioon pelkällä tiedostoviittauksella. Menettely on käytännöllinen useissa tutkimushankkeissa esiintyvien muuttujien (metsätyyppi, veroluokka jne.) yhteydessä ja johtaa ajatukseen standardoinnista. Tärkeimmät luokitukset ja niiden yleisesti hyväksytyt tulkinnat olisi koottava johonkin yhteiseen hakemistoon yleistä käyttöä varten.

Kolmannessa vaiheessa edellytetään ohjelmointitaitoa. TUTKAn apuohjelmisto ei sisällä yleistä tietokannan muodostusrutiinia. Tietokanta täytyy ladata datan käsittelyyn tarkoitetuista DML-aliohjelmista kootulla luonti- tai päivitysohjelmalla. Jos tietokannan tietosisällöksi tarkoitettu data on valmiiksi tarkistettu ja järjestetty, luontiohjelma on varsin helppo muodostaa. Jokanen TUTKA-tietokanta edellyttää omaa tukiohjelmistoa, jolla tietokantaa voidaan päivittää ja jolla kyetään myös lukemaan tietokannan tietosisältöä. Luontiohjelma luo perustan tälle ohjelmistolle.

Tutkimusaineistoja sisältävän tietokannan käytön päätarkoitus on poimia tietokantaan talletettua tietoa ja muodostaa siitä havaintomatriisi. TUTKAn DAP-ohjelma (Data Access Program) mahdollistaa tutkimusaineiston tämänkaltaisen valinnan käytettävissä olevasta tietomassasta. DAP kysyy käyttäjältä joukon ehtoja, joilla halutaan rajata tietokannan tietosisältöä, sekä joukon muuttujia, jotka halutaan poimia tulostustiedostoon. Ohjelma käy läpi tietokannan tietosisältöä ja poimii ehdot täyttävät tietueet tulostustiedostoon. Tulostustiedoston lisäksi käyttäjä saa tiedoston kuvauksen automaattisesti. Havaintomatriisin poiminta tietokannasta erilliseksi tiedostoksi on perusteltua kahdesta syystä: 1) tiedon säilyttäminen ja prosessointi ovat kaksi luonteeltaan toisistaan poikkeavaa tietojenkäsittelytapahtumaa ja siksi syytä pitää selvästi erillään toisistaan, 2) tilastollisen käsittelyn kannalta on edullista, jos koko havaintoaineistosta vain tarpeellinen osa on käsillä. DAPilla voidaan tehdä myös kyselyjä, joiden tarkoituksena on antaa palautetietoa tietokannan tietosisällöstä. Tässä suhteessa ohjelma on kuitenkin turhan hidas. Tietokantakohtaiset palveluohjelmat on syytä laatia DML-ohjelmia käyttäen.

## KEHITTÄMISNÄKÖKOHTIA METSÄNTUTKIMUKSEN TIEDONHALLINTAAN

### Päälinjat

Metsäntutkimuslaitoksen tutkimustoiminnasta voidan erottaa ainakin kolme päälinjaa, joihin liittyvät tiedonhallinnan muodot ovat hieman toisistaan poikkeavat. Päälinjat ovat 1) inventointitutkimus, 2) koetoiminta ja 3) tilaustutkimustoiminta.

Päälinjojen lisäksi on olemassa suuri joukko erilaisia tiedonhallintatarpeita, jotka liittyvät tutkimustyön hallinnollisiin yksityiskohtiin.

Nyt esitelty tietokannan hallintaohjelmisto tukee parhaiten kahta ensinmainittua päälinjaa. TUTKAn käyttöperiaatteet sopivat hyvin jäsentyneeseen tapahtumakokonaisuuteen, jollaiseksi metsäntutkimuksen prosessi yleensä muodostuu sisä- ja ulkotyökauden vaihdellessa.

Mittavissa inventoinneissa kerätään runsasti sellaista tietoa, jota useat käyttäjäryhmät voisivat ja haluaisivat hyödyntää. Hyödyntämisen esteenä ovat usein olleet suurten tietomäärien säilytys-, dokumentointi- ja käyttöönotto-ongelmat. TUTKAn rakentaminen on ratkaisu osaan näistä ongelmista. Suurimpien aineistojen kannalta TUTKA merkitsee kuitenkin vasta periaatteellista ratkaisua. Käytössä olevan ohjelmiston sekä tiedon talletusvälineistön täytyy vielä kehittyä nykyisestä, jotta esimerkiksi valtakunnan metsien inventoinnin kokoluokkaa olevat aineistot voidaan konstruoida reaaliajassa käsiteltäväksi tietokannaksi.

Koejärjestelyjen osalta tiedonhallinnan ogelma liittyy pikemminkin datan laadullisiin



ominaisuuksiin kuin teidon määrään. Yleiskäyttöisten muuttujien arvojen standardoinnin merkitys korostuu, kun ajatellaan kokeen sisäisten mittaustapahtumien välistä integrointia. Joidenkin subjektiiviseen päätöksentekoon perustuvien luokitusten käyttö sattaa muuttuvissa olosuhteissa olla epävarmaa. Joissakin tapauksissa myös kokeiden välisen muuttujien valinnan tulisi tukeutua jo olemassaolevaan sopimukseen (esimerkiksi samoissa olosuhteissa suoritettavat erilliset kokeet). Ns. vanhojen aineistojen käsittelyn yhteydessä standardien käyttö parantaa usessa tapauksessa aineiston luotettavuutta. Muuttujia koskevat tulkinnat on tehtävä joiltakin osin uudelleen. Standardin valinta tässä tilanteessa saattaa merkitä hienoista virhetulkintaa, mutta vain samassa määrin kuin minkä tahansa muun tulkinnan yhteydessä. Standardi kuitenkin lisää tulkinnan invarianssia jatkossa. TUTKAlla toteutettavaan tietokantajärjestelmään voidaan luoda tukijärjestelmä, jolla standardointia edistetään.

Tietokantojen muodostaminen tai tietokantajärjestelmän ylläpito voi hyvin perustein muodostua erääksi Metsäntutkimuslaitoksen ulospäin suunnatun palvelufunktion muodoksi. Metsätilaston ylläpito ja eräät tilaustutkimuksen piiriin liittyvät hankkeet synnyttävät jo nyt paineita online -periaatteella käytettävissä olevien tietokantojen rakentamiseksi. Toisaalta tulevaisuudessa mm. kirjallisuushaut tullaan kohdistamaan käyttäjän päätteeltä maailmalla kehittyviin tietopankkeihin ja saatu palaute, kirjallisuusviitteet, tullaan tallettamaan omaan tietokantaan. Kommunikointi saattaa muodostua jopa tietokantojen väliseksi. Nämä käyttömuodot kuuluvat TUTKAn sovellusalueen ulkopuolelle ja vaativat TUTKAa monipuolisempaa tietokannan hallintajärjestelmää. Tässä yhteydessä voidaan kuitenkin todeta, että tietokantojen käyttö eri muodoissaan tulee nopeasti lisääntymään ja samalla lisäämään erilaisia mahdollisuuksia tutkimustyön suoritustavoissa. Samalla tutkimuskäytäntö muuttuu.

### Varoituksen sana

Metsätieteellinen tutkimus on luonteeltan aineistokeskeistä. Piirre on vaikeasti vältettävissä, jos kokeet kestävät vuosikymmeniä. Siihen liittyy omat vaaransa. Teorian ja hypoteesien muodostusta ei saa perustaa aineistoihin. Hyvään tutkimustapaan kuuluu, että tieteelliset havainnot tehdään yksiselitteisten hypoteesien rajaamina. Tässä esitetty tiedonhallinta-ajattelu sisältää teorian aineistojen käsittelystä, mikä ohjaa tutkimusta ja auttaa teoriaperusteisen otteen säilyttämisessä.

## VIITE

Kaila, E. & Taipale, M. 1984. TUTKA-tiedonhallinta-ohjelmisto. Tietokannan muodostus ja käyttö. Metsäntutkimuslaitoksen tiedonantoja 157. 113 s.

## SUMMARY

*OBSERVATION MANIPULATION AND DATA MATRIX DERIVATION FROM THE VIEWPOINT OF DATA MANAGEMENT*

Data management is ponted out as a phase of research process. It covers research activities from observation recording to data matrix derivation. Some concepts of logical and physical database implementation are introduced.

The TUTKA – software system is introduced and observed as an instrument of research data management. Attention is paid to TUTKA's five design criteria. Database implementation and usage with TUTKA are observed precisely. Three different phases of implementing a data base are specified.

Finally, some aspects to the data management of forest research are presented.