KALERVO JÄRVELIN

Numeeriset tietokannat ja niiden käyttökustannukset: uusia ongelmia informatiikalle ja informaatiopalvelulle

Järvelin, Kalervo, Numeeriset tietokannat ja niiden käyttökustannukset: uusia ongelmia informatiikalle ja informaatiopalvelulle. [Numeric online databases and their user charges: New problems in information science and information service]. Kirjastotiede ja informatiikka 5 (2): 33—55, 1986.

Numeric online databases (NDBs) which form a relatively new but rapidly developing area in information retrieval (IR), are considered in this article. NDBs differ from the traditional bibliographic databases (BDBs) with respect to their content and use, structural complexity, data manipulation capabilities, user interfaces and user changes. Therefore their evolution is likely to have many effects on the information seeking behavior of their users. Due to the characteristics of NDBs, user charges have an essential role for their users. Recent trends in user charging policy for all online IR have been toward charging the users for what they actually retrieve from the databases, or their retrieval cost, rather than for their connect-time to the database, as is traditional in IR. Although the former are economically sound and justifiable, they cause severe problems if the users cannot predict the charges in advance, during the query negotiation phase. Due to the characteristics of NDBs, it is especially difficult for their users to estimate the charges correctly in advance. The number of items retrieved, their unit charge and their retrieval cost cannot be estimated as simply as in the BDBs, e.g. by utilizing the inverted file and by setting a fixed price per retrieved reference. As a consequence, the users must be supported by charge estimation tools which must be available in the user interfaces to NDBs. The requirements to be met by such tools and the methodology for building them are presented in the article.

Address: University of Tampere, Department of Library and Information Science, P.O. Box 607, SF-33101 Tampere, Finland.

1. Johdanto

Julkisesti saatavissa olevat online viitetietokannat ovat viimeisten viidentoista vuoden aikana nopeasti yleistyneet kirjasto- ja informaatiopalvelutyössä. Niiden käytöstä tutkimustiedon hakemisessa ja välittämisessä on tullut ammatin arkirutiinia. Lähivuosina niiden käyttö yleistyy voimakkaasti myös yleisissä kirjastoissa. Viitetietokantojen käytön arkipäiväistymisen ohella myös niiden rajoittuneisuus tiedonhaussa on tullut yhä ilmeisemmäksi: sisältäväthän ne lopulta kuitenkin vain viitteitä, joista itsestään vain harva lienee kiinnostunut. Yleensä vasta viitteiden osoittamien tietojen tai julkaisujen käyttöönsaanti tyydyttää tiedontarpeen. Vaikka viitetietokannat tarjoavatkin tiedot käyttäjiään mahdollisesti kiinnostavista viitteistä joutuisasti, saattaa varsinaisten julkaisujen yms. lähteiden hankinta osoittautua vaivalloiseksi ja hitaaksi. Vaivalloisuutta lisää se, että pelkkien viitteiden (ja mahdollisten tiivistelmien) perusteella on usein vaikeaa päätellä

viitatun dokumentin lopullista sopivuutta tiedonhakijan tarpeisiin. Lopuksi jää hakijalle vielä vaiva etsiä haettu tieto niistä dokumenteista, jotka hän päättää hankkia ja on onnistunut saamaan käyttöönsä.

Näiden vaikeuksien voittamiseksi on kehitetty uusi lähestymistapa tietokannoista tapahtuvaan tiedonhakuun. On alettu tuottaa *tieto*kantoja, joista käyttäjä saa viitteiden sijasta suoraan tarvitsemansa *tiedot* käytettäväkseen. Näitä tietokantoja on monentyyppisiä. *Tekstitietokannat* (full-text databases) sisältävät tekstimuotoisia dokumentteja, kuten esim. lakiasioita (ennakkotapauksia, lakitekstiä) ja muita hallinnollisia päätöksiä koskevia dokumentteja, uutisaineistoa, aikakauslehtiä, hakuteoksia jne. [Lern-83] [Teno-84]. Tavallisesti nämä ovat, ainakin toistaiseksi, saatavissa myös paperijulkaisuina. *Numeeriset tietokannat* (numerical databases) sisältävät pääosin tai pelkästään numeromuotoista tietoa esim. luonnontieteen ja tekniikan (kemiallisten yhdisteiden omaisuudet), yritystoiminnan ja kaupan (osakkeiden kurssit, tuotteiden valmistajat) sekä julkisen vallan ja yhteiskuntatieteiden aloilta (tilastoaineistot) [ChHe-84] [Fost-84] [Heim-82] [RuHa-84]. *Muut faktatietokannat* (other fact databases) sisältävät einumeerista mutta kuitenkin yleensä strukturoitua tietoa, samoista aihepiireistä [Cuad-82] [Cuad-84].

Tuoreet hakemistot ja katsaukset osoittavat selkeästi, että numeeriset ja muut faktatietokannat ovat nopeasti tulossa yhä suosituimmiksi ja keskeisemmiksi tiedonhaussa: vuonna 1982 USA:ssa noin puolet käytettävissä olleista noin tuhannesta online tietokannasta oli muita kuin viitetietokantoja [Cuad-82]; vuonna 1984 käytettävissä oli noin 2700 online tietokantaa, joista 900 oli numeerisia ja 300 muita faktatietokantoja, 550 tekstitietokantoja ja loput erilaisia viitetietokantoja [Cuad-84]. Euroopassa käytettävissä olevien viitetietokantojen määrä on kaksinkertaistunut vuodesta 1975 vuoteen 1983 kun taas muiden käytettävissä olevien tietokantojen määrä on *kaksikymmenkertaistunut* samana aikana [Eusi-83]. Tässä artikkelissa tarkastellaan pelkästään numeerisia tietokantoja, jotka määritellään tarkemmin luvussa 2. Valtaosa esitettävistä näkökohdista tosin sopii muihinkin faktatietokantoihin, edellyttäen kuitenkin, että ne sisältävät strukturoitua tietoa. Numeeristen tietokantojen saatavuuden, tietosisällön ja katteen, tietojen ajantasaisuuden, käyttömukavuuden ja -kustannusten kehityslinjat viittaavat siihen, että ne tulevat lähivuosina yhä suosituimmiksi tiedonhaussa verrattuna sekä viitetietokantoihin että numeeristen tietokantojen paperille painettuihin vastineisiin (esim. tilastojulkaisut). Jo nykyisin niiden taloudellinen merkitys (kuten tuottama tulo) tietokantoja käyttöön tarjoaville yrityksille on suurempi kuin viitetietokantojen [Fost-84].

Tiedonhaun kustannukset ovat perinteisesti määräytyneet sen *yhteysajan* (connect-time) perusteella, jonka tietokannan käyttäjä on ollut yhteydessä tietokannan johonkin tiedostoon [Inte-84]. Tämä merkitsee, että mitä kauemmin haku vie aikaa, sitä enemmän se maksaa täysin riippumatta haun tuloksista — tuloshan voi olla vaikka negatiivinen: tietokannassa ei ehkä ole haluttua tietoa. Tämä tilanne on viime vuosina alkanut muuttua siihen suuntaan, että haun kustannus määräytyykin yhteysajan sijasta sen tuottamien tulosten ja/tai näiden tulosten tuottamiskustannusten perusteella (esim. [Aitc-84] [DuBo -84] [Hull-84] [Hunt-84]). Tätä kehityssuuntaa voidaan hyvin perustella esimerkiksi sillä, että tällöin kustannus lasketaan sen mukaan mistä tietokannan käyttäjä itse asiassa haluaa maksaa, eikä tästä täysin riippumattoman kriteerin nojalla. Etenkin numeeristen tietokantojen yhteydessä nämä kehityssuunnat aiheuttavat kuitenkin uusia ongelmia, koska haun tuottamien tulosten tai näiden tuottamiskustannusten perusteella laskutettaessa kustannukset riippuvat sellaisista tekijöistä, joita käyttäjien ei voida edellyttää hallitsevan, ja ne voivat vaihdella varsin laajoissa rajoissa. Tuntumaa siitä, kuinka *kello käy ja lasku kasvaa,* ei voi käyttää kustannusten arviointiperusteena. Varsin yksinkertaiselta näyttävän haun suorittaminen voi tulla hyvin kalliiksi, ehkä jopa kalliimmaksi kuin se hyöty, mitä tuloksien avulla on saatavissa. Siksi on tärkeää täsmentää kustannusten muuttuvien laskentatapojen aiheuttamat uudet ongelmat ja kehittää keinot niiden ratkaisemiseen.

Numeerisia tietokantoja koskeva tutkimus ja selvitystoiminta on kansainvälisestikin tarkastellen vielä nuorta ja siksi myös varsin vähäistä. Tämän takia artikkelissa kuvataan numeerisia tietokantoja, niiden tietosisältöjä, rakennetta, käyttömahdollisuuksia, käyttäjäliitäntöjä, käyttökokemuksia ja -vaikutuksia. Tarkastelu tehdään pääasiassa tietokantojen käyttökustannusten kannalta. Lisäksi tarkastellaan käyttökustannusten laskentatapojen kehitystä ja merkitystä numeeristen tietokantojen käytölle. Käyttökustannusten laskentatapojen kehityksellä on merkitystä myös perinteiselle viitteenhaulle, koska samat kehityslinjat näkyvät myös niiden käyttökustannusten laskennassa. Artikkelissa osoitetaan ja määritellään kustannusten laskentatavasta johtuvia uusia ongelmia ja tarkastelleen niiden keskeistä merkitystä numeeristen tietokantojen käytölle. Asiaan liittyy sekä pelkästään taloudellisia tekijöitä että myös tiedon saannin tasa-arvotekijöitä. Näitä ongelmia ei ennen ole tarkasteltu

numeeristen tietokantojen yhteydessä. Artikkelissa myös osoitetaan, että käyttökustannusten laskentatavasta johtuvien ongelmien ratkaisu edellyttää sellaisten välineiden kehittämistä ja liittämistä numeeristen tietokantojen käyttäjäliitäntöihin, joiden avulla voidaan etukäteen *ennustaa* tiedonhaun kustannukset. Lopuksi tarkastellaan tällaisille välineille asetettavia vaatimuksia ja esitetään metodologia niiden kehittämiseksi.

Artikkelin tarkoituksena on siis kuvata numeeristen tietokantojen ja niiden käytön ja käyttökustannusten laskennan nykytila ja kehitysnäkymät lähitulevaisuudessa sekä määritellä erityisesti käyttökustannusten laskennan aiheuttamat ongelmat näiden tietokantojen käytössä. Lisäksi tavoitteena on määritellä täsmällisesti käyttökustannusten ennustamiseen liittyvät ongelmat ja metodologia niiden ratkaisemiseen.

Tietokantojen tuotantoon ja käyttöön liittyvien käsitteiden nimityksissä esiintyy horjuvuutta, jonka takia täsmennetään tässä kirjoituksessa käytettävä käsitteistö seuraavasti:

- *tietokanta* (database) on tietojen varasto, joka koostuu joukosta tiedostoja, jotka puolestaan koostuvat joukosta tietueita;
- *tiedonhallintajärjestelmä* (database management system, DBMS) on ohjelmisto, jonka avulla huolehditaan tietokannan ylläpidosta ja käytöstä;
- *kyselyjärjestelmä* (query subsystem) on se tiedonhallintajärjestelmän osa, jonka avulla tietokannan sisältämiä tietoja voidaan hakea (viitetietokannoissa käytetään usein termiä hakuohjelma (search program, search system));
- *kyselyllä* (query) tarkoitetaan sekä tietontarpeen sanallista kuvausta että sen esitystä kyselyjärjestelmän hyväksymällä kyselykielellä (viitetietokannoissa käytetään usein termiä hakupyyntö tai -lausuma (search statement));
- *loppukäyttäjä* (end-user, client) on se henkilö tai taho, joka tietokannasta haettavia tietoja tarvitsee, esimerkiksi työssään;
- *välittäjä* (intermediary, searcher) on henkilö, joka ammattitaitonsa nojalla suorittaa tiedonhaun loppukäyttäjän puolesta (esim. kirjastonhoitaja tai informaatikko), mikäli loppukäyttäjä ei itse suorita hakua;
- *käyttäjä* (user, system user), on kuka tahansa henkilö, joka käyttää kyselyjärjestelmää, siis joko välittäjä tai loppukäyttäjä (joskus myös *välitön* (immediate) käyttäjä);
- tietokannan *myyjä* (vendor) on mikä tahansa taho, joka tavalla tai toisella hankkii tietokannan käyttöönsä, organisoi ja ylläpitää sitä ja tarjoaa sen käyttäjien käyttöön, tavallisesti maksua vastaan;
- tietokannan *tuottaja* (producer) on mikä tahansa taho, joka kerää, organisoi ja tallentaa tietokannan sisältämät tiedot, tavallisesti magnettinauhalle, ja luovuttaa ne, yleensä korvausta vastaan, tietokannan myyjälle. [Ullm-80] [BoMC-84] [Henr-80]

Usein puhuttaessa kirjasto- ja tietopalvelualan henkilöstöstä tiedonhaun yhteydessä ajatellaan vain välittäjiä, jotka toimivat loppukäyttäjien *puolesta*. Tämä käsityskanta on valitettavan rajoittunut. On selvää, että esimerkiksi indeksoijat kuuluvat, tiedonhaun kannalta tarkastellen, alan henkilökuntaan, vaikka he toimisivatkin tietokantojen tuottajien tai myyjien organisaatioissa. Mutta myös ne henkilöt, jotka tietokantojen tuottajien organisaatioissa tai tutkimuslaitoksissa *kehittävät* kysely- ja järjestelmiä ja niihin liitettäviä apuneuvoja ja välineitä paremmin käyttäjien tiedonhakuun soveltuviksi, siis heitä *varten*, ovat tätä henkilökuntaa. Tällaisen kehitystyön tuloksia viitetietokantojen kyselyjärjestelmissä ovat mm. *termin katkaisu* (truncation) ja *termikaaviot* (term templates), *läheisyysoperaattorit* (proximity operators) ja *termiryhmähaku* (expand tai explode-komennot) [BoMC-84] [Henr-80] [SaMc-84]. Alan tutkimus ja opetus eivät voi olla vain *käytön* tutkimista ja opettelua. Sen tulee olla myös käytön aktiivista kehittämistä ja helpottamista.

2. Numeeriset tietokannat

2.1. Suhde muuhun julkaisemiseen

Seuraavaksi tarkastellaan numeeristen tietokantojen määrittelyä ja suhdetta toisaalta muun tyyppisiin tietokantoihin ja toisaalta niihin pääasiassa paperimuotoisiin julkaisuihin, joita ne vastaavat, korvaavat tai täydentävät. Tavallisesti numeeriset tietokannat määritellään seuraavaan tapaan: ne ovat *tietokoneella käsiteltävissä olevia tietojen kokoelmia, jotka ovat pääasiassa numeerisia.* Tavallisesti erotetaan kolme alaryhmää: 1) *puhtaat numeeriset* tietokannat, kuten tilastoja aikasarjatietokannat, 2) *tekstinumeerit* tietokannat, jotka sisältävät esim. numeerisen aineiston lisäksi niiden tulkintaa, 3) *ominaisuustietokannat,* kuten elektroniikan komponenttien, kemiallisten yhdisteiden tai saastelähteiden ominaisuuksia kuvaavat tietokannat. Nämä määritelmät ovat varsin löyhiä, eivätkä perustu erottelevien piirteiden systemaattiseen soveltamiseen, kuten esim. luokitusten teossa on tapana vaatia. Toisaalta ovat numeeriset tietokannatkin niin monisyinen ilmiö, ettei yhtä kaikenkattavaa ja tarkkaa määritelmää voida esittää. Keskeisiä yhdistäviä piirteitä numeerisille tietokannoille ovat a) se, että ne sisältävät tietoja viitteiden sijasta, b) numeeristen tietojen keskeinen osuus tietosisäl-

lössä ja c) se, että tiedot ovat strukturoituja.

Seuraavassa on joitakin esimerkkejä tärkeimmistä numeeristen tietokantojen myyjistä ja joistakin näiden myymistä, pääasiassa taloustietoa sisältävistä tietokannoista lähteiden [Fost-84] ja [KaLM-84] perusteella:

- SIA-Computer Services (Lontoo): 16 numeerista tietokantaa, mm.
  - Financial Times Currency and Share Index Databank
  - IMF International Financial Statistics
  - OECD Main Economic Indicators
  - CSO UK Central Statistical Office Databank
  - ...
- ADP Network Services International (Lontoo): 32 numeerista tietokantaa, mm.
  - BCD Business Condition Digest
  - CPI Consumer Price Index (USA)
  - PPI Producer Price Index (USA)
  - BANK Bank of England Database
  - ...
- I.P. Sharp Associates: 94 numeerista tietokantaa, mm.
  - OECD Indicators of Industrial Activity
  - OECD Quarterly National Accounts
  - Business International Economic Forecasts
  - Eastern Block Countries Economic Data
  - ...
- Dialog Information Retrieval Services (Palo Alto): 15 numeerista tietokantaa, mm.
  - Predicasts tietokannat
  - Business International Data Time Series
  - ...
- Chase Econometrics: 125 numeerista tietokantaa, mm.
  - United Nations Demographies Database
  - Far East Forecast Database
  - World Agriculture Supply and Disposition Database
  - ...
- Data Resources Inc. (DRI): yli 75 numeerista tietokantaa, mm.
  - International Trade Information Secvice Database
  - ...

Usein numeerisen tietokannan tuottaja tuottaa myös paperimuotoista tietokannan vastinetta, kuten esimerkiksi tilastojulkaisua, yritys- tms. hakemistoa, käsi- tai taulukkokirjaa tai muuta hakuteosta. Nämä paperimuotoiset julkaisut ovat osa kirjasto- ja tietopalvelujen perinteistä aineistoa, jota on käytetty palvelujen tuottamisessa ja jonka tuntemus kuuluu keskeisenä kirjantonhoitajan ja informaatikon ammattitaitoon kunkin alan kirjasto- ja informaatiopalvelutyössä. Vastaavien numeeristen tietokantojen tuntemus on kuitenkin yleisesti varsin vähäistä. Suomessakin kirjastonhoitajat ja informaatikot tuntevat hyvin Suomen Tilastollisen Vuosikirjan, mutta sen tietokantavastineiden, Tilastokeskuksen tiedostojen, tuntemus on varsin vähäistä. *Syitä* tähän ovat mm. numeeristen tietokantojen suhteellinen uutuus; se, että loppukäyttäjät ovat joissakin tapauksissa tottuneet hakemaan tarvitsemansa tiedot itse; se, että kirjasto- ja informaatiopalvelualan koulutusohjelmissa ei aivan viime vuosia lukuunottamatta ole tarjottu niitä koskevaa opetusta; numeeristen tietokantojen käyttö yleensä poikkeaa viitetietokantojen käytöstä jonkin verran vaivalloisempana [ChHe-84].

Chenin ja Hernonin kokoomateoksen [ChHe -84] mukaan numeeriset tietokannat tulevat yhä yleistymään ja osittain myös korvaamaan paperimuotoiset vastineensa, tietokantojen *tuoreempi* tieto tulee yhä tärkeämmäksi ja kirjastonhoitajien ja informaatikkojen osuus niiden käyttäjäkunnasta tulee lisääntymään huomattavasti. *Pelkästään se, että tietojen esittämiseen käytetty väline vaihdetaan paperista tietokannaksi, ei kelpaa syyksi siihen, etteivät numeeriset tietokannat enää kuuluisi kirjastonhoitajien ja informaatikkojen keskeisiin työvälineisiin.* Aluksi vieroksuttiin myös viitetietokantoja, kun perinteen mukaisesti oli totuttu painettuihin bibliografioihin. Nykyisin viitetietokannat, vajaan kahdenkymmenen vuoden kehityksen jälkeen, kuitenkin ovat informaatiopalvelun 'leipäpuu'. Samat syyt, jotka aikanaan pakottivat luopumaan laajojen bibliografioiden painamisesta ja siirtymään viitetietokantojen tuottamiseen, pakottavat vähitellen yhä laajempaan numeeristen tietokantojen tuotantoon ja käyttöön. Julkaiseminen paperilla on liian hidasta, kallista ja joissain tapauksissa lähes mahdotonta (esim. lähes päivittäin muuttuvat tiedot). Kirjastonhoitajien ja informaatikoiden osaamisen tasoon kohdistuvat vaatimukset ovat kuitenkin edelleen vähintään entisen suuruiset: oman toimintasektorin tiedonlähteet — siis tietokannat — on edelleen tunnettava ja niitä on osattava myös käyttää. Tämän takia on nopeasti kehitettävä numeerisiin tietokantoihin liittyvää opetus- ja tutkimustoimintaa myös Suomessa.

Numeeriset tietokannat tarjoavat *lukuisia etuja* verrattuna paperimuotoisiin vastineisiinsa:

- tietojen *täydellisyys:* usein tiedot tietokannoissa ovat täydelliset tai ainakin laajemmat kuin painetuissa julkaisuissa;
- tietojen *tuoreus:* yleensä tiedot tietokannoissa ovat tuoreempia kuin painetuissa julkaisuissa; joissain tapauksissa painamisen hitaus on kokonaan estänyt paperimuotoisen julkaisemisen;
- tietojen *käsittelymahdollisuudet:*
  - *tietokannoista voidaan poimia* käsiteltäviksi

juuri ne tiedot, joista ollaan kiinnostuneita (esim. painetun taulukon tiedoista ei enää voida erotella vain havaintoaineiston jotakin osapopulaatiota koskevia tietoja);

- tietokannan tietoja voidaan *täydentää* käyttäjän itse syöttämillä tiedoilla ja molempia voidaan sitten *käsitellä yhdessä;*
- *tietoja voidaan tutkia ja käsitellä käyttäjän omien luokitusten* määrittämissä ryhmissä (esim. painetun tilastojulkaisun tietojen uudelleenluokittelumahdollisuudet rajoittuvat pelkästään luokkien yhdistämiseen);
- tietoja voidaan *yhdistellä* useista eri lähteistä (tiedostoista, tietokannoista) ja käsitellä yhdessä (manuaalisti tämä on hyvin vaivalloista);
- eri lähteistä saadut tiedot voidaan *yhdenmukaistaa* esimerkiksi käytettyjen indeksien perusvuosien tai valuuttojen suhteen;
- tietoja voidaan *käsitellä* sopivilla analyysiohjelmistoilla (kuten esim. tilastolliset tai taloudelliset analyysit ja mallit);
- tietojen *tulostusmahdollisuudet:* poimittua ja käsiteltyä numeerista aineistoa voidaan automaattisesti havainnollistaa graafisina esityksinä kuten käyrinä tai histogrammeina.

Tiedot saadaan siis huomattavasti vaivattomammin kerättyä juuri niistä seikoista, joista ollaan kiinnostuneita, käsiteltyä asiaankuuluvalla tavalla sekä vielä tulostettua tiedon hyväksikäyttöä tukevassa muodossa. Kaikki tämä tietenkin onnistuu myös manuaalisti, mutta on usein hyvin vaivalloista.

Perinteisiin viitetietokantoihin verrattuna on numeerisissa tietokannoissa monia eroja, jotka koskevat sisältöä, rakennetta, käyttötapoja ja käyttäjäliitäntöjä. Tärkein eroista koskee tietokantojen *sisältöä.* Numeeristen tietokantojen sisältämiä tietoja voidaan tavallisesti *välittömästi käyttää tiedontarpeen tyydyttämiseen.* Tiedot ovat todellisuutta tai abstraktia maailmaa koskevia tosiasiaväitteitä tai ennusteita, vastauksia itsessään, eivätkä viittauksia mahdollisten vastauksien lähteille jonnekin toisaalle. Viitetietokannoista taas tavallisesti saadaan vain viittauksia tiedon lähteille; viitteet kokonaisuuksina tai niihin sisältyvät tiedot ovat harvoin haettuja lopullisia vastauksia.

Numeeristen ja viitetietokantojen *rakenteelliset* erot jakaantuvat *loogisen* ja *teknisen* rakenteen eroihin. Tyypillinen viitetietokanta koostuu useista — jopa kymmenistä — viitetiedostoista ja niiden käänteistiedostoista. Vaikka dokumenttien kuvailutavat vaihtelevat tiedostosta toiseen (esim. tietojen lukumäärä, tyyppi ja esitystapa), aina kuitenkin on kysymys dokumentin jonkinlaisesta kuvailusta. Vaikka viitetietokannassa olisi kuinka monta tiedostoa tahansa, niin *loogiselta kannalta* tarkastellen tiedot koskevat vain *yhtä todellisuuden objektityyppiä,* dokumenttia. Tavallisesti numeerinen tietokanta kuvaa *monia todellisuuden objektityyppejä, niiden suhteita tai tapahtumia* eri tiedostoissaan. Tiedot voivat koskea henkilöitä, väestöjä, organisaatioita, talouselämää, tuottajia, tuotteita, markkinoita, ostajia, ostoja, toimituksia jne. Kun viitetietokannan eri tiedostojen sisältämien viitteiden välillä ei ole tärkeitä *loogisia yhteyksiä,* niin numeerisen tietokannan eri objekteja koskevat tiedot liittyvät toisiinsa tavoilla, jotka vastaavat näiden objektien todellisia suhteita (esim. suhteet tietyn tuottajan, tuotteen, markkinoiden ja ostajan kesken).

Teknisesti viitetietokannat ovat myös hyvin samanlaisia keskenään. Jokseenkin kaikki kaupalliset online viitetietokannat perustuvat käänteistiedoston käyttöön, vaikka eroja onkin siinä, mistä ja kuinka monista viitetiedoston kentistä (tiedoista) käänteishakemisto(t) tehdään. Käänteistiedosto tarjoaa online-viiteenhaussa monia etuja, jotka kuitenkin perustuvat siihen, ettei viitetiedostojen välillä ole loogisia yhteyksiä, ja siihen, että tietoja päivitetään suhteellisen harvoin (esim. joka toinen viikko) verrattuna kyselyjen lukumäärään. Numeeristen tietokantojen tekninen toteutustapa on monimutkaisempi.[2] Käytössä on useita tietomalleja (mm. relaatiomalli, hierarkkinen tietomalli ja verkkomalli [Ullm-80]) ja monia eri tiedostorakenteita käänteistiedostorakenteen ohella, esimerkiksi peräkkäisrakenne, taulukoitu peräkkäisrakenne, hajarakenne ja monilistarakenne (näiden kuvailu, ks. esim. [Hans-82] [TeFr-82] [Wied-77] [Ullm-80]). Tietokannan käyttö- ja ylläpitotavoista riippuu, millainen tiedostorakenne millekin tiedoille parhaiten soveltuu. Missään tapauksessa käänteistiedosto ei aina ole paras ratkaisu numeerisiin tietokantoihin. Erot tiedostorakenteissa aiheuttavat eroja myös kyselyjen suorittamisessa. Kun viitetietokannan käänteistiedostosta voidaan yleensä tutkia, *viitteitä vielä hakematta,* kuinka moni viite sisältää hakusanat NUCLEAR ENERGY *ja* SAFETY, ei tällainen onnistukaan esim. peräkkäisrakenteessa: jokainen viite tulisi tutkia.

Viitetietokannan tavallinen *käyttötapa* koostuu pelkistettynä a) kyselyn alustavasta muotoilusta jotakin tiedostoa varten, b) käänteistiedoston tutkimisesta sen selvittämiseksi, montako viitettä kullekin termille ja termiyhdistelmälle tiedostossa löytyy, c) kyselyn mahdollisesta uudelleenmuotoilusta tämän perusteella (tarkentaminen, laajentaminen) ja d) viitteiden poiminnasta ja tulostamisesta sekä e) kyselyn mahdollisesta toistamisesta tietokannan muissa tiedostoissa sellaisenaan tai muokattuna [BoMC-84] [Henr-80]. Yhtä useamman tiedoston käyttö ei ole välttämätöntä — niiden käyttö tuottaakin usein varsin

paljon sellaisia viitteitä, jotka löytyivät jo ensimmäisestä tiedostosta. Numeerisen tietokannan tavallinen käyttötapa poikkeaa tästä olennaisesti, koska a) tavalliseti kyselyä ei tehdä jotakin tiedostoa varten, vaan se kohdistuu moneen tiedostoon, b) käänteistiedostoa ei ole tai sitä ei voida käyttää sen tutkimiseen, montako tietuetta kysely tuottaa vastauksenaan, c) relevanssiongelmaa, joka vaatisi kyselyn laajentamista tai tarkentamista, ei ole, ja d) tietojen poiminnan jälkeen niitä tavallisesti vielä yhdistellään ja jalostetaan (esim. tilastollinen käsittely) ja tulostetaan usein graafisina esityksinä sekä d) kyselyä ei voida toistaa saman tietokannan eri tiedostoissa (sellaisia ei ole). *Tiedostojen käsittelyn suhteen* kysely viitetietokannassa on *yksivaiheinen:* hakusanat sisältävät tietueet poimitaan hakemiston kautta viitetiedostosta. Tyypillinen kysely numeerisessa tietokannassa on *monivaiheinen:* tiedot poimitaan vaiheittain useista tiedostoista, ja poiminnan jälkeenkin on ehkä tuotettava useita välituloksia ennen lopullisen vastauksen valmistumista. Viitetietokannasta vastaus saadaan jokseenkin aina muutamassa sekunnissa, kun vastauksen saanti numeerisesta tietokannasta kestää muutamasta sekunnista useisiin minuutteihin, jopa kymmeniin minuutteihin, riippuen siitä, millaisia tietoja haetaan ja miten niitä yhdistellään ja käsitellään.[3]

Numeeristen ja viitetietokantojen sisällön, rakenteen ja käyttötapojen erot heijastuvat myös niiden *käyttäjäliittymiin.* Kyselyä määriteltäessä on periaatteessa aina määriteltävä haluttu *toimenpide,* sen *kohde* tai *kohteet* ja toimenpiteen suorittamista ohjaavat *parametrit.* Viitetietokannasta haettaessa voidaan tavallisesti suorittaa seuraavia toimenpiteitä [BoMC-84] [Henr-80]:

- *Hakujoukon määritys:* käytetään *komentosanaa,* kuten **hae** tai **select,** ja sitä seuraavia *hakusanoja* ja *loogisia operaattoreita,* esimerkiksi **select** NUCLEAR ENERGY **and** SAFETY; tässä hakusanat ja operaattorit ovat toimenpiteen parametreja. Jos hakusanojen halutaan esiintyvän viitteiden tietyssä kentässä, voidaan tämä ilmaista *kenttämäärityksen* avulla: merkeillä **au:** ilmaistaan tekijäkenttä, merkeillä **ti:** nimekekenttä jne., esimerkiksi vaikka **select au:** VENNAMO, V. **and ti:** NUCLEAR ENERGY **and ti:** SAFETY. Tämän toimenpiteen kohteena on aina se (yksi) viitetiedosto, joka kyselyä varten on avattu.
- *Välitulostus- tai tulostus:* esimerkiksi komentosana **display** tai **print** parametreinaan *tulostusformaatti* ja tulostettavien viitteiden *lukumäärä* sekä kohteenaan viimeksi määritelty tai tunnuksellaan ilmaistu (yksi) hakujoukko.
- Muita tavallisia toimenpiteitä ovat *sanaston selaus* sekä kyselyn *talletus, toisto,* ja *lopetus.*

Hieman pelkistäen voidaan siis sanoa, että viitteiden käsittelyyn on tarjolla kaksi toimenpidettä: haku ja tulostus. Numeerisessa tietokannassa tietojen käsittelyyn on tarjolla enemmän toimenpidevaihtoehtoja. Pelkän haun (poiminnan tiedostosta) ja tulostuksen lisäksi tietoja voidaan yhdistellä ja käsitellä eri tavoin. Koska toimenpiteiden kohteina voi olla yksi tai useampi tiedosto tai muiden toimenpiteiden tulos, on niiden kohde tai kohteet aina täsmällisesti ilmaistava. Kyselyissä ei myöskään riitä hakusanojen ja loogisten operaattoreiden luettelu, vaan kukin hakuehto koostuu aina kentän nimestä, vertailuoperaattorista ja arvosta tai toisesta kentän nimestä, esimerkiksi YEAR = 1986 tai MARKET-SHARE = 0.30. Hakuehtoja voidaan yhdistellä loogisten operaattoreiden avulla samaan tapaan kuin viitteenhaussa. Yksinkertainen kysely, jolla etsitään tuotteita, joiden markkinaosuus on yli 30 %, voi näyttää esimerkiksi seuraavalta: **select** PRODUCT-NAME **from** MARKET-FILE **where** YEAR = 1986 **and** MARKET-SHARE = 0.30. Tässä **select**-osa ilmaisee toimenpiteen ja tulostettavat kentät, **from**-osa toimenpiteen kohteen ja **where**-osa hakuehdot. [Ullm-80] [SaMc-83]. Vaikka viitetietokantojen kyselykielet kokonaisuutena poikkeavatkin toisistaan varsin paljon, ovat perustoimenpiteet ja periaatteet kuitenkin varsin samanlaisia. Numeerisissa tietokannoissa erot ovat paljon suuremmat, mikä johtuu toisaalta tietokantojen aihepiirien ja käyttötarkoitusten ja toisaalta käsittelymahdollisuuksien ja teknisten toteutusratkaisujen vaihtelusta. Pelkästään komentojen parametriluettelot voivat olla varsin monimutkaiset. Numeeristen tietokantojen nuoruudesta johtuu myös niiden käyttöä helpottavien apuneuvojen ja opastusvälineiden kehittymättömyys [Gaul-84]. Niiden käyttö saattaa näinollen edellyttää, ainakin toistaiseksi, aihepiirin tuntemuksen lisäksi myös perusteellista harjoittelua [KaLM-85].

Yritysten ja muiden organisaatioiden toimintaansa varten käyttämistä *taloudellis-hallinnollisista* tietokannoista numeeriset tietokannat eroavat myös monissa suhteissa. Gaultin [Gaul-84] mukaan on kyse erosta tietokantatekniikan *yrityssovellusten* (business applications) ja *tieteellisten sovellusten* (science applications) välillä. Taloudellis-hallinnollisia sovelluksia voidaan luonnehtia seuraavasti:

- Tietojen *kontrolli* ja *suojaus* ovat tärkeitä: usein kyseessä on organisaation toiminnan kannalta arkaluontoisia tietoja, joiden saanti ja käyttö on toiminnalle keskeisen tärkeää ja joiden joutuminen vääriin käsiin tulee estää. [Gaul-84]

- Tietojen pääkäyttö liittyy organisaation *päivittäisten rutiinitehtävien* suorittamiseen.
- Tavallisimpia ovat *yksinkertaiset* kyselyt, joiden tyyppi voidaan ennakoida ja joihin tulee löytää vastaus välittömästi.
- Kyselyjen *kustannuksilla ei ole merkitystä* — kyselyt ovat osa toiminnan päivittäistä rutiinia ja tietokanta on suunniteltu palvelemaan tätä käyttöä mahdollisimman tehokkaasti.

Numeerisia tietokantoja (tieteellisiä sovelluksia) taas voidaan luonnehtia seuraavasti:

- Tietojen *kommunikointi* ja *vapaa käyttö* (tai *myynti)* ovat keskeisiä: julkista resurssia halutaan levittää mahdollisimman laajaan käyttöön. [Gaul-84]
- Tietojen pääkäyttö liittyy *tutkimus- ja kehitystyöhön* sekä organisaation *strategiseen ja taktiseen päätöksentekoon,* jolloin kyselytyyppejä on vaikea ennakoida.
- Kyselyt ovat usein *monimutkaisia* eikä vastauksen välttämättä tarvitse tulla välittömästi; usein voidaan odottaa esim. puoli tuntia.
- Kyselyjen *kustannuksilla on keskeinen merkitys,* koska usein voidaan vastaavaa tietoa hakea myös muista lähteistä tai voidaan päättää tulla toimeen ilman tietokannan tarjoamaa tietoa [CaMS-75] [Järv-86a].

Vaikka tiedonhallintajärjestelmien käytön helposti sanotaankin olevan tiedon tai informaation kontrollointia, ovat ne kuitenkin keskeisiä välineitä tieteellisen kommunikaation toteutumisessa ja tullevat siinä yhä merkittävämmiksi.

2.2. Vaikutukset tiedonhakuun ja tiedon käyttöön

On paikallaan tarkastella myös numeeristen tietokantojen vaikutuksia tiedonhakuun ja tiedon käyttöön sekä joitakin näistä saatuja kokemuksia. Numeeristen tietokantojen vaikutukset voidaan tyypitellä seuraavasti:

- *Välittömät vaikutukset:*
  - *tiedon saatavuus: sisällöllinen saatavuus* (mitä asioita koskevia tietoja on saatavilla ja kuinka laadukkaita ne ovat), *maantieteellinen saatavuus* (mitä alueita koskevia tietoja on saatavilla ja missä niitä on saatavilla), *ajallinen saatavuus* (mitä ajanjaksoja koskevia tietoja on saatavilla ja kuinka nopeasti ne ovat saatavilla), *tekninen saatavuus* (mitä tietoja, taitoja ja välineitä tietojen hankinta edellyttää), *taloudellinen saatavuus* (mitä tietojen hakeminen maksaa);
  - *tiedon hankintatavat:* miten tietojen eri käyttäjäryhmien tiedonhankintatavat tai -käytännöt kehittyvät.
- *välilliset vaikutukset:*
  - *työn tekemisen käytäntö;*
  - *työn tietoperustat ja tiedontarpeet.*

Koska numeeriset tietokannat ovat kasvava osa yhä kasvavaa *kaupallista* tiedonvälitystä, näkyy kaupallisuus myös tietojen *sisällöllisessä* saatavuudessa. Tietokannat kattavat parhaiten aihepiirejä, joilla tällainen tiedonvälitys on taloudellisesti kannattavinta, siis ennenkaikkea yksityisen talouselämän toiminnan, luonnontieteiden, lääketieteen ja tekniikan kannalta tärkeitä tietoja. Humanistiset tieteet, yhteiskuntatieteet ja filosofia tullevat hitaasti perässä, koska maksukyky ja tietojen potentiaalinen markkina-aluekin on pienempi. Kehitys tässä lienee siis sama kuin viitetietokantojenkin aihepiirittäisen katteen kehitys: numeeriset tietokannat edistävät tiedon saatavuutta ensiksi ja eniten aloilla, joihin liittyvät suurimmat taloudelliset intressit, tiedon keruuja tutkimusponnistelut. Toinen tietojen sisällölliseen saatavuuteen liittyvä keskeinen ongelma on *tietojen luotettavuus* l. validiteetti ja *vastuu* siitä [Fost-84]. Viitetietokannoissa tämä ei ole merkittävä ongelma, koska virheelliset viitetiedot harvoin voivat johtaa merkittäviin vahinkoihin. Paperijulkaisun kustantajakin voi helposti suojautua nimiölehdelle painetun julkaisuvuoden taakse: asiat ovat voineet sittemmin muuttua, ja lukijan tulee osata ottaa se huomioon. Toisin on numeerisissa tietokannoissa. Online-tietoihin liittyy helposti ainakin ajantasaisuuden illuusio, vaikkei aihetta olisikaan. Virheet valuuttojen ja osakkeiden myyntikurssitiedoissa, lääkkeiden sivuvaikutustiedoissa tai kemiallisten yhdisteiden toksisuustiedoissa saattavat helposti aiheuttaa suurta vahinkoa. Mitä tarkemmin tiedot validoidaan, sitä kalliimmiksi ne hinnoiteltaneen. Tämä piirre näkyy paperijulkaisujenkin hinnoissa.

Numeeristen tietokantojen tietojen *maantieteellinen kate* on nykyisin jokseenkin ainoastaan teollistuneet länsimaat. Näin on voitu havaita mm. tilastotietokannoista [KaLM-85]. Syynä tähän lienevät niin tilastotoimen kuin tietotekniikankin kehittymättömyys esimerkiksi kehitysmaissa. Tosin kaikkien tietojen kannalta maantieteellisellä katteella ei ole suurta merkitystä: jos kyse on kemiallisten yhdisteiden ominaisuuksista, ovat nämä samoja kaikkialla, jos vain tiedot on osattu kerätä tietokantaan maailman eri kolkista. *Tietokantojen sijaintipaikat* (myös saman tietokannan eri kopioiden) ovat myös tärkeitä, koska tietokannan käyttöä voi sen haltija kontrolloida paljon tarkemmin kuin kustantaja julkaisemiensa painotuotteiden käyttöä. Ylikansalliset tietovirrat saadaan nykyaikaisella tietoliikennetekniikalla kyllä vuolaiksi, mutta vain, jos maailmanpoliittinen tilanne on suotuisa. Viime vuosilta on esimerkkejä erilaisista vientirajoituk-

sista ja -kielloista esimerkiksi länsimaiden ja sosialististen maiden välillä. Toisaalta tietoliikennetekniikka tarjoaa mahdollisuuden »kiertää mutkan kautta» siten, että maa **A** hakee tiedot maassa **C** olevan bulvaanin kautta maassa **B** olevista tietokannoista, mikäli **A**:n ja **B**:n välit ovat huonot, mutta muut kahdenväliset suhteet kunnossa. Näiden mutkien kautta kulkevia tietovirtoja on vaikea valvoa. [Cawk-80] [Surp-85] Tällainen riski (**B**:n näkökannalta) voisi aiheuttaa ongelmia **B**:n ja **C**:n välisiin tietoliikenne- ja tietokantojen käyttösopimuksiin. Suomessakin on keskusteltu siitä, aiheuttavatko Neuvostoliiton parantuvat tietoliikenneyhteydet Suomeen mahdollisesti Suomen informaatio- palveluille ongelmia tietokantojen käyttösopimusten teossa amerikkalaisten tietokantojen myyjien kanssa. Valtaosa numeerisista tietokannoista, kuten viitetietokannoistakin, sijaitsee Länsi-Euroopassa ja USA:ssa.

Tietokantojen *ajallinen kate* on samantapainen kuin viitetietokannoissakin: pääosa niistä sisältää tietoja 1970-luvulta ja sitä myöhemmistä ajoista. Vanhemman aineiston hakemisessa paperijulkaisut ovat korvaamattomia. Esimerkiksi tilastotietokannoissa useimmat aikasarjat alkavat vasta 70-luvulta ja niiden *ajantasaisuus* oli monessa tapauksessa vain vastaavien paperijulkaisujen luokkaa [KaLM-85]. Jos vertaillaan paperijulkaisujen ja vastaavien tietokantojen *hakunopeutta* tilanteessa, jossa *molemmat ovat käytettävissä,* lienee paperijulkaisujen käyttö nopeampaa, jos haetaan tietoa, joka sellaisenaan on painettu julkaisuun, ja muulloin yleensä nopeampaa — jopa huomattavasti nopeampaa — tietokannasta.

Numeeristen tietokantojen tietojen *tekninen* saatavuus riippuu sopivien päätelaitteiden ja tietoliikenneyhteyksien olemassaolosta sekä tiedoista ja taidoista näiden sekä tiedonhallintajärjestelmien käytössä. Sekä tekniset, tiedolliset että taidolliset vaatimukset ovat selvästi suuremmat kuin tietokantojen paperivastineiden käytössä. Kynnysraha on korkeampi, joskin tarjolla olevat mahdollisuudetkin ovat suuremmat. Suurin kynnys pysyy kuitenkin entisellään: tiedot ja taidot tarpeiden määrittelyyn ja löydettyjen tietojen käyttöön tai käytön neuvontaan.

Numeeristen tietokantojen tiedot ovat kauppatavaraa. Siksi niiden käyttö maksaa vielä kynnysrahan tultua maksetuksi, mikä edelleen rajoittaa tietojen *taloudellista* saatavuutta. Kyselyjen kustannusten laskentaa tarkastellaan lähemmin luvussa 4. Kustunnustasosta numeeristen tietokantojen käytössä verrattuna paperijulkaisujen aiheuttamiin kustannuksiin ei ole tutkimuksia. Tietokantojen käytön hinnoittelu ja laajuus muuttuvat nyt niin nopeasti, ettei tällaisilla tasotiedoilla voine olla edes pitkää käyttöarvoa. Paperijulkaisun hankinta aiheuttaa kertakustannuksen hankinnan yhteydessä ja jatkuvia varastointikustannuksia (osuus henkilökunnan ja tilojen kuluista), kun taas tietokantojen käytöstä maksetaan vain käyttökertakohtainen kustannus.

Seuraavassa on tiivistelmä kahdessa kotimaisessa selvityksessä ([KaLM-85] [Auvi-85]) kerätyistä ulkomaisten numeeristen tietokantojen käyttökokemuksista:

- sekä saman myyjän että eri myyjien tietokantojen kesken esiintyy päällekkäisyyttä;
- tietokantojen sisältöä ja rakennetta ei ole suunniteltu erityisesti millekkään käyttäjäryhmälle; ne eivät siten välttämättä yksinään riitä tyydyttämään kenenkään tarpeita;
- tiedot eivät aina ole paperijulkaisuja täydellisempiä eivätkä tuoreempia; joissakin tapauksissa taas tietokannat ovat huomattavasti parempia tietolähteitä, jopa korvaten vastaavat hakuteokset;
- tietojen maantieteellinen kate on hyvä vain teollistuneiden länsimaiden osalta;
- tietojen käsittelymahdollisuudet tarjoavat merkittäviä etuja paperijulkaisuihin verrattuna;
- käyttäjän on hallittava monenlaisia ja monimutkaisia kyselyjärjestelmiä ja tunnettava niiden erikoispiirteet;
- kyselyjärjestelmien tarjoamat apuneuvot ovat kehittymättömiä; muistettavaa on paljon;

Kullakin ammattialalla on oma *työn tekemisen käytäntönsä.* Sitä pitävät ammattikunnat yllä koulutuksen, menettelytapasääntöjen, standardien, keskinäisen arvioinnin jne. avulla. Kunkin ammattikunnan keskuudessa työn käytäntö määrää mm. seuraavista seikoista [JäRe-83]:

- mitkä ovat alalle kuuluvia ongelmia tai tehtäviä,
- millä tavoin niitä lähestytään ja tarkastellaan,
- mitä tietoja ongelmasta tai tehtävästä tarvitaan,
- minkä tyyppisiä ammattialalla tunnettuja tietoja tulisi työssä käyttää apuna,
- mistä näitä tietoja kannattaa ja pitää hankkia, ja
- minkä tyyppisiin ratkaisuihin tai tuloksiin pyritään.

Työn käytäntö kehittyy lukuisten tekijöiden, kuten filosofisten, sosiaalisten, kulttuuristen, teknisten tekijöiden sekä luonnonolojen vaikutuksesta. Numeeriset tietokannat kuuluvat teknisten tekijöiden luomiin mahdollisuuksiin, mutta vaikuttavat välillisesti tai välittömästi kaikkiin työn käytännön puoliin. Työn käytännössä ilmeneviin *välittömiin vaikutuksiin* kuuluvat vaikutukset tiedonhankintaan ja *välillisiin vaikutuksiin* ne, jot-

ka liittyvät työn varsinaisten perustoimintojen suorittamiseen [JäRe-83].

Kullakin ammattikunnalla on *tiedonhankinnan* yleinen *käytäntönsä* ja kullakin ammattikunnan jäsenellä siitä muokattu oma tiedonhankinnan käytäntönsä. Molemmat muuttuvat tietojen saatavuuden, hankintapaikkojen ja -välineiden muuttumisen myötä. Yleisesti voidaan sanoa, että numeeriset tietokannat tarjoavat sekä uuden tavan *organisoida* tietoja että uuden tavan tai *välineen* hankkia niitä käyttöön. Ne ovat vaihtoehto esim. tilasto- ja hakemistojulkaisulle ja tulevat varmasti vähitellen muuttamaan totunnaisia tiedon hankintatapoja. Vielä ei ole tutkimustietoa siitä, miten nämä muutokset toteutuvat eri ammattikunnissa. [JäRe-82] [JäRe-83]

Tietotekniikan työn tekemisen käytäntöön kohdistuvien *välillisten vaikutusten* monimuotoisuudesta on esitetty malli [Järv-86a], joka kattaa työn päämäärät, perus- ja tukitoiminnan, sovellettavan tekniikan ja tiedot. Mallista voidaan johtaa vaikutushypoteeseja eriteltynä eri tehtävä-, ongelma-, tieto-, tiedonlähde-, työvälinesekä työskentelytapatyyppien suhteen. Vaikka numeeriset tietokannat vaikuttavatkin työn *tietoperustoihin* ja sitä kautta työn käytäntöön ja *tiedontarpeiden muotoutumiseen,* ovat vaikutukset niin monimuotoiset ja erilaiset eri tilanteissa, ettei niistä ole yleistyksiä esitettävissä. Seuraavassa kuitenkin tyypitellään vaikutuksia ja tarkastellaan tekijöitä, jotka vaikutuksia voimistavat, ja tekijöitä, jotka vaikutuksia rajoittavat.

Tietotekniikan vaikutukset työprosessiin voidaan tyypitellä seuraaviin asteittain voimistuviin vaikutuksiin: tehtävän suorittajan vaihtuminen (työvaiheen automatisointi sellaisenaan), työmenetelmän muutos tuloksen laadun pysyessä ennallaan, työn tuloksen tavoitetason muutos kohti optimaalista, työn tulosten arviointikriteerien muutos, työn perimmäisten päämäärien tai ongelmien muutos [JäRe-82] [Järv-86a]. Nämä vaikutukset liittyvät informatiikan kannalta keskeisiin ilmiöihin, kuten tiedon käyttö, tarpeet ja hankinta, ts. vaikutusten kohteena on koko tieteellinen ja ammatillinen kommunikaatio. Numeeristen tietokantojen käytön vaikutuksista yleensä voidaan tämän tyypittelyn valossa olettaa ainakin seuraavaa: 1) Numeeriset tietokannat korvannevat organisaatioissa erilaisten kortistojen ja rekistereiden ylläpitoon ja käyttöön kuuluvaa työtä (esim. tietojen keruuta ja tunnuslukujen laskemista). Tapahtuu siis suorittajan vaihdoksia. 2) Käsiteltävissä olevan numeerisen tietoaineiston saataville tulo aiheuttanee kokemukseen perustuvan harkinnan ja intuition käytön vähentymistä laskentamallien ja -tekniikoiden hyväksi *(miksi pohtia — kalkyloidaan!);* tapahtuu siis työmenetelmien muutoksia. 3) Laskentatekniikoiden käyttöön ottaminen merkitsee yleensä myös pyrkimystä mahdollisimman hyviin tai optimaalisiin tuloksiin työssä (esim. parempiin tuote-, tuotanto- tai markkinointisuunnitelmiin ja -päätöksiin). Tapahtuu siis myös työn tavoitetason muutoksia. Voidaan kuvitella myös tilanteita, joissa pelkästään, tai ainakin merkittävässä määrin, numeerisen tietoaineiston saataville tulo ja käyttö johtaisi työn tulosten arviointikriteerien tai päämäärien muutoksiin. Tällaiset tilanteet lienevät kuitenkin harvinaisempia. Vaikutusten yksityiskohtaisempi tarkastelu edellyttäisi sekä tarkasteltavien tietokantojen että niiden käyttötilanteiden täsmentämistä.

Numeeristen tietokantojen käyttöä lisäävät ja siten vaikutuksia voimistavat edellä tarkasteltujen etujen lisäksi seuraavat seikat: Ainakin periaatteessa numeerista tietokantaa voidaan sopeuttaa eri käyttäjäryhmille siten, että kukin saa itselleen relevantit tiedot mielekkäällä ja totutulla tavalla esitettynä. Lisäksi tietojen keruu ja organisointi käsiteltävissä oleviksi tietokannoiksi tukee tietoja jalostavien analyysiohjelmistojen kehittämistä. Näin voidaan lisätä järjestelmien kykyä tukea työntekoa [JäRe-83] [JäRe-84].

Numeeristen tietokantojen vaikutuksia rajoittavat monet muutkin seikat niiden ohella, jotka liittyvät tietojen saatavuuteen. Yleisesti ennen tietojen keruuta ongelmanratkaisua tai päätöksentekoa varten tulee ongelma jäsentää, ts. harkinta, *mistä ilmiöistä ja niiden suhteista ja mistä niiden piirteistä* (muuttujista, ominaisuuksista) ollaan kiinnostuneita, ja vasta tämän jälkeen seuraa numeerisen tai muun tietoaineiston (= muuttujien arvojen) keruu. Mitä aidommasta ongelmasta (määrittely, ks. [Elor-74] [KuRS-77]) on kyse, sitä keskeisempi merkitys on näillä tietojen keruuta edeltävillä ja sille vaatimuksia asettavilla vaiheilla. Numeeristen tietokantojen tarjoama tieto on ns. neutraalia, kovaa, ongelma- tai ongelma-aluetietoa (käsitteistä esim. [CaMS-75] [JäRe-83] [Järv-86a] [Rich-83]), eikä jäsentävää, ideoivaa, pehmeää, arvopitoista tai metodista tietoa, joilla on keskeinen merkitys ongelman jäsentämisessä. Vasta kun ongelma on pitkälle jäsennetty ja rajattu, eli mahdollisten ratkaisujen joukko on pitkälle rajattu, tulee tiedon keruun, esimerkiksi numeeristen tietokantojen käytön, aika. Isojen ongelmien ratkaisemisen tukemisessa ei yleensä saada paljon aikaan vain numeerisen tietoaineiston saatavuutta parantamalla. Tärkeimmät parannukset kohdistuvat tällöin jäsentävään, ideoivaan, pehmeään, arvopitoiseen ja metodiseen tiedon saatavuuteen ja käyttöön. Numeerisissa tietokannoissa ei tällaista aineistoa ole.

Muut vaikutukset voidaan tyypitellä seuraavasti: *yksilötasolla:* työn tulosten laatu ja niiden

tuottamisen kustannukset, työn kokoonpano, taitovaatimukset, stressi sekä työtyytyväisyys; *organisaatiotasolla:* muutokset päätöksenteossa, organisaatiorakenteissa, tuottavuudessa, organisaatioiden keskinäisissä ja organisaatioiden ja yleisön välisissä suhteissa; *yhteiskuntatasolla:* vaikutuksia työllisyydessä, kansantaloudessa, yhteiskunnan turvallisuudessa ja haavoittuvuudessa, tietosuojassa ja tasa-arvossa sekä poliittisessa osallistumisessa ja vallassa. Näitä tarkastellaan lukuisissa kansantaloudellisissa, yhteiskuntatieteellisissä ja tietojenkäsittelytieteellisissä tutkimuksissa, esim. [Alte-80] [AtRu-84] [Cron-85] [Lepp-85] [MåOW-84]. Nämä vaikutukset sivuutetaan.

3. Relaatiomalli

Nykyaikaiset tiedonhallintajärjestelmät tarjoavat käyttäjilleen ns. *korkean tason näkemyksen* tietokannan rakenteeseen, sisältöön ja käyttöön. Tämä tarkoittaa, ettei tietokannan käyttäjän tarvitse välttämättä tuntea tietojen todellista teknistä talletustapaa tietokannassa eikä niiden todellisia hakumenetelmiä. Tietokantojen kyselyjärjestelmät ja käyttäjäliitännät saadaan näin *käyttäjäystävällisiksi ja joustaviksi.* Tämä onkin perusedellytys sille, että tietokantoja voidaan tarjota julkiseen online-käyttöön. Tiedonhallintatekniikkaan tavallisesti perehtymättömiltä loppukäyttäjiltä tai välittäjiltä ei voida edellyttää teknisten yksityiskohtien hallintaa.

Relaatiomalli täyttää erityisen hyvin vaatimukset korkean tason käyttäjäliittymästä. Relaatiomallin kehitys alkoi 1970-luvun alusta [Codd-70] ja sittemmin se on saanut keskeisen aseman tietokantoihin liittyvässä teoreettisessa tutkimuksessa (esim. [Ullm-80]). Nykyisin on tarjolla useita relaatiomalliin perustuvia kaupallisia tiedonhallintajärjestelmiä. Se tarjoaa hyödyllisen lähestymistavan numeeristen tietokantojen luomiseen, kuten voidaan todeta useista CODATA-raportin [RuHa-84] artikkeleista. Sen käyttöä viitetietokantojen yhteydessä ja yleisemminkin kirjastoautomaatiossa on myös tarkasteltu useissa tutkimuksissa (esim. [Atki-79] [Craw-81] [Kurt-84] [McLe-77]). Esimerkiksi Crawford [Craw-81] esittää relaatiomallin eduiksi seuraavat:

- relaatiomallin *yksinkertaisuus* on vakuuttavaa: se tarjoaa yhdenmukaisen, yksinkertaisen ja selkeän näkemyksen tietokannasta;
- sillä on vankka *matemaattinen perusta:* kaikki tiedot ja niiden käsittely voidaan täsmällisesti määritellä joukko-opin ja logiikan avulla (esim. [NiJä-85]); tämän takia voidaan myös relaatiomalliin perustuvien tietokantojen ja niiden käytön ominaisuuksia täsmällisesti tutkia;
- käyttäjän kannalta *kaikkien tietojen haku ja käsittely* tapahtuu *yhdenmukaisella* tavalla;
- relaatiomallin tarjoama *tietoriippumattomuus* on huippuluokkaa: tiedot ja niiden käsittely kuvataan tavalla, joka on täysin riippumaton niiden teknisestä toteutustavasta.

Relaatiomallissa *tiedot esitetään* matemaattiseen relaation käsitteeseen perustuvina *relaatioina,* joita käyttäjille tavallisesti havainnollistetaan kaksiulotteisina taulukkoina. Kuvassa 1 esitetään kolmesta relaatiosta koostuva kuvitteellinen markkinatietokanta taulukkoina. Tässä tietokannassa kuvataan tuotteita (tuotenumero, tavaramerkki, tuotetyyppi, valmistaja ja myyntimäärä), tuottajia (tuottajanumero, tuottajan nimi, pääkonttorin sijaintipaikka, liikevaihto ja voitto) sekä markkinoita (tuotenumero, maa, vuosi ja markkinaosuus). Tiedoista ilmenee esimerkiksi, että tuote 1512 on tyyppiä 19500 (esim. partavesi), tavaramerkiltään Gillette, sen valmistajan valmistajanumero 260011 ja sitä myytiin tarkasteluvuonna noin puoli miljoonaa kappaletta. Valmistaja 260011 osoittautuu olevan GILLETTE UK, toimipaikaltaan Lontoo, ja sen liikevaihto on 10 miljoonaa. Tuotteen 1512 markkinaosuus esim. vuonna 1984 Kongossa näyttää olleen 10 %.

Taulukkoesitys on käyttäjille luonnollinen tapa relaatioiden esittämiseen. Numeerinen tietoaineisto, esim. tilastot, esitetään perinteisestikin juuri kaksiulotteisina taulukoina. Lisäksi taulukot muistuttavat rakenteelliselta kannalta läheisesti perinteisiä tiedostoratkaisuja, joita edelleen käytetään monissa tietokantaympäristöissä (relaatiomallin ulkopuolella). Tässä artikkelissa käytetään relaatioista nimitystä (epähierarkkinen) *tiedosto* (flat file) [Ullm-80], koska tarkastelu kattaa sekä relaatiot että niiden toteutuksen, joka tapahtuu tiedostoina relaation matemaattisen käsitteen alueen ulkopuolella.

Tietojen *käsittely määritellään* relaatiomallissa *relaatioalgebran tai relaatiokalkkyylin* (relational algebra, relational calculus) avulla. Relaatiomalliin perustuvien tiedonhallintajärjestelmien kyselykielet (esim. SQL [Astr-76] [AsCh-75] [Craw-81] [SaMc-83]) perustuvat näistä jompaankumpaan. Kyselykielten ilmaisut voidaan aina kääntää vastaaviksi relaatioalgebran ilmaisuiksi. Relaatioalgebralla määritellään relaatioiden käsittely täsmällisesti (esim. [Nijä-85] [Ullm-80]). Se on ns. *relationaalisesti täydellinen* kyselykieli, ts. sen avulla voidaan annetuista relaatioista (taulukoista) johtaa mikä tahansa relaatio, joka niistä periaatteessa on johdettavissa 1. kertaluokan predikaattilogiikan puitteissa. Havainnollisemmin sanottuna: voidaan johtaa mikä tahansa taulukko, joka annetuista lähtötau-

Kuva 1. Esimerkkitietokanta

PRODUCTS	(PRODUCT-NO,	TRADEMARK,	TYPE,	MANUF-NO,	QSALES)
	1512	GILLETTE	19500	260011	502000
	1586	GILLETTE	19190	260011	107000
	376203	BRAUN	19190	7005	408000
	95051	BLUE STRATOS	19500	530286	200000
	556556	TABAC	19500	1050	900000
	...	...			
	...				
	...				

COMPANIES	(COMPANY-NO,	C-NAME,	HQ-LOC,	TURNOVER	REVENUE)
	260011	GILLETTE UK	LONDON	10000000	800000
	7005	BRAUN AG	FRANKFURT	5000000	500000
	530286	SHULTON LTD	NEW YORK	8000000	200000
	1050	MÄURER+WIRTZ	STOLBERG	7000000	1000000
	...	...			
	...				
	...				

MARKETS	(PRODUCT#,	COUNTRY,	YEAR,	MARKET-SHARE)
	1512	TAHITI	1984	0.5
	1512	CONGO	1984	0.1
	1512	CONGO	1983	0.12
	95051	TAHITI	1984	0.02
	556556	BURMA	1983	0.75
	556556	TAHITI	1984	0.42
	...	...		
	...			
	...			

lukoista periaatteessa voidaan tuottaa rivejä ja sarakkeita leikkelemällä ja yhdistelemällä edellyttäen, että kaikki tulostaulukon rivit sisältävät samat sarakkeet. Tässä artikkelissa rajoitutaan tarkastelemaan relaatioalgebraa tietokantojen kyselykielenä. Seuraava yksinkertainen esimerkkikysely tosin esitetään myös SQL-kielellä.

Oletetaan, että relaatiomalliin perustuvan numeerisen tietokannan käyttäjä haluaa tietää, mitkä olivat eri partavesien markkinaosuudet Tahitilla vuonna 1984. Oletetaan lisäksi, että tuotetyyppi -tiedoissa TYPE = 19500 on partavesien tuotetyyppi. Vastaukseen tulee saada tiedot tuotenumerosta, tavaramerkistä ja markkinaosuudesta. Tämä hyvin yksinkertainen kysely voidaan esittää SQL-kielen avulla seuraavasti:

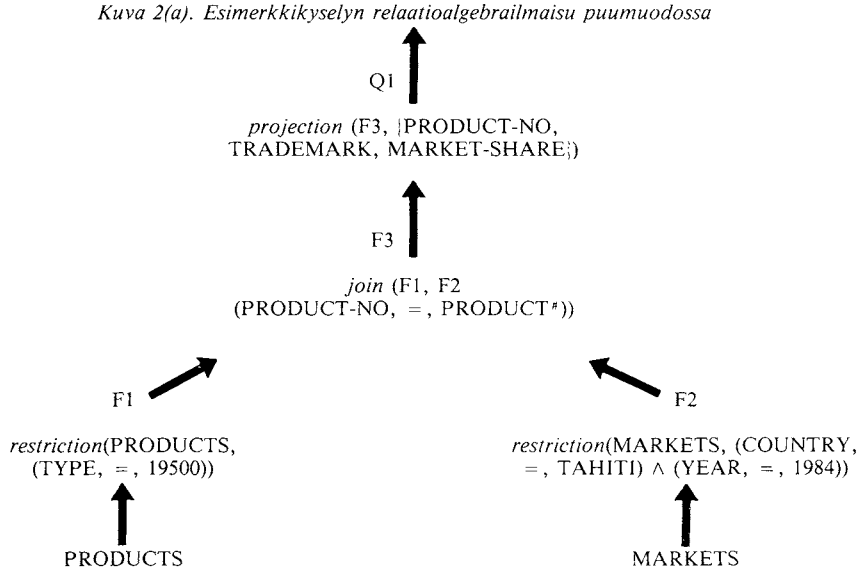
```
select  PRODUCT-NO, TRADEMARK, MARKET-
        SHARE
from    PRODUCTS, MARKETS
where   PRODUCTS.TYPE = 19500 and
        MARKETS.COUNTRY = TAHITI and
        MARKETS.YEAR = 1984 and
        PRODUCTS.PRODUCT-NO =
        MARKETS. PRODUCT#
```

Tässä **select**-osa määrittelee tulokseen tulevat sarakkeet *(attribuutit* relaatiomallin sanastossa), **from**-osa ne taulukot, joista tiedot poimitaan, ja **where**-osa tietojen poiminta- ja yhdistelyehdot. Poimintaehtoja ovat PRODUCTS.TYPE = 19500, MARKETS.COUNTRY = TAHITI ja MARKETS.YEAR = 1984, joista kukin ilmaisee, mitä relaatiota ja mitä sen attribuuttia poiminta koskee, ja mitkä attribuutin arvot ovat sallittuja. Ehto PRODUCTS.PRODUCT-NO = MARKETS.PRODUCT# on yhdistelyehto, joka ilmaisee, että relaatioiden PRODUCTS ja MARKETS rivit tulee yhdistää siten, että niillä olevat tuotenumerot täsmäävät. Relaatioalgebralla esitettynä sama kysly näyttäisi seuraavalta:

> *projection (join (restriction* (PRODUCTS, (TYPE, =, 19500)), *restriction*(MARKETS, (COUNTRY, =, TAHITI) ∧ (YEAR, =, 1984)), (PRODUCT-NO, =, PRODUCT#)), {PRODUCT-NO, TRADEMARK, MARKET-SHARE}).

Kuvassa 2(a) tämä ehkä hiukan hankalalta näyttävä ilmaisu on purettu puumuotoon, josta paremmin ilmenevät käytettävät eri operaatiot, niiden parametrit ja järjestys. Puun lehtinä ovat esimerkkitietokannan tiedostot (relaatiot) PRODUCTS ja MARKETS. Seuraavalla tasolla ovat näitä tiedostoja käsittelevät *rajoitusoperaatiot* (restriction), joiden parametreina ilmoitetaan kohdetiedoston nimi ja *rajoitusehto* (restriction predicate). Vasemmanpuoleinen rajoitus poimii PRODUCTS-tiedostosta ne tietueet, joiden tuotetyyppiattribuutin TYPE arvo on 19500 (partavesi). Tuloksena on välitiedosto **F1.** Oikeanpuoleinen rajoitus poimii MARKETS-tiedostosta ne tietueet, jotka koskevat TAHITIa ja vuotta 1984 (symboli ∧ tarkoittaa samaa kuin **and**). Tuloksena on välitiedosto **F2**. Seuraavaksi tehdään *liitosoperaatio* (join), jolla liitetään välitiedostojen **F1** ja **F2** tietueet toisiinsa tuotenumerojen mu-

Kuva 2(a). Esimerkkikyselyn relaatioalgebrailmaisu puumuodossa



Kuva 2(b). Esimerkkikyselyn tulos

Q1	(PRODUCT-NO,	TRADEMARK,	MARKET-SHARE)
	1512	GILLETTE	0.5
	95051	BLUE STRATOS	0.02
	556556	TABAC	0.42
	...	...	
	...		
	...		

kaan; näin saadaan liitettyä tuotetyyppitietoihin niiden markkinoita koskevat tiedot. Liitoksen parametreina ovat välitiedostojen nimet ja *liitosehto* (join predicate). Lopuksi kyselyssä on *projektio-operaatio* (projection), jolla liitoksen tulos karsitaan sisältämään vain ne attribuutit, joita käyttäjä pyysi vastaukseen: PRODUCT-NO, TRADEMARK ja MARKET-SHARE. Parametreina ovat liitoksen tuottama välitulos **F3** ja haluttujen attribuuttien nimet. Projektio tuottaa kyselyn tuloksen **Q1**, joka sekin on tiedosto. Kuvassa 2(b) on osa siitä.

Kaikkia numeerisia tietokantoja ei koskaan tultane toteuttamaan relaatiomallin tarjoamalta pohjalta. Kun tässä artikkelissa tarkastellaan numeeristen tietokantojen käyttökustannusten ennustamista relaatiomallin avulla, eivät saavutettavat tulokset ole sellaisenaan sovellettavissa kaikkiin numeerisiin tietokantoihin. Sellaisenaan ne käyvät vain relaatiomalliin perustuviin numeerisiin tietokantoihin, ja lähes sellaisenaan perinteisiin ei-hierarkisiin tiedostoihin perustuviin tietokantoihin. Muihin malleihin perustuvien tietokantojen osalta saadaan osoitettua, mitä kyselykustannusten ennustamisessa on saatavissa aikaan ja minkä tyyppinen lähestymistapa ennustamismenetelmien kehittämisessä tarvitaan. *Relaatiomalli on vankan teoreettisen perustansa takia erityisen sovelias tarkatelukehykseksi, jossa voidaan demonstroida käyttökustannusten ennustamismetodologia ja sillä saavutettavat tulokset täsmällisesti.* Relaatiomallin puitteissa on myös tehty paljon sellaista tutkimusta, jota voidaan *soveltaa* käyttökustannusten ennustamismenetelmissä. Muut tietomallit eivät tarjoa näitä etuja.

4. Käyttökustannusten laskenta

4.1. Perinne ja kehityslinjat

Tavallisin viitetietokantojen käyttökustannusten määräytymisperuste on jo perinteisesti ollut kyselyn yhteysaika, joka alkaa käyttäjän avatessa jonkin niistä tiedostoista, jotka kyselyjärjestelmä tarjoaa käytettäväksi. Yhteysajan perusteella määräytyvästä summasta on osa mennyt tietokannan (tiedoston) tuottajalle royalty-maksuna ja osa on jäänyt tietokannan (tiedoston) myyjälle

toiminnan kulujen katteeksi ja voitoksi.

Tämän käyttökustannusten laskentatavan *ongelmana* on se, että kustannus ei määräydy sen perusteella, mistä loppukäyttäjä varsinaisesti on halukas maksamaan ja minkä takia hän itse tai välittäjän avustamana tietokantaa käyttää, vaan toisarvoisen ja edellisistä riippumattoman tekijän nojalla. Harva käyttänee tietokantoja nykyisin vain huvin tai statuksen vuoksi. Loppukäyttäjää kiinnostavat tietokannasta saatavissa oleva hyöty, tulokset, kirjallisuusviitteet, ja niitä hänelle viimekädessä myydäänkin. Yhteysaikalaskutuksessa käyttökustannukset eivät kuitenkaan riipu haun tuottamasta tuloksesta — lasku kasvaa, vaikkei yhtään relevanttia viitettä löytyisi. Tämä tilanne ohjaa tietokannan käyttäjää suoriutumaan tiedonhausta mahdollisimman joutuisasti, jopa saatavien tulosten kustannuksella. Aikaavieviä kyselyjen muotoilutapoja, jotka mahdollisesti johtaisivat onnistuneeseen hakuun, ei suosita. Käyttäjä suostuu tulostamaan myös epärelevantteja viitteitä, siis roskaa, kunhan vain joukosta löytyy muutama relevantti viite ja kaikki saadaan äkkiä paperille. *Kyselyjärjestelmien koko potentiaalia kyselyn asteittaiseen ja vuorovaikutteiseen mahdollisimman hyvään muotoiluun ei käytetä hyväksi.* Tämä on käyttökustannusten laskentatavan luonnollinen seuraus. Tämän takia yhteysaikalaskutus on *tuloksellisen tiedonhaun vastainen* (counter-productive). [DuBo -81] [Hull-84]

Yhteysaikalaskutuksen vaikutukset ilmenevät myös useissa viime vuosien tutkimuksissa, joissa on pyritty kehittämään tietokantojen käyttäjille tiedonhaun apuvälineitä. Nämä apuvälineet ovat *oikoteitä* hakukomentojen kirjoittamisessa (esim. komentojen ryhmitys, termin katkaisu, termiryhmähaku [BoMC-84] [Henr-80]) tai *palveluja,* jotka suorittavat osan kyselyn määrittelyyn ja järjestelmän käyttöön kuuluvista tehtävistä käyttäjän puolesta (esim. automaattinen yhteyden luonti, tiedoston valinta tai hakukomentojen muotoilu [ClCr-83] [Poll-85] [Marc-83]). Usein näiden oikoteiden ja palvelujen kehittämisen yhtenä keskeisenä motiivina esitetään yhteysajan lyhentäminen ja rahan säästäminen tällä tavalla (esim. [ClCr-83] [Poll-85] [Marc-83]). Osaltaan käyttökustannusten laskutusratkaisut vaikuttavat siihen, mihin alan tutkimuksessa pyritään.

Viime vuosina viitetietokantojenkin käyttökustannusten laskentaperusteita on kehitetty siihen suuntaan, että yhteysajan osuutta alennetaan ja tulostettavien viitteiden määrän osuutta loppusummasta lisätään. Kunkin viitteen kustannusvaikutus voi vielä riippua sen tulostusmuodosta: jos tulostetaan esim. vain nimekkeitä haun muotoiluvaiheessa, tai vain suppeat bibliografiset tiedot, on kustannus pienempi kuin, jos tulostetaan täydelliset tiedot; usein online-tulostus, josta ei voida identifioida dokumenttia, on ilmainen (yhteysaikaa tietenkin kuluu). [Aitc-84] [Henr-80] [Inte-84] [KaWa-85] [Will-82] Tämä menettelytapa toimii viitetietokannoissa, koska kyselyn tuloksena on aina viitteitä, muodossa tai toisessa, jolloin viite muodostaa luonnollisen laskutusyksikön. Joissakin artikkeleissa esitetään viitteiden tai tekstin tulostuskustannusten perustamista tulostettaviin kenttiin [Hull-84] [Hunt-84]. Tulostettavien tietojen perusteella muodostuvaa käyttökustannusten laskentatapaa kutsutaan tässä artikkelissa *tulosperusteiseksi* laskentatavaksi.

Tulosperusteisen laskentatavan ohella myös viite- ja tekstitietokannoissa on harkittu *käsittelyperusteisen* käyttökustannusten laskentatavan soveltamista. Tällöin käyttäjä maksaisi *käyttämistään, erihintaisista resursseista,* kuten tietokoneen keskusyksikköaika, tukimuisti, kirjoittimet, erilaiset ohjelmat ja palvelut, niiden käytön mukaan. Tällöin miettiminen ei maksaisi, eikä sekään, että mahdollisesti joutuu odottamaan tietokoneen palvelua muiden käyttäjien aiheuttaman ruuhkan takia. Tällainen käyttökustannusten laskentatapa edellyttää kuitenkin mahdollisuutta ennustaa kustannukset ennen hakua, eikä tällainen ennustaminen ole helppoa. [Hull-84]

Tiedonvälitys tulee edelleen kaupallistumaan tietoyhteiskunnan (esim. [Cron-85] [Koch-83] [Kort-85] [MåOW-84] [Savo-84]) kehittymisen myötä: tietoyhteiskunnassa tiedon myynti ja ostaminen ovat keskeisiä toimintoja. Tämä aiheuttanee sen, että kaikkien tietokantojen käyttökustannusten laskentatapoja kehitetään tulos- ja käsittelyperusteisiin suuntiin, jotka varsin yksimielisesti on todettu yhteysaikalaskutusta paremmin tiedonhakuun soveltuviksi. Kehitys johtanee Lancesterin ennakoimien *online neuvontapalvelujen* (online referral centres) tarjontaan tiedonhaun avuksi. Tällaisesta palvelusta tiedonhakija saisi neuvoja siitä, mistä hänen tarvitsemiaan tietoja kannattaa hakea ja mitä niiden hakeminen eri tahoilta todennäköisesti tulisi maksamaan [Lanc-78].

4.2. Numeeristen tietokantojen vaatimukset

Nykyisin numeeristen tietokantojen käyttökustannusten laskenta perustuu eri yhdistelmiin seuraavista tekijöistä: yhteysaika, eri resurssilajien kulutus (tietokoneen keskusyksikköaika, levytilan käyttö, tulostuspalvelujen käyttö, eri palvelujen, kuten grafiikka- yms. erikoisohjelmistojen käyttö), hallinnolliset yleiskulut, vuosimaksut ja tietoliikennemaksut [Fost-84]. Kuten luettelosta näkyy, on käsittelyperusteinen käyttökustannusten laskentatapa jo mukana laskentaperus-

teissa. Tämä on välttämätöntä, koska joidenkin kyselyjen suorittaminen kuluttaa, ehkä lyhyenkin yhteysajan puitteissa, runsaasti tietokonelaitteiston resursseja ja viivyttää muiden kyselyjen käsittelyä. Yhteysaika on jokseenkin riippumaton kyselyn kuluttamien resurssien määrästä. Aivan samoin perustein kuin viite- ja tekstitietokantojenkin yhteydessä, myös numeeristen tietokantojen yhteydessä yhteysajan käyttö pääasiallisena laskutusperusteena on haitallista. Aivan samoin perustein on myös tulosperusteinen käyttökustannusten laskentatapa toivottava. Tällöin on kuitenkin mahdotonta käyttää viitetietokantojen tapaan kiinteitä, tai tulostusformaatista riippuvia yksikköhintoja tulostettaville tietueille. Tämä johtuu seuraavista tekijöistä:

- Numeerisissa tietokannoissa *ei ole yhtä* sellaista luonnollista *kohdetta,* kuten dokumentti viitetietokantojen yhteydessä, *jota koskevia tietoja aina haettaisiin* ja tulostettaisiin. Päinvastoin: käyttäjät ovat kiinnostuneita vaihtelevista kohteista vaihtelevin tavoin. Luvun 3 esimerkkitietokannasta voidaan hakea tietoja, jotka koskevat tuotteita, tuottajia, markkinoita, tuotetyyppien myyntimääriä, eri paikkakuntien tuotantotoiminnan profiilia tai yritystoiminnan kannattavuutta, jne.
- Samankin kohteen suhteen voidaan olla kiinnostuneita hakemaan numeerisista tietokannoista varsin *erilaisia tietoja ja eri laajuudessa.* Esimerkiksi viitetietokantojen tarjoamille painovuosi- ja luokitustiedoille ei ole juuri käyttöä ilman viitteiden sisältämiä muita tietoja. Numeeristen tietokantojen yhteydessä ei voida yhtä helposti sanoa, mitä tietoja välttämättä tulee hakea yhdessä, ja mitkä irrallisina ovat vailla merkitystä.
- Kaikki tietoalkiot *eivät ole* numeerisissa tietokannoissa *yhtä arvokkaita,* kun taas viitteisiin sisältyvät tiedot ovat kaikki jokseenkin tasa- ja vähäarvoisia. Eroja tietoalkioiden arvoon ja hintaan aiheuttavat toisaalta tietokannan *tuottamiseen* liittyvät seikat ja toisaalta sen *käyttäjiin liittyvät seikat.* Edellisiin kuuluvat ainakin erot tietojen saatavuudessa tuottajan käyttöön ja erot niiden keruun vaivalloisuudessa ja kalleudessa. Jälkimmäisiin kuuluvat erot tietojen kysynnässä, luotettavuudessa, täydellisyydessä ja kattavuudessa, sekä erot niiden arvossa ja käyttökelpoisuudessa käyttäjien tarpeisiin.

On ilmeistä, että jotkut tietoalkiot numeerisissa tietokannoissa ovat 'kuumia' ja siksi mahdollisesti kalliita. Tietokannan käyttäjä voi sisällyttää ne kyselynsä tuloksiin tai jättää ne pois. Joidenkin tietojen käyttökelpoisuus loppukäyttäjälle johtuu siitä, että niiden avulla voidaan *suoraan ratkaista* ongelmia ja suorittaa muita tehtäviä. Toisten tietojen käyttökelpoisuus voi perustua niiden kykyyn *yksilöidä* kohteitaan, esimerkkinä tuotenumero, henkilötunnus, kaupparekisteritunnus, jne. Tällöin niiden avulla voidaan yhdistää eri tahoilta kerättäviä tietoja laajemmiksi kokonaisuuksiksi (tämän kolikon toisena puolena ovat tietosuojaongelmat). Kolmanneksi tietojen käyttökelpoisuus lisääntyy, jos ne tarjoavat *yhteydenottomahdollisuuden* kohteeseen, esimerkkinä osoitetiedot.

Mikäli luovutaan tietue- ja formaattiperusteisesta kustannusten laskutavasta ja *hinnoitellaankin* näiden sijasta *erikseen kukin* tietokannan eri tiedostoissa esiintyvä *attribuutti,* voidaan tulosperusteinen kustannus laskea kyselyille. Tämä on suoraviivainen laajennos formaattiperusteisesta hinnoittelusta: riippuuhan siinäkin hinta tulostettavien tietojen täydellisyydestä ts. tulostettavien tietoalkioiden määrästä, eikä ole juuri sen vaikeampaa kuin viite- ja formaattiperusteinen hinnoittelu viitetietokannoissa. Kun viitetietokannassa määritellään viitteen hinnaksi 100 penniä, niin vastaavasti voidaan numeerisessa tietokannassa määritellä yhden tuotenumerotiedon hinnaksi 20 penniä, yhden tuotetyyppitiedon hinnaksi 15 penniä, yhden liikevaihtotiedon hinnaksi 150 penniä, jne. Näiden tietojen hintataso voidaan määritellä tutkimalla toisaalta tietojen keruun ja organisoinnin kustannuksia ja taloudellisia tavoitteita ja toisaalta tietojen kysyntää ja markkinoita. Kun näin menetellään, voidaan kyselyn tuloksesta tutkia, kuinka monta minkin attribuutin tietoalkiota tulokseen sisältyy, ja sitten laskea kyselyn tulosperusteinen kustannus viitetietokantojen käytäntöä vastaavalla tavalla. Tämä menettelytapa kattaa tietue- ja formaattiperusteisen laskutavan yhtenä erikoistapauksena, mutta pystyy ottamaan huomioon sekä eri kohteita koskevien tietojen vaihtelevat hinnat että tulostettavien attribuuttien lukumäärän ('formaatin') vaihtelun. Lasku määräytyy vain niiden tietojen perusteella, jotka tulos sisältää. Tällaista menettelytapaa voidaan soveltaa myös viitetietokannoissa. Se *ei ole tuloksellisen tiedonhaun vastainen.*

Tarkastellaan esimerkin vuoksi luvun 3 esimerkkitietokantaa ja -kyselyä. Oletetaan, että eri attribuuttien tietoalkioille on annettu seuraavat hinnat:

• PRODUCT-NO:	100	COMPANY-NO:	80
• TRADEMARK:	50	C-NAME:	80
• TYPE:	40	HQ-LOC:	20
• MANUF-NO:	80	TURNOVER:	120
• QSALES:	120	REVENUE:	120
• PRODUCT#:	100	COUNTRY:	60
• YEAR:	60	MARKET-SHARE:	150

Nyt voidaan laskea esimerkiksi, että yhden kokonaisen tietueen hinta PRODUCTS-tiedostossa on 390 yksikköä[4] ja MARKETS-tiedostossa 370 yksikköä. Esimerkkikyselyn tuloksessa kunkin tietueen hinta on 300 yksikköä. Jos kyselyn koko tulos sisältäisi esimerkiksi 30 tietuetta, muodostuisi sen hinnaksi 9000 yksikköä.

4.3. Syntyvät ongelmat

Tarkastellaan seuraavaksi tulos- ja käsittelyperusteisten tietokannan käyttökustannusten laskentatapojen aiheuttamia ongelmia, niiden syitä ja merkitystä online tiedonhaun eri osapuolille. *Pääongelma on seuraava: koska voidaan osoittaa, ettei pelkästään tietokannan käyttäjän esittämän kyselyn perusteella voida ennustaa kyselyn kustannuksia ja että nämä kustannukset voivat vaihdella varsin suurissa rajoissa, käyttäjän tulee voida etukäteen tietää, mitä kysely tulee maksamaan.* Ennuste kustannuksista tarvitaan kyselyä muotoiltaessa, siis ennen kyselyn suoritusta [Hull-84] [Hunt-84]. Tällöin ennustetta voidaan käyttää hyväksi kyselyn muotoilussa. Käyttäjät ovat kustannustietoisia eivätkä halua sitoutua etukäteen tuntemattomiin kustannuksiin tuolloin vielä tuntemattomien tulosten hankkimisessa — he eivät halua ostaa sikaa säkissä.

Kyselyn *muotoilussa* kustannusennusteita, ja niiden myötä mahdollisesti saatavia muita tietoja (esim. tuloksen koko), voidaan käyttää kyselyn laajentamiseen ja/tai supistamiseen viitetietokannoista tutulla tavalla. Mikäli odotettavissa olevat tulokset osoittautuvat liian kalliiksi, voidaan ennusteen perusteella luopua koko kyselystä. Ennusteita voidaan käyttää myös *myyjien valinnassa:* kuten viitetietokannatkin niin myös monet numeeriset tietokannat (ja vielä useammin ainakin vastaavat tiedot) ovat saatavissa useamman myyjän välityksellä. Usein tällaiset eri myyjien tietokannat ovat käytettävissä eri laajuudessa, eri tavoin organisoituina tai eri tavoin hinnoiteltuina. Jos saatavilla on ennusteita eri tietokantojen käyttökustannuksista ja tulosten laajuudesta, voidaan kysely kohdistaa sille myyjälle, jolta saatavien tietojen hinta/laatu -suhde on sopivin. *Ennusteiden puute* tai *epäonnistuneet ennusteet* johtavat helposti valituksiin, jotka kohdistuvat myyjiin tai heidän tarjoamiinsa kyselyjärjestelmiin. Esimerkiksi NEXIS-tekstitietokannassa sovelletaan käsittelyperusteista laskutusta, eikä kyselyjen kustannuksista ole saatavissa ennusteita. Tämä on johtanut arvosteluun ja kehittämisehdotuksiin jopa tieteellisten aikakauslehtien tasolla [Lomi-85]. Mikäli ennusteita kustannuksista tarjotaan, tulee käyttäjällä olla takeet siitä, että ennusteet ovat kohtuullisen tarkkoja todellisiin kustannuksiin verrattuna tai että myyjä sitoutuu laskuttamaan ennustetun kustannuksen, olipa todellinen kustannus mikä tahansa.

Numeeristen tietokantojen käyttökustannusten suureen *vaihteluun vaikuttavat* olennaisesti seuraavat *tekijät:* kyselyjen ominaisuudet ja tietokannan kunkinhetkinen todellinen sisältö, sekä tiedostorakenteet. Nämä tulee ottaa huomioon, mikäli kustannukset aiotaan ennustaa. Seuraavassa tarkastellaan kutakin näistä tekijöistä.

Kyselyjen ominaisuudet ja tietokannan kunkinhetkinen todellinen tietosisältö määräävät, kuinka suurta tietojen volyymiä kysely käsittelee ja kuinka suuren tuloksen se tuottaa. Ilmiö on tuttu jo viitetietokannoissa: mitä suppeammin kysely rajataan tarkalla (specific) ja tyhjentävällä (exhaustive) ilmaisulla (kyselyn ominaisuus), sitä vähemmän viitteitä poimitaan viitetiedostosta ja sitä pienempi on kyselyn tulos (esim. [Lanc-68]). Viitetietokannoissa voidaan kunkin kyselyn saamien *osumien* (hits) lukumäärä yleensä määritellä käänteistiedoston avulla tarkasti ennen viitteiden poimintaa. Määrä riippuu kyselyn lisäksi siitä, kuinka monta viitettä tiedostossa kullekin kyselyssä esitetylle termille on kirjattu. Tämä onnistuu, koska kyselyt ovat yksivaiheisia. Kokeneet käyttäjät voivat osumien perusteella arvioida myös tarvittavan yhteysajan ja siihen perustuvan kyselyn kustannuksenkin, joka tällaisissa tilanteissa riippuu pääasiassa tietoliikenneverkon ja kyselypäätteen nopeudesta. Numeerisissa tietokannoissa tällainen ei tavalliseti onnistu: tavallisesti kyselyt ovat monivaiheisia eikä tällöin ole yhtä sellaista käänteistiedostoa, josta vastauksen koon l. *kardinaalisuuden* (cardinality) [BGWR-81] [PiCo-84] tai siihen poimittavien tietueiden talletusosoitteet voisi tarkistaa. Tämän takia ei kyselyn välitulosten tai lopputuloksen kokoja voida täsmällisesti määritellä eikä sen arviointikaan ole helppoa (esim. [BGWR-81] [Chri-83] [Järv-86b] [PiCo-84] [YuLi-82]). Niiden arviointi edellyttää tietoa siitä, miten kyselyssä käsiteltävien attribuuttien arvot ovat jakaantuneet vaihteluvälillään ja miten tietueet ovat jakaantuneet näiden arvojen kesken tietokannassa.

Tiedostorakenteet määräävät tiedostojen käsittelykustannukset osittain riippumattomasti poimittavien ja käsiteltävien tietojen määrästä. Tiedostorakenteet nimittäin määräävät, millaiset tietojen *poimintastrategiat* (access strategies) ovat mahdollisia kunkin kyselyn suorittamisessa (esim. [TeFr-82] [Wied-77]). Tästäkin löytyy esimerkkejä viitetietokantojen käytössä [Henr-80]. Oletetaan, että viitetietokannan jonkin viitetiedoston viitteistä on käänteistiedosto laadittu ainoastaan nimeke- ja indeksitermikenttien perusteella. Jos nyt haetaan viitteitä nimekkeen sanoil-

la tai indeksitermeillä, käy haku joutuisasti käänteistiedoston välityksellä: käänteistiedostosta voidaan määritellä osumat ja viitetiedostosta tarvitsee poimia vain niitä vastaavat viitteet. Mutta jos ollaankin kiintostuneita tietyn henkilön kirjoittamista dokumenteista, käykin haku hitaasti: koska tekijäkenttää ei ole otettu käänteistiedostoon, ei ole muuta mahdollisuutta kuin lukea koko viitetiedosto alusta loppuun läpi ja tutkia joka viitteestä, onko haetun henkilön nimi tekijäkentässä. Toisin sanoen, muutaman tietueen hakeminen voi vaatia koko tiedoston tutkimisen. Yhdessä laitteiston ominaisuuksien kanssa tiedostorakenteet määräävät tiedostojen käsittelykustannukset. Tilanne on aivan sama numeerisissa tietokannoissa. Viitetietokannoissa tietueiden poiminta tiedostoista ei yleensä ole kyselyjärjestelmän pullonkaula. Numeerisissa tietokannoissa kuitenkin tiedostorakenteet vaihtelevat tiedostosta toiseen ja tietojen poimintastrategiat kyselystä toiseen. Tietojen yhdistely useasta tiedostosta vaatii usein niiden käsittelyä määrätyssä järjestyksessä ja käsittelyn synkronointia; tämä voi edellyttää esimerkiksi tietueiden lajittelua. Tietojen poimintastrategia voidaan usein valita monen vaihtoehdon joukosta — eikä tämä valinta ole yksioikoinen — ja kyselyn kustannukset riippuvat tästä valinnasta varsin suuressa määrin (esim. [BGWR-81] [SACL-79] [SmCh-75] [TeFr-82] [Wied-77] [Yao-79]). Numeerisissa tietokannoissa tietojen varsinainen poiminta tukimuistista ja niiden käsittely muodostavat usein suuren, mutta sangen vaihtelevan osan koko kyselyn vaatimista resursseista.

Näitä seikkoja *ei* tietokannan *käyttäjien,* jotka eivät ole tiedonhallinnan asiantuntijoita, *voida edellyttää hallitsevan.* Vaikka heille nämä tiedot tarjoittaisiinkin, ei heidän voitaisi kohtuudella edellyttää tietävän, *miten* niitä käytetään kyselyn kustannusten ennustamiseen. Tehtävä on asiantuntijoillekin vaivalloinen, ellei käytettävissä ole sopivia apuvälineitä. Tiedonhallintajärjestelmien avulla on tietoisesti pyritty helpottamaan tietokantojen käyttöä niin, ettei käyttäjien tarvitse tuntea tietojen teknistä talletustapaa ja hakumenetelmiä. On suhteellisen yksinkertaista esittää kysely, johon vastaaminen on hyvin vaivalloista. On välttämätöntä tehdä kyselyjen esittäminen näin helpoksi, mutta samalla käyttäjältä menee periaatteellinenkin mahdollisuus käyttökustannusten ennustamiseen ennen kyselyn suorittamista. Siksi on välttämätöntä kehittää apuvälineitä käyttökustannusten ennustamiseen.

Vaikka kyselyjen yhteysaikaan perustuva kustannusten laskeminen johtaakin monessa suhteessa haitallisiin vaikutuksiin, ovat sen vaihtoehdot, tulos- ja käsittelyperusteinen kustannusten laskentapa, myös ongelmallisia tiedonhaun kaikille osapuolille. *Loppukäyttäjän* kannalta ongelma syntyy suoranaisesti epävarmuudesta kyselyn toteutuvien kustannusten suhteen. Kustannustietoisuus ja tulosvastuullisuus panevat aprikoimaan kyselyn järkevyyttä ja tarpeellisuutta: kustannuksia ja tuloksen hyötyä tulisi voida punnita, mutta kummastakaan ei ole selvää kuvaa ennen kyselyn suoritusta. Kun tietokannan käyttö vielä vaatii vaivannäköä, joko kyselyn omatoimisessa suorittamisessa tai tarpeiden selittämisessä välittäjälle, lisääntyy mahdollisten loppukäyttäjien taipumus jättää tietokannat käyttämättä. Usein samantapaisia tietoja on tarjolla muillakin tahoilla, ei ehkä yhtä täydellisinä, tarkkoina ja käyttökelpoisina, mutta kuitenkin riittävästi. Usein voidaan myös tulla toimeen sormituntumalla, kokemuksen tai epätäydellisten tietojen varassa. Tällaisia loppukäyttäjän harkintaan kuuluvia seikkoja sekä tiedonhankintatapojen muotoutumista ja tietokantojen vaikutusta niihin tarkastellaan lähemmin useissa julkaisuissa (esim. [CaMS-75] [Järv-81] [Järv-86b] [JäRe-83] [KuRS-77] [Rich-83]). Epätietoisuus aiheutuvista kustannuksista lisännee etenkin niiden, joilla on vähiten rahaa ja muita voimavaroja käytettävissä työssään ja vähiten kokemusta tietokantojen käytöstä, *haluttomuutta* tietokantojen käyttöön. Kehitysmaissa, kehitysalueilla ja yleensä haja-asutusalueilla tämä tullee selvästi esiin: näin ollen kulttuuri- ja tiedontasoerot näiden ja kehittyneiden maiden tai alueiden kesken kasvavat ja tasa-arvo-ongelmat kärjistyvät.

Välittäjille tulos- ja käsittelyperusteinen kustannusten laskentatapa aiheuttaa ongelmia asiakkaiden hankinnassa ja palvelujen markkinoinnissa. Etenkin niille kirjastoille, jotka laskuttavat asiakkailtaan kyselyjen kustannukset ja joiden asiakaskunta on vaihteleva ja epämääräinen, on tärkeää tietää kyselyjen kustannukset jo ennen kyselyn suorittamista. Muutoin voi olla vaikeaa, jopa mahdotonta, »myydä» hakua mahdolliselle loppukäyttäjälle. Toisaalta kirjastojen omat budjetit ovat niin tiukkoja, etteivät kirjastot voi ottaa suuria riskejä kyselyjen maksamisessa. Ongelma on samanlainen kaikentyyppisten tietokantojen käytössä. [Hunt-84]

Ennakoimattomissa olevat kustannukset ovat ongelma myös tietokannan *myyjälle* ja *tuottajalle.* Siinä määrin kuin tulos- ja käsittelyperusteinen kustannusten laskentatapa estää mahdollisia käyttäjiä käyttämästä tietokantaa, siinä määrin vähenee myyjän ja tuottajan tulojen kertymä. Toisaalta halpojen hintojen ohella on käyttäjän mahdollisuus ennustaa kyselyjensä kustannukset myös *kilpailutekijä:* jos joku tällaisen mahdollisuuden tarjoaa, on muidenkin se tarjottava, etteivät käyttäjät alkaisi kartuttaa kilpailijan tilejä.

Sellaiset apuvälineet, jotka kykenevät ennustamaan kyselyjen kustannukset tarjoavat ratkaisun näihin ongelmiin. Vasta niiden kehittäminen antaa täyden hyödyn tulos- ja käsittelyperusteisille kyselyjen kustannusten laskentatavoille, jotka sinänsä ovat saaneet suuren kannatuksen.

5. Käyttökustannusten ennustaminen

5.1. Osaongelmat ja vaatimukset

Numeerisen tietokannan käyttäjän ongelmana ovat kyselyjen vaihtelevat ja hankalasti ennustettavissa olevat kustannukset noudatettaessa käyttökustannusten tulos- tai käsittelyperusteisia laskentatapoja. Tutkittavaksi ongelmaksi muodostuu siten, *miten voidaan ennustaa käyttäjän muotoileman kyselyn kustannukset numeerisessa tietokannassa etukäteen jo kyselyn muotoiluvaiheessa, mikäli kustannusten määräytymisperusteena lopullisessa laskutuksessa käytetään tulosja käsittelyperustetta.* Ongelmana on siis *metodologian kehittäminen* käyttökustannusten ennustamista varten. Seuraavat kolme osaongelmaa voidaan johtaa kustannusten laskentatapojen ja numeeristen tietokantojen käyttöominaisuuksien perusteella:

- *Kuinka suuria ovat kyselyn tulos ja välitulokset, ts. mikä on niiden sisältämien tietueiden lukumäärä (kardinaalisuus)?* Tätä ongelmaa kutsutaan *tuloksen koon arviointiongelmaksi* (cardinality estimation problem). Tämän ongelman ratkaisusta riippuvat molempien muiden osaongelmien ratkaisut.
- *Mikä on kyselyn tuloksen tulosperusteinen hinta?* Tätä ongelmaa voidaan kutsua *tuloksen hinnoitteluongelmaksi* (princing problem). Sen ratkaiseminen edellyttää ensimmäisen osaongelman ratkaisua, koska tuloksen koko on keskeinen tekijä sen hinnan muodostumisessa, ja lisäksi tietokannan attribuuttien arvojen hinnoittelua. Sen ratkaiseminen tarjoaa käyttäjälle tiedon, joka vastaa viitetietokannoissa saatavaa tietoa kyselyn osumien lukumäärästä, josta voidaan edelleen laskea tulostettavien viitteiden hinta.
- *Mikä on kyselyn käsittelyperusteinen hinta?* Tätä ongelmaa voidaan kutsua *kyselyn käsittelykustannusten arviointiongelmaksi* (processing cost estimation problem). Sen ratkaiseminen edellyttää ensimmäisen osaongelman ratkaisua, koska kyselyn väli- ja lopputulosten koko on keskeinen tekijä sen käsittelykustannuksen muodostumisessa: mitä suurempia tietomassoja käsitellään, sen suuremmaksi käsittelykustannus todennäköisesti muodostuu. Lisäksi tulee kuitenkin ottaa huomioon tietokannan tiedostorakenteet ja niiden vaikutus käsittelykustannuksiin.

Ollakseen käyttökelpoisia, metodologian ja siihen perustuvien apuvälineiden tulee täyttää seuraavat vaatimukset:

- Saman *metodologian tulee kattaa koko ongelma-alue* ja sopia vielä kyselyjen *muidenkin* kustannusten arviointiin. Esimerkiksi hajautettujen tietokantojen yleistyminen aiheuttaa sen, että tietoliikennekustannusten arviointi tulee tärkeäksi; tietoliikennepalvelujen kehittäjät ovat jo suunnitelleet tietoliikenneveloituksen muuttamista suoraan siirrettävien tulosten ja välitulosten koon perusteella määräytyväksi [Hull-84] [Hunt-84].
- Kustannusennusteiden lisäksi metodologian tulee tarjota tietokannan käyttäjille *riittävän selkeitä ja monipuolisia kuvauksia kyselyjen tuloksista,* jotta käyttäjät voivat niiden perusteella arvioida kyselyjen hyötyä ja sopivuutta ja tarpeen mukaan muotoilla niitä uudelleen. Tämä on tärkeää, koska tulokset kuvaavat vaihtelevia kohteita, ilmiöitä ja/tai suhteita, eikä pelkästään dokumentteja, kuten viitetietokannoissa.
- Koska kyselyt voivat vaihdella kohdetietojensa ja monimutkaisuutensa suhteen varsin laajoissa rajoissa, tulee metodologian olla *yleispätevä;* sen tulee kyetä *minkä tahansa* esitettävissä olevan kyselyn kustannusten ennustamiseen.
- Koska käyttäjät muokkaavat kyselyjään ja päättävät niiden suorittamisesta kustannusennusteiden perusteella, pitää ennusteiden olla riittävän *tarkkoja.*
- Metodologian tulee tehdä kyselyjen kustannusten ennustaminen *vaivattomaksi* tietokannan käyttäjälle; mitään uusia ponnistuksia tai teknisiä tietoja käyttäjiltä ei saa edellyttää.
- Metodologian tulee olla *täsmällisesti määritelty,* sillä muutoin sen toimivuudesta ja yleispätevyydestä ei voida olla vakuuttuneita. Tulee siis määritellä täsmällisesti, *mitä* tietoja kustannusten ennustamisessa käytetään ja *kuinka* niitä siinä käytetään.
- Metodologian tulee olla *suoraan sovellettavissa* relaatiomalliin perustuviin numeerisiin tietokantoihin ja sen tulee tarjota *suoria vihjeitä* siitä, kuinka sitä voidaan soveltaa muihin tietokantoihin, joissa sovelletaan jotakin muuta tietomallia tai jotka eivät ole numeerista (esim. viitetietokannat).

Informatiikan piirissä on perinteisesti kehitetty tiedonhakua helpottavia ja yksinkertaistavia apuvälineitä ja menetelmiä. Tietokantojen käyttökustannusten ennustamismetodologian ja -apu-

välineiden kehittäminen on tämän perinteen jatkamista: näiden avulla voidaan numeeristen tietokantojen käyttöä yksinkertaistaa ja helpottaa, koska ne tarjoavat käyttäjälle toisaalta arvioita kyselyjen kustannuksista ja toisaalta muitakin tietoja, joiden nojalla kyselyjen muotoilu ja uudelleenmuotoilu tulevat mielekkäiksi. Käyttäjä saa tietoa, joka auttaa häntä rajaamaan tai laajentamaan kyselynsä tarpeitaan paremmin vastaavaksi sekä sisältönsä että hintansa puolesta.

5.2. Metodologia

Kyselyjen kustannusten ennustamismetodologian osaongelmia on tarkasteltu kirjallisuudessa aikaisemmin, tosin erillään ja muissa yhteyksissä kuin numeeristen tietokantojen käyttökustannusten ennustamisen kannalta. Joka tapauksessa metodologian kehittämistä varten on tarjolla aikaisempia tutkimuksia, joissa kehitettyjä menetelmiä voidaan *soveltaa* nyt tarkasteltavaan tehtävään ja jotka tarjoavat hyvin perusteltuja vihjeitä siitä, mitä tietoja kyselyjen kustannusten ennustamisessa tarvitaan. Tarkastellaan tätä osaongelmittain.

Kyselyjen *tuloksen koon arviointiongelman* ratkaisemiseen voidaan soveltaa relaatiomallin puitteissa taloudellis-hallinnollisia tietokantasovelluksia varten aikaisemmin kehitettyjä menetelmiä. Lähestymistapoja on tarjolla useita. Yleisimmässä lähestymistavassa oletetaan, että 1) tiedostojen *tietueiden jakaantuminen* kunkin attribuutin ari arvojen kesken on *tasainen,* 2) attribuuttien *arvojen jakaantuminen* attribuutin arvoalueella on *tasainen* ja 3) eri attribuuttien *arvot* ovat *toisistaan riippumattomia* (esim. [BGWR-81] [GeGa-82] [GaPu-84] [Järv-84] [Järv-86b] [Rich-81] [Rose-81] [SACL-79] [YuLi-82]). Kutsutaan tätä lähestymistapaa *standardilähestymistavaksi.* Tämän lähestymistavan etuna on sen helposti saavutettava yleispätevyys ja yksinkertaisuus: kyselyjen tulosten koon arviointiin tarvitaan varsin vähän parametritietoja. Tietokantojen tiedot eivät kuitenkaan tavallisesti täytä tämän lähestymistavan yksinkertaistavia oletuksia, kuten on osoitettu useissa tutkimuksissa (esim. [Chri-83a] [Chri-83a] [MeOt-79] [PiCo-84]). Jakaumat ovat usein vinoja, ja mikäli attribuuttien välillä vielä on *funktionaalisia riippuvuuksia* [Ullm-80], ei riippumattomuusoletus pidä kirjaimellisesti paikkaansa. Tämän takia standardilähestymistapa johtaa usein konservatiivisiin (liian suurin) ennusteisiin tulosten koosta [Chri-83a]. Muita lähestymistapoja on kehitetty tarkempien ennusteiden saavuttamiseksi. Christodoulakis [Chri-83a] [Chri-83b] ehdottaa käytettäväksi mm. tilastotieteen *parametrisia menetelmiä* ja osoittaa niiden edut verrattuna standardilähestymistapaan. Merret ja Otoo [MeOt-79] ehdottavat *jakaumamallien* (distribution models) käyttöä tarkempien ennusteiden saamiseksi ja Piatetsky-Shapiro ja Connell [PiCo-84] suosittelevat *jakaumavälimenetelmää* (distribution step method).

Mikäli sovelletaan standardilähestymistapaa tai sen kehitelmiä, voidaan kyselyjen tulosten koon ennustamisessa tarvittavat parametrit yhdistää aiemmista tutkimuksista seuraavasti:

- tiedoston *nimi* tarvitaan ilmaisemaan, mitä tiedostoa parametrit koskevat (esim. [BGWR-81] [YuLi-82]);
- tiedoston *tietueiden lukumäärä l. kardinaalisuus;* ilman tätä tietoa ei kyselyjen tulosten kokoa voida ollenkaan arvioida;
- tiedoston *attribuuttien kuvaus:* kustakin *nimi* tunnisteeksi (esim. [BGWR-81] [YuLi-82]), *eri arvojen lukumäärä l. selektiivisyys* (esim. [SACL-79] [YuLi-82]), *arvoalue* (domain, esim. [BGWR-81] [YuLi-82]), sekä *arvojen ala- ja ylärajat* (range, esim. [BGWR-81] [YuLi-82]); näitä tietoja tarvitaan, kun arvioidaan esim. kyselyn rajoitusoperaatioiden rajoitusehdot täyttävien tietueiden lukumääriä [Järv-84] [Järv-86b];
- tiedoston attribuuttien välisten *funktionaalisten riippuvuuksien kuvaus* [GeGa-82] [GaPu-84]; funtionaaliset riippuvuudet ovat keskeisiä projektio-operaatioiden tulosten koon ennustamisessa.

Tulosten *hinnoitteluongelmaa* ei ole tarkasteltu relaatiomallin puitteissa eikä muutenkaan numeeristen tietokantojen yhteydessä. Ongelman ratkaisemiseksi voidaan kuitenkin soveltaa ja kehittää edelleen viitetietokannoista tuttua tapaa viitteiden hinnoittelussa [Aitc-84] [Henr-80] [KaWa-85]. Tämä edellyttää edellä esitettyä tietokantaan talletettujen tietojen tietoalkiokohtaista hinnoittelua tietuekohtaisen hinnoittelun sijasta, kuten Hull [Hull-84] ja Hunter [Hunt-84] ehdottavat viite- ja tekstitietokantojen yhteydessä, ja näiden hintatietojen laskemista myös kyselyn tulosta varten. Tulosten koon ennustamisessa tarvittujen tietojen lisäksi tarvitaan siis tieto kunkin attribuutin arvojen yksikköhinnasta.

Kyselyn *käsittelykustannusten arviointiongelman* ratkaisemisessa voidaan soveltaa *tietokantojen ja tiedostojen suunnitteluun kehitettyjä menetelmiä* (esim. [Chan-76] [Hans-82] [TeFr-82] [Schk-74] [Wied-77] [YaMe-75] [Yao-77]) sekä relaatiomallin puitteissa kehitettyjä *kyselyjen optimointimenetelmiä* (esim. [AsCh-75] [Astr-76] [BGWR-81] [BlEs-77] [Chri-83a] [Hall-76] [Kim-80] [Merr-83] [SACL-79] [SmCh-75] [Yao-79]). Koska tässä kirjallisuudessa ovat jat-

kuvati olleet huomion kohteena lähes yksinomaan taloudellishallinnolliset tietokantasovellukset, joissa kyselyjen kustannuksilla ei ole välitöntä merkitystä käyttäjille, ei huomiota ole kiinnitetty kyselyjen kustannusennusteiden kertomiseen käyttäjille. Näissä sovelluksissa kyselyt suoritetaan kustannuksista riippumatta, joskin mahdollisimman tehokkaasti, käyttäjän haluamassa muodossa. Tämän takia käsittelykustannuksiakaan ei tarvitse tarkkaan arvioida: riittää, kun kyselyjen mahdollisten suoritustapavaihtoehtojen *suhteellinen* tehokkuus tai edullisuus voidaan arvioida. Varsin karkeat arviot riittävät tällöin. Tiedostojen suunnittelun menetelmät taas tarjoavat usein tarkkoja menetelmiä yhden tiedoston käsittelykustannusten arviointiin, mutta eivät yleispäteviä menetelmiä kyselyjä varten. Yhdistämällä tiedostojen suunnittelun ja kyselyjen optimoinnin menetelmiä voidaan tietokantojen käyttökustannusten ennustamiselle edellä asetetut vaatimukset täyttää käsittelykustannusten arvioinnin osalta.

Käsittelykustannusten arviointi edellyttää toisaalta kyselyjen laitteisto- ja ohjelmistoympäristön kuvaamista ja toisaalta tiedostojen tiedostorakenteiden kuvaamista niiden tietojen lisäksi, joita tarvitaan edellisten osaongelmien ratkaisemiseen [Järv-85a] [Järv-85b]. Tiedostorakenteiden kuvaaminen edellyttää seuraavan tyyppisien tietojen esittämistä:

- tiedoston *tiedostorakennetyyppi,* esim. onko tiedosto lajiteltu peräkkäistiedosto, käänteistiedosto jne.; tämän tiedon avulla voidaan toisaalta määritellä ja toisaalta tulkita muut tiedostorakenteen kuvaustiedot;
- *tiedostorakennetyypin parametrit,* esim. tietueen pituus, lajitteluavain, hakemistojen kuvaus, ylivuotoalueiden kuvaus jne. (esim. [TeFr-82] [Wied-77]).

Lisäksi käsittelykustannusten ennustamista varten tulee kerätä ja tallettaa kuvauksiin tieto kunkin välituloksen tuottamisen ennustetuista kustannuksista. Käsittelykustannus ei ole tiedoston tietojen ominaisuus, eikä se niiden kuvaukseen sisälly. Se vaatii oman parametrinsa.

Kaikkien osaongelmien ratkaisut pitää saada toimimaan samassa viitekehyksessä ja siten, ettei ennusteiden laskeminen edellytä kyselyn suorittamista ensin. Lisäksi tulee tuottaa käyttäjälle ymmärrettävät ja hyödylliset tulokset kyselyjen muotoilua varten. Tämä onnistuu laatimalla tiedostoista sopivat kuvaukset samaan tapaan kuin bibliografioiden bibliografiassa kuvataan bibliografioita: on kuvattava tietokannan kunkin tiedoston tietueiden lukumäärä (vrt. bibliografian kate ja koko) ja attribuutit (vrt. bibliografian kuvailun taso ja siihen sisältyvien kenttien ja niiden sisällön (kuten käytettyjen luokituskaavojen) kuvaus) sekä rakenne (vrt. bibliografian järjestyksen ja hakemistojen jne. kuvaus). Jos kyseessä olisi viitetietokanta, olisi analogia lähes täydellinen — nyt tietosisällön erot aiheuttavat eroja kuvauksissakin. Sekä metodologian kehittämisen että siihen perustuvien välineiden kannalta on tiedostojen kuvaus keskeinen apuväline.

Tiedostokuvauksen tulee sisältää kaikki ne parametrit, joita kyselyjen kustannusten ennustamisessa tarvitaan. Lisäksi näiden parametrien tulee olla sellaisia, että ne voidaan johtaa minkä tahansa kyselyn kaikille välituloksille ja lopputulokselle. Näin taataan metodologian yleispätevyys. Tällaisen tiedostokuvauksen esitystavaksi sopii *n-jono -esitys* (n-tuple representation), josta on tullut suosittu esitystapa tiedonhallintatutkimuksessa sellaisten ongelmien ratkaisemisessa, jossa tarvitaan täsmällistä, eksplisiittistä ja monipuolista kuvailua tai määrittelyä (esim. [AuBM -80] [Niemi-83] [NiJä-85]). Yhden tiedoston käsittelykustannusten ennustamista varten on määritelty n-jono -esitykseen perustuva tiedostokuvaus [Järv-85a] ja sitä on sovellettu kyselyjen välitulosten koon ennustamiseen numeeristen tietokantojen yhteydessä [Järv-84] [Järv-86b]. Kuvaus on, hiukan yksinkertaistettuna, seuraavanlainen. Esimerkkinä kuvataan luvun kolme esimerkkitiedosto PRODUCTS:

> **P** = (PRODUCTS, 250000, {(PRODUCT-NO, 7, 250000, int, 1000, 9999999), (TRADEMARK, 20, 200000, char, λ, λ), (TYPE, 5, 17000, int, 100, 50000), (MANUF-NO, 6, 12000, int, 1000, 999999), (QSALES, 7, 20000, int, 0, 9999999)}, {{PRODUCT-NO} —> {PRODUCT-NO, TRADEMARK, TYPE, MANUF-NO, QSALES}, {TRADEMARK} —> {MANUF-NO}, ...}, (indexed-file, (45, λ, {(TRADEMARK, 7813, 3, 200000), (TYPE, 2715, 2, 17000), (PRODUCT-NO, 4483, 3, 250000)})), 0)

Kuvauksen nimi on **P** ja sen komponentit ovat seuraavat (lukuarvot ovat kuviteltuja):

- *Tiedoston nimi:* PRODUCTS;
- *Tiedoston tietueiden lukumäärä:* 250000;
- *Tiedoston attribuuttien kuvausjoukko:* {(PRODUCT-NO, 7, 250000, int, 1000, 9999999), (TRADEMARK, 20, 200000, char, λ, λ), (TYPE, 5, 17000, int, 100, 50000), (MANUF-NO, 6, 12000, int, 1000, 999999), (QSALES, 7, 20000, int, 0, 9999999)}. Tässä kunkin attribuutin kuvaus koostuu kuudesta komponentista, jotka ovat *attribuutin nimi* (esim. TYPE), sen arvot sisältävän *kentän pituus* (esim. 5), *attribuutin eri arvojen lukumäärä* (esim. 17000), *attribuutin arvojen tyyppi* (esim. int), *attribuutin arvojen alaraja* (esim. 100), ja *att-*

ribuutin arvojen yläraja (esim. 50000). Merkkijonotyyppisiltä (char) attribuuteilta (esim. TRADEMARK) ala- ja ylärajat puuttuvat (merkintä: λ). Nämä kuusikkokuvaukset yhdessä muodostavat attribuuttien kuvausjoukon.

- *Tiedoston attribuuttien funktionaalisten riippuvuuksien kuvausjoukko:* {{PRODUCT-NO} —> {PRODUCT-NO, TRADEMARK, TYPE, MANUF-NO, QSALES}, {TRADEMARK} —> {MANUF-NO}, ...}. Kukin funtionaalinen riippuvuus esitetään parina, esim. {TRADEMARK} —> {MANUF-NO}, jonka vasemman puolen muodostaa niiden attribuuttien nimien joukko, jotka funktionaalisesti määrittävät oikean puolen joukossa nimetyt attribuutit. Kolme pistettä (...) kertoo, ettei joukkoa ole esitetty kokonaisuudessaan.
- *Tiedoston tietorakenteen kuvaus:* (indexed-file, (45, λ, {(TRADEMARK, 7813, 3, 200000), (TYPE, 2715, 2, 17000), (PRODUCT-NO, 4483, 3, 250000)})). Tässä ilmaistaan ensin tiedostorakenteen tyyppi (indexed-file) ja annetaan sitten sen kuvausparametrit, jotka ovat *tietueen pituus* (45), *tietueen lajitteluavain* (λ, tarkoittaa lajittelun puuttumista), ja *käänteishakemistojen kuvausjoukko,* jossa kukin kuvaus on nelikkö. Nämä neliköt kuvaavat *käänetyn attribuutin nimen* (TYPE), *hakemiston leveyden* (2715), *hakemiston korkeuden* (2), ja *hakemiston sisältämien hakusanojen lukumäärän* (17000).
- *Tiedoston käsittelykustannus kyselyä varten:* 0. PRODUCTS-tiedosto on tietokantaan pysyvästi talletettu ja sellaisenaan sillä ei vielä ole mihinkään kyselyyn liittyvää käsittelykustannusta, mitä luku 0 kuvaakin.

Tätä kuvausta ja sen komponentteja voidaan käsitellä täsmällisesti määritellyin matemaattisin funktioin ja siten voidaan kyselyjen kustannusten ennustamismenetelmät määritellä täsmällisesti. Kuvaukseen kuuluu joukko varsin teknisiäkin komponentteja, joita ei ole, ainakaan Suomessa, ehkä totuttu tarkastelemaan tiedon tallennuksen ja haun yhteydessä. *Ainoastaan eksplisiittisesti nämä komponentit esittämällä voidaan kehittää menetelmiä ja välineitä, jotka vapauttavat tietokantojen käyttäjät tarpeesta tuntea ja käsitellä näitä komponentteja.* Käyttäjäystävällisyyden aikaansaaminen merkitsee usein joko sitä, että käyttäjille vaikeat tehtävät määritellään täsmällisesti ja automatisoidaan, tai sitä, että rajoitetaan mahdolliset tehtävät niin, että vain yksinkertaiset jäävät jäljelle. Tässä on valittu edellinen tapa, joka on ainoa mahdollisuus monimutkaisten kyselyjen kustannusten ennustamista ajatellen. Kun esitettyä kuvausta täydennetään kyselyjen tulosten hinnoitteluongelman vaatimilla lisäkomponenteilla, saadaan tiedostokuvaus, jonka avulla voidaan ratkaista kaikkien relaatiomallin puitteissa esitettyjen kyselyjen tulos- ja käsittelyperusteiset laskutusongelmat.

6. Johtopäätökset

Artikkelissa on tarkasteltu numeerisia tietokantoja, niiden eroja ja etuja perinteisiin viitetietokantoihin, taloudellis-hallinnollisiin tietokantoihin ja paperimuotoiseen julkaisemiseen verrattuna, sekä vaikutuksia tiedon saatavuuteen ja hankintaan. Vaikutusten kannalta keskeiseksi seikaksi on osoitettu numeeristen tietokantojen käyttökustannusten laskentatapa ja käyttökustannusten ennustamismahdollisuus jo hakutehtävän eli kyselyn muotoiluvaiheessa. Tämä johtuu toisaalta niistä kehityslinjoista, jotka painottavat tulos- ja käsittelyperusteisten kyselykustannusten laskentatapojen keskeisyyttä laskutuksessa ja perinteisten laskutustapojen sopimattomuutta, ja toisaalta numeeristen tietokantojen ominaispiirteistä, jotka aiheuttavat suurta vaihtelua kyselyjen kustannuksiin, mikäli tulos- ja käsittelyperusteisia kustannusten laskentatapoja sovelletaan.

Kyselyjen kustannusten ennustamismahdollisuus voidaan toteuttaa ainoastaan tarjoamalla käyttäjille sellaiset välineet tietokantojen käyttäjäliitännöissä, jotka laskevat kustannukset automaattisesti käyttäjän kyselyn ja tietokannan kuvauksen perusteella ennen kyselyn suorittamista. Käyttäjät eivät itse pysty kustannuksia ennustamaan, koska se on varsin monimutkaista, ja edellyttäisi tietoja ja taitoja, joita heiltä ei voi vaatia. Tietokantojen kehittämisessä on nimenomaan pyritty siihen, ettei tietokantojen käyttö tällaisia tietoja tai taitoja edellytä. Vain näin on laajamittainen online-tiedonhaku ylipäätään tullut mahdolliseksi. Ennustamisvälineiden puute aiheuttaa ongelmia tiedonhaun kaikille osapuolille ja vaikuttaa toisaalta loppukäyttäjien halukkuuteen käyttää tietokantoja sekä kirjastojen ja informaatiopalvelujen mahdollisuuksiin tarjota niitä asiakkailleen.

Kyselyjen kustannusten ennustamismetodologian ja -välineiden kehittämistä tarkasteltiin relaatiomallin puitteissa. Vankan teoreettisen perustansa ja laajan suosionsa takia se on muita tietomalleja selvästi parempi tarkastelukehys, jossa voidaan kehittää ja demonstroida tietokantojen käyttökustannusten ennustamismetodologia ja sillä saavutettavat tulokset täsmällisesti.

Kyselyjen kustannusten ennustamismenetelmän ja -välineiden kehittämisen osaongelmiksi osoitettiin kyselyn tuloksen ja välitulosten koon

ennustaminen, ja tähän perustuen toisaalta tuloksen tulosperusteisen ja toisaalta käsittelyperusteisen kustannuksen ennustaminen. Myöhemmissä tutkimuksissa tulee selvittää, miten esitettyä metodologiaa täsmennetään ja sovelletaan kunkin osaongelman ratkaisemiseen relaatiomallin puitteissa. Tiedostokuvaukset tarjoavat monipuolisen välineen osaongelmien ratkaisemisen edellyttämien tietojen jäsentämiseen ja esittämiseen monimutkaisuudeltaan millaisen tahansa kyselyn analysoinnissa. Metodologiaa voidaan edelleen soveltaa kyselyjen muidenkin kulujen, kuten tiedonsiirtomaksujen tai vaikka yhteysajan, ennustamiseen.

Informatiikan piirissä on perinteisesti kehitetty tiedonhakua helpottavia ja yksinkertaistavia apuvälineitä ja menetelmiä. Yhä useammin näiden kehittäminen edellyttää tietojenkäsittely- ja tiedonhallintatutkimuksen ja matemaattisten menetelmien tuntemusta, mikä ilmenee mm. tiedon tallennuksen ja haun perusoppikirjoissakin (esim. [SaMc-83]), vaikka näitä ei ole, ainakaan Suomessa, ehkä totuttu tarkastelemaan tiedon tallennuksen ja haun yhteydessä. Vain tuntemalla nämä tekniikat tarkoin voidaan ne piilottaa tiedonhakijalta ja siten vapauttaa tietokantojen käyttäjät tarpeesta tuntea ja hallita ne. Näin voidaan kehittää joustavia, käyttäjäystävällisiä ja yleispäteviä välineitä tiedonhaun helpottamiseksi ja tehostamiseksi.

Numeeristen tietokantojen käyttökustannusten ennustamismetodologian kehittäminen tuottaa 'sivutuotteenaan' taitoja ja välineitä kaikenlaisten tiedon tallennukseen ja hakuun liittyvien ilmiöiden täsmälliseen kuvaamiseen ja siinä esiintyvien ongelmien ratkaisemiseen. Tämän kirjoituksen ajatuksia voidaan soveltaa esimerkiksi viitetietokantojen täsmälliseen kuvaamiseen joukko-opin ja logiikan välinein ja viitteenhaun täsmälliseen määrittelyyn tältä pohjalta.

Hyväksytty julkaistavaksi 17. 6. 1986

Viitteet:

[1] Strukturoidulla tiedolla tarkoitetaan tietoa, joka on jäsennetty määrämuotoisiin kenttiin ja esitetty määrämuotoisella tavalla. Tällaista tietoa voidaan tarkastella esimerkiksi taulukkona, jonka rivit kuvaavat havaintoyksiköitä ja sarakkeet ovat kuvauksen ulottuvuuksia. Esimerkiksi sopii vaikka väestörekisteri. Tekstimuotoiset dokumentit eivät ole tässä mielessä strukturoituja.

[2] Jotkin tilastotietokannat on kuitenkin toteutettu samaan tapaan kuin viitetietokannat: esimerkiksi Dialogin tarjoamat Predicasts-tietokannat sisältävät mm. aikasarjoja, jotka on sanallisesti indeksoitu ja joita voidaan hakea käänteistiedoston avulla totuttuun tapaan. Samalla on kuitenkin huomattavasti jouduttu rajoittamaan tietojen käsittely- ja tulostusmahdollisuuksia [KaLM-85].

[3] Käyttäjän *päätteelle näkyvä* vastausaika viitetietokannasta on yleensä enintään muutama sekunti. Siihen vaikuttaa myös tiedonsiirtoverkon tyyppi ja siirtonopeus. Viitetietokannan kyselyjärjestelmä ei poimi kaikkia tulostettavia viitteitä kerralla keskusmuistiin, vaan muutaman kerrallaan, koska tiedonsiirto ja tulostus ovat suhteellisen hitaita verrattuna tietojen poimintaan. Jos kaikki viitteet kuitenkin poimittaisiin ennen ensimmäisen lähettämistä tulostettavaksi käyttäjälle, olisi vastausaika suoraan verrannollinen tuloksen kokoon, tavallisesti yksi sekunti per 10...20 viitettä. Vastaavaa riippuvuutta ei numeerisissa tietokannoissa ole: vastaus voi koostua vaikka yhdestä luvusta, mutta sen laskeminen saattaa viedä useita minuutteja, jos kysely on monivaiheinen; toisaalta, jos kysely on yksivaiheinen, vastausaika voi määräytyä samaan tapaan kuin viitetietokannoissa.

[4] Tämän kirjoituksen kannalta ei ole merkitystä sillä, onko tietoalkioiden hinnan yksikkönä pennit tai pfenningit, centit tai pencet tai niiden murto-osat.

Lähteet:

[Aitc-84] Aitchison, T. M. »Online and the database producer», Journal of Information Science. 9 (2): 75—80; 1984.

[Alte-80] Alter, S. Decision Support Systems: Current Practice and Continuing Challenges. Reading, Mass: Addison-Wesley; 1980.

[Astr-76] Astrahan, M. M. et al. »System R: Relational Approach to Database Management», ACM TODS. 1 (2): 97—137; 1976.

[AsCh-75] Astrahan, M. M.; Chamberlin, D. D. »Implementation of a Structured English Query Language», Communications of the ACM. 18 (10): 580—588; 1975.

[Atki-79] Atkinson, M. P. »Database Systems», Journal of Documentation. 35 (1): 49—91; 1979.

[AtRu-84] Attewell, P.; Rule, J. »Computing and Organizations: What We Know and What We Don't Know», Communications of the ACM. 27 (12): 1184—1192; 1984.

[AuBM-80] Ausiello, G.; Batini, C.; Moscarini, M. »On the equivalence among data base schemata» In: Proceedings of the International Conference on Data Bases, Aberdeen, Scotland, 2—4 July, 1980. London: Heyden, 1980: 34—46.

[Auvi-85] Auvinen, A. Kansainväliset markkinatietokannat viejän tietolähteinä. Helsinki, Finland: Suomen Ulkomaankauppaliitto, Tietopalveluryhmä; 1985.

[BlEs-77] Blasgen, M. W.; Eswaran, K. P. »Storage and access in relational data bases», IBM Systems Journal. 16 (4): 363—377.

[BGWR-81] Bernstein, P. A.; Wong, E.; Reeve, C. L.; Rothnie, J. B. Jr. »Query Processing in a System for Distributed Databases (SDD-1)», ACM TODS. 6 (4): 602—625; 1981.

[BoMC-84] Borgman, C. L.; Moghdam, D.; Corbett, P. K. Effective Online Searching: a Basic Text. New York, NY: Marcel Dekker; 1984.

[CaMS-75] Caplan, N.; Morrison, A.; Stambaugh, R. J. The Use of Social Science Knowledge in Policy Decisions at the National Level: a Report to the Respondents. Ann Arbor, Mich.: University of Michigan, Institute for Social Research; 1975.

[Cawk-80] Cawkell, A. E. »Information Technology and Communications» In: Williams, M. E. ed. Annual Review of Information Science and Technology. Vol. 15. White Plains, NY: Knowledge Industry Publications; 1980: 37—65.

[Chan-76] Chan, A. Y. Index Selection in a Self-Adaptive Relational Data Base Management System. Cambridge, MA: MIT, Lab. for Computer Science, Report MIT/LCS/TR-166; 1976.

[ChHe-84] Chen, C.-c.; Hernon, P. eds. Numeric Databases. Norwood, NJ: Ablex, 1984.

[Chri-83a] Christodoulakis, S. »Estimating Record Selectivities», Information Systems. 8 (2): 105—115; 1983.

[Chri-83b] Christodoulakis, S. »Estimating block transfers and join sizes», ACM SIGMOD Record. 13 (4): 40—54; 1983.

[ClCr-83] Clarke, A.; Cronin, B. »Expert Systems and Library/Information Work», Journal of Librarianship. 15 (4): 277—292; 1983.

[Codd-70] Codd, E. F. »A Relational Model for Large Shared Data Banks», Communications of the ACM. 13 (6): 377—387; 1970.

[Craw-81] Crawford, R. »The Relational Model in Information Retrieval». Journal of the American Society for Information Science. 32 (1): 51—64; 1981.

[Cron-85] Cronberg, T. Työ, aika ja asuminen tietoyhteiskunnassa. Helsinki, Finland: Asuntohallitus; 1985.

[Cuad-82] Directory of Online Databases. Santa Monica, CA: Cuadra Associates; 1982.

[Cuad-84] Directory of Online Databases. Santa Monica, CA: Cuadra Associates; 1984.

[DuBo-84] Dunn, R. G.; Boyle, H. F. »Online Searching: an Analysis of Marketing Issues», Information Services & Use. 4: 147—154; 1984.

[Elor-74] Eloranta, K. T. Heuristiikat ja heuristisuus. Tampere, Finland: University of Tampere, Dept. of Administrative Sciences; 1974.

[Eusi-83] EUSIDIC Database Guide 1983. Learned Information, Oxford, 1983.

[Fost-84] Foster, A. »Business Information from Databanks: the Potential of Online Numeric Databases», Business Information Review. 1 (1): 38—45; 1984.

[Gaul-84] Gault, F. D. »Database Management Systems for Science and Technology», In: [RuHa-84] : 39—73.

[GeGa-82] Gelenbe, E. and Gardy, D. »On the Size of Projections: I», Information Processing Letters. 14 (1): 18—21; 1982.

[GaPu-84] Gardy, D. and Puech, C. »On the Size of Projections: A Generating Function Approach», Information Systems. 9 (3/4): 231—235; 1984.

[Hall-76] Hall, P. A. V. »Optimization of a Single Relational Expression in a Relational Data Base System», IBM Journal of Research and Development. 20 (3): 244—257; 1976.

[Hans-82] Hanson, O. Design of Computer Data Files. London: Pitman; 1982.

[Heim-82] Heim, K. M. ed. Data Libraries for the Social Sciences. A Special Issue of Library Trends. 30 (3); 1982.

[Henr-80] Henry, W. M. et al. Online Searching: An Introduction. London, UK: Butterworth; 1980.

[Hull-84] Hull, D. »Marketing and Pricing of Full-text End-user Services», Information Services & Use. 4: 167—170; 1984.

[Hunt-84] Hunter, J. A. »What Price Information», Information Services & Use. 4: 217—223; 1984.

[Inte-84] »International Comparative Price Guide to Databases», Online Review. 9 (1): 77—84; 1984.

[Järv-81] Järvelin, K. »Tiedontarpeiden tutkimisesta informatiikassa: viitekehysten arviointi» In: Järvelin, K.; Vakkari, P. Tiedontarpeiden ja kirjastonkäytön tutkimisesta: kaksi tutkielmaa. Helsinki, Finland: Kirjastopalvelu; 1981: 15—64.

[Järv-82] Järvelin, K. »Finding functional dependencies for intermediate relations of relational algebra expressions» In: Kangassalo, H. ed. Proceedings of the First Scandinavian Research Seminar on Information Modeling and Data Base Management. Acta Universitatis Tamperensis ser. B 17. Tampere, Finland: University of Tampere: 1982: 407—441.

[Järv-84] Järvelin, K. Cardinalities and attribute descriptions of result relations of relational algebra operations. Tampere, Finland: University of Tampere, Dept. of Mathematical Sciences, Report A134, 1984.

[Järv-85a] Järvelin, K. A systematic approach to modelling the costs of flat files. Tampere, Finland: University of Tampere, Dept. of Mathematical Sciences, Rep. A152, 1985.

[Järv-85b] Järvelin, K. »A systematic approach to query cost modelling» In: Kangassalo, H. ed. Proceedings of the 4th Scandinavian Research Seminar on Information Modeling and Data Base Management, Ellivuori, Finland, June 5—7, 1985 (in press).

[Järv-86a] Järvelin, K. »On Information, Information Technology and the Development of Society: An Information Science Perspective» In: Ingversen, P.; Kajberg, L.; Mark Pejtersen, A. eds. Information Technology and Information Use: towards a unified view of information and information technology. London, UK: Taylor Graham; 1986: 35—55.

[Järv-86b] Järvelin, K. »Estimation of Query Cardinalities in Numeric Databases» In: Brookes, B. C. ed. Intelligent Information Systems for the Information Society: Proceedings of the IRFIS 6 Conference, Frascati, Italy, Sept. 16—18, 1985. Amsterdam, NL: North-Holland, 1986 (in press).

[JäRe-82] Järvelin, K.; Repo, A. J. »Knowledge work augmentation and human information seeking», Journal of Information Science. 5 (6): 79—86; 1982.

[JäRe-83] Järvelin, K.; Repo, A. J. »On the Impacts of Modern Information Technology on Information Needs and Seeking: A Framework» In: Dietschmann, H. J. ed. Representation and Exchange of Knowledge as a Basis of Information Processes. Amsterdam, NL: North-Holland; 1983: 207—230.

[JäRe-84] Järvelin, K.; Repo, A. J. »A Taxonomy of Knowledge Work Support Tools» In: 1984: Challenges to an Information Society. Proc. 47th ASIS Annual Meeting, Vol. 21, Philadelphia, PA, Oct. 21—25, 1984. White Plains, NY: Knowledge Industry Publications; 1984: 59—62.

[KaLM-85] Kahima, M.; Lukkari, M.; Myllys, H. Ulkomaisia suorakäyttöisiä tilastotietokantoja. Helsinki, Finland: Tilastokeskus, muistio N:o 98; 1985.

[KaWa-85] Karlsson, U.; Wallin, M. Att söka i databaser: Interaktiv informationssökning: metoder och möjligheter. Nordinfo-publikation 8. Ballerup, Sweden: Bibliotekcentralen; 1985.

[Kim-80] Kim, W. »A new way to compute the product and join of relations» In: Chen, P. P.; Sprowls, R. C. eds. Proceedings of ACM-Sigmod 1980, Santa Monica, CA, May 14—16, 1980: 179—187.

[Koch-83] Kochen, M. »Information and Society» In: Williams, M. E. ed. Annual Review of Information Science and Technology. Vol. 18. White Plains, NY: Knowledge Industry Publications; 1983: 277—304.

[Kort-85] Kortteinen, M. »Uusi yhteiskuntamuoto?», Sosiologia. 22 (2): 87—105; 1985.

[KuRS-77] Kunz, W.; Rittel, H. W. J.; Schwuchow, W. Methods of Analysis and Evaluation of Information Needs: A Critical View. München: Verlag Dokumentation; 1977.

[Kurt-84] Kurtz, L. A. »An Introduction to the Database Management Systems», Program. 18 (1): 1—15; 1984.

[Lanc-78] Lancaster, F. W. Toward Paperless Information Systems. New York, NY: Academic Press; 1978.

[Lanc-68] Lancaster, F. W. Information Retrieval Systems: Characteristics, Testing and Evaluation. New York, NY: Wiley; 1968.

[Lepp-85] Leppänen, A. »Ihminen tietokoneistetussa työssä: katsaus ihmisen ja tietokoneistetun työn vuorovaikutuksen psykologisiin tutkimusperinteisiin», Työterveyslaitoksen tutkimuksia. 3 (3): 287—303; 1985.

[Lern-83] Lerner, R. G. et al. »Primary Publication Systems and Scientific Text Processing» In: Williams, M. E. ed.

Annual Review of Information Science and Technology, vol. 18. White Plains, NY: Knowledge Industry Publications; 1983: 127—149.

[Lomi-85] Lomio, J. P. »The High Cost of NEXIS and What a Searcher Can Do About It», Online. 9 (5): 77—84; 1984.

[Marc-83] Marcus, R. S. »An Experimental Comparison of the Effectiveness of Computers and Humans as Search Intermediaries», Journal of the ASIS. 34 (6): 381—404; 1983.

[McLe-77] McLeod, I. A. »Towards an Information Retrieval Language Based on the Relational View of Data», Information Processing & Management. 13: 167—175; 1977.

[McOt-79] Merret, T. H.; Otoo, E. »Distribution Models of Relations» In: Proceedings of the 5th International Conference of Very Large Data Bases, Rio de Janeiro, Brazil, 1979: 418—425.

[Merr-83] Merret, T. H. »Why Short-Merge Gives the Best Implementation of the Natural Join», ACM SIGMOD Record. 13 (2): 39—51; 1983.

[MåOW-84] Mårtenson, G.; Otala, M.; Wiio, O. A. Tietotekniikka 1990-luvulla. Helsinki, Finland: SITRA, sarja B Nro 78; 1984.

[Niem-83] Niemi, T. »A Seven-tuple Representation for Hierarchical Data Structures», Information Systems. 8 (3): 152—157; 1983.

[NiJä-85] Niemi, T.; Järvelin, K. »A Straightforward Formalization of the Relational Model», Information Systems. 10 (1): 65—76; 1985.

[PiCo-84] Piatetsky-Shapiro, G.; Connell, C. »Accurate estimation of the number of tuples satisfying a condition» In: Proceedings of the ACM SIGMOD Conference, Boston, MA, June 18—21, 1984: 256—276.

[Poll-85] Pollit, S. »The CANSEARCH Approach to End User Information Retrieval» In: Brookes, B. C. ed. Intelligent Information Systems for the Information Society: Proceedings of the IRFIS 6 Conference, Frascati, Italy, Sept. 16—18, 1985. Amsterdam, NL: North-Holland; 1986 (in press).

[Rich-81] Richard, P. »Evaluation of the Size of a Query Expressed in Relational Algebra». In: Lien, Y. E. ed. Proceedings of the ACM SIGMOD Conference, Ann Arbor, Mich., 1981: 155—163.

[Rich-83] Rich, R. F. »Management and Problem Solving Styles: An Assessment on Information System Designs» In: Debons, A.; Larson, A. G. ed. Information Science in Action: System Design. Boston, Mass.: Nijhoff; 1983: 240—264.

[Rose-81] Rosenthal, A. S. »A Note on the Expected Size of a Join», ACM SIGMOD Record. 11 (4): 19—25; 1981.

[RuHa-84] Rumble, J. R., jr.; Hampel, V. E. eds. Database management in science and technology: A CODATA sourcebook on the use of computers in data activities. Amsterdam: North-Holland; 1984.

[SACL-79] Selinger, P. G.; Astrahan, M. M.; Chamberlin, D. D.; Lorie, R. A.; Price, T. G. »Access path selection in a relational database management system». In: Bernstein, P. A. ed. Proceedings of the ACM SIGMOD Conference, Boston, MA, May 30 — June 1, 1979: 23—34.

[SaMc-83] Salton, G.; McGill, M. J. Introduction to Modern Information Retrieval. New York, NY: McGraw-Hill; 1983.

[Savo-84] Savolainen, R. »'Informaatioammatit' ja 'informaatioyhteiskunnan' kehittyminen», Kirjastotiede ja informatiikka. 3 (2): 35—45; 1984.

[Schk-74] Schkolnick, M. Optimizing partial inversions for files. San Jose, CA: IBM Research Laboratory, Rep. RJ 1477 (#22576), 1974.

[SmCh-75] Smith, J. M.; Chang, P. Y.-T. »Optimizing the Performance of a Relational Algebra Database Interface», Communications of the ACM. 18 (10): 568—579; 1975.

[Surp-85] Surprenant, T. T. »Global Threats to Information» In: Williams, M. E. ed. Annual Review of Information Science and Technology. Vol. 20. White Plains, NY: Knowledge Industry Publications; 1985: 3—25.

[TeFr-82] Teorey, T. J.; Fry, J. R. Design of Database Structures. Engelwood Cliffs, NJ: Prentice-Hall; 1982.

[Teno-84] Tenopir, C. »Full-Text Databases» In: Williams, M. E. ed. Annual Review of Information Science and Technology, vol. 19. White Plains, NY: Knowledge Industry Publications; 1984: 215—246.

[Ullm-80] Ullman, J. D. Principles of Database Systems. London: Pitman; 1980.

[Wied-77] Wiederhold, G. Database Design. New York, NY: McGraw-Hill; 1977.

[Will-82] Williams, M. E. »Relative Impact of Print and Database Products on Database Producer Expences and Income — A Follow-Up», Information Processing and Management. 18 (6): 307—311; 1982.

[YaMe-75] Yao, S. B.; Merten, A. G. »Selection of file organization using an analytic model» In: Proceedings of the first VLDB Conference, Sept. 1975, pp. 255—267.

[Yao-77] Yao, S. B. »An Attribute Based Model for Database Access Cost Analysis», ACM TODS. 2 (1): 45—67; 1977.

[Yao-79] Yao, S. B. »Optimization of Query Evaluation Algorithms», ACM TODS. 4 (2): 133—155; 1979.

[YuLi-82] Yu, C. T.; Lin, Y. C. »Some estimation problems in distributed query processing» In: Proceedings of the IEEE Distributed Computing Systems Conference, 1982: 13—19.