Tutkimustiedonhallinnan peruskurssi

Hannu Toivonen, Marko Salmenkivi, Inkeri Verkamo

hannu.toivonen, marko.salmenkivi, inkeri.verkamo@cs.helsinki.fi

Helsingin yliopisto – Tietojenkäsittelytieteen laitos

Metadata

Metadata: osan sisältö

- Mitä metadata on?
- Mihin sitä tarvitaan?
- Mitä vaikutusta metadatalla on tutkimustiedonhallinnan ratkaisuihin?
- Lähde: Z. Michalewicz (toim.): Statistical and Scientific Databases; luku 4
- Esimerkkisovellus: Pubudu Wariyapola et al.: Ontology and Metadata Creation for the Poseidon Distributed Coastal Zone Management System, Advances in Digital Libraries, 1999, 180–189.

Mitä metadatalla tarkoitetaan?

- Metadata on tietoa, jota tarvitaan datan (raaka- ja analysoidun)
  - hakemiseen, käsittelemiseen, jäsentämiseen, tulkitsemiseen
- Metadatan keruu ja talletus on välttämätöntä
  - muuten varsinainen data ei ole käyttökelpoista!
- tavoittena varsinaisen datan
  - ymmärrettävyys, käytettävyys, ylläpidettävyys
  - muillekin ja myöhemminkin
  - + tutkimuksen toistettavuus

Tutkimuksen metadataa

- raakadatan tuottaminen ja kerääminen
  - tutkimus- tai koejärjestelyt
  - datan keruun tai mittausprosessin kuvaus
  - kuka, missä, milloin?
- datan jalostus ja analyysi
  - suoritetut muunnokset, suodatukset, korjaukset
  - muunnosten yms. perustelut
  - tehdyt analyysit, niiden parametrit
  - sovelletut menetelmät ja työkalut

- tutkimustulosten käyttöohjeet
  - opastusta tiedon ymmärtämiseen, lisätiedon hankkimiseen
  - tulkinnassa käytetyt mallit ja otokset
- viittauksia ulkoisiin artikkeleihin, taksonomioihin, standardeihin
- sovellusaluekohtainen osaaminen
- ei välttämättä käytettävissä koneellisesti
- äärimmillään koko tutkimushistoria rekonstruoitavissa

Metadataa on myös...

- tiedon fyysinen rakenne
  - tiedostojen nimet ja formaatit, tietotyypit
- tiedon looginen rakenne
  - tietoalkioiden nimet ja attribuutit
  - mahdolliset operaatiot

I: Ohjaava metadata

- käsittelijöinä tietokannan ja järjestelmän operaatiot
- sovellusten välisen tiedonsiirron rajapinnat
- määrämuotoinen esitystapa
- ohjaavan metadatan käyttö suositeltavaa
  - laadun varmistus
  - tiedon eheys
  - sekä raakadatan että käsitellyn datan laatu
  - automatisointi, kustannusten ja resurssien säästö

: Opastava metadata@

- käsittelijöinä ihmiset
  - etsintä; selailu
- esitys luonnollisella kielellä, kuvina,
- opastava metadata on välttämätöntä
  - datan ymmärtäminen
  - rajapinta tiedon tuottajan ja käyt

Mistä tutkimustieto koostuu

- mikrodata eli raakadata:
  - havainnot
  - simuloinnin tulokset
- makrodata eli analysoitu data = käsittelyn tuloksena syntynyt data
  - periaatteessa konstruoitavissa uudelleen, mutta voi olla työlästä
  - käsittely voi kohdistua raakadataan tai jo analysoituun dataan: esim. yhteenvedot
- metadata

Tutkimustiedon ominaisuuksia

Raakadata:

- havaintoja tai simulointituloksia
- havaintojen säännöllisyys:
  - havaintoja tehdään usein jonkin (aikaan tai paikkaan liittyvän) säännön mukaan
  - koordinaattien talletus?
- havaintojen tiheys:
  - hylätyt arvot (mittaustarkkuus)
  - havaintotiedon tiivistys

- ajan vaikutus koordinaatteihin:
  - havaintopisteiden koordinaatit voivat muuttua ajan funktiona

Parametridata:

- raakadataa kuvaaviin parametreihin liittyy usein luokitteluhierarkioita
  - samat tai eri operaatiot hierarkian eri tasoilla
  - yhteenvedot hierarkian eri tasoilla

Datan käsittelyn tasoja

- raakadata
- kalibroitu data
- tarkistettu data
- johdettu data
- tulkinta

Tutkimustiedon rakenne

- moniulotteisuus monessakin mielessä:
  - kymmeniä tai satoja attribuutteja
  - eri tietojoukoissa voidaan käyttää samoja attribuuttien nimiä (samoista tai samantapaisista käsitteistä)
  - attribuuttien merkitykset (luokittelukriteerit) voivat muuttua ajan kuluessa
  - datajoukosta voi olla useita eri-ikäisiä versioita yhtä aikaa käytössä

- rakenteellinen monimutkaisuus:
  - alkioiden välillä on monenlaisia (ja monimutkaisia) riippuvuuksia
  - alkio voi sisältää rakenteeltaan ja käsittelytavaltaan erilaisia osia
- luokitteluhierarkiat
- erityiset tietotyypit kuten
  - aikasarjat (dataa toistuvista mittauksista)
  - merkkijonot, sekvenssit (esim. DNA)
  - paikkatieto
  - kaaviot, kuvat

Metadatatietokanta

- rakenteeltaan ja käsittelytavaltaan erilaisia tietotyyppejä, kuten
  - tekstiä
  - kaavioita
  - tekstidokumentteja
  - ohjelmia
  - sääntöjä
  - taulukoita
  - viittauksia muihin lähteisiin
  - omat talletustavat ja operaatiot (työvälineet)

- voidaan rakentaa kuten tietokanta
- parhaimmillaan käsitellään samassa ympäristössä kuin varsinaista dataa
- tavoitteena on automatisoida metadatan käsittelyä
- mahdollisia työkaluja myös XML, RDF, ontologiat
  - tietojen vaihdettavuus, yhteiset käsitteistöt

Tutkimusmetadatan piirteitä

- metadata on hyvin monimuotoista
- metadata muuttuu usein
  - kerätään uudentyyppistä dataa
  - käsitellään kerättyä dataa uusilla tavoilla
- monimuotoisuus rakenteessa ja käytössä

  $\Rightarrow$ ei keskitettyä auktoriteettia

  $\Rightarrow$ ei kaikenkattavia standardeja

  $\Rightarrow$ vain rajallinen automatisointi
- datan ja metadatan tuottavat tutkijat
  - tutkijan pitäisi päästä keskittymään sisältöön, ei rakenteeseen

Metadata: yhteenveto

- metadata kuvaa dataa
- metadata on välttämätöntä datan käsittelemiseksi ja ymmärtämiseksi
- tutkimustiedon tuottaminen ja jalostaminen: paljon vaihtoehtoja, täsmällisyys tärkeää

$\Rightarrow$ tarkka ja monipuolinen metadata

- miten raakadata on tuotettu ja kerätty?
- miten dataa on käsitelty ja analysoitu?
- kytkentä tunnettuihin taksonomioihin ja standardeihin

Tiedon mallitus ja talletus

Tiedon mallitus: osan sisältö

- Miten tietoa jäsennetään talletusta varten?
- Millaiset esitystavat ja talletusrakenteet sopivat erityisesti tutkimustiedolle?
- Lähde (esim.): Harri Laine: Johdatus sovellussuunnitteluun ja Tietokantojen perusteet (TKTL:n opetusmonisteita)

Tiedosto

- kaiken tiedon talletuksen perusta
  - monimutkaisemmat talletusrakenteet toteutetaan (yhden tai useiden) tiedostojen avulla
- hierarkkinen muistijärjestelmä
  - keskusmuisti: hakuaika X ns
  - levy: X ms
  - cd-pakka: X s
- käsittelyä varten tieto tuodaan muistiin
- tehokkuuden kannalta on edullista
  - tuoda tietoa kerralla paljon
  - käsitellä muistissa oleva osa kerralla

Järjestämätön peräkkäisrakenne

- tietueet peräkkäin lisäysjärjestyksessä
- lisäys aina tiedoston loppuun
- ei hukkatilaa
- hyvä ratkaisu jos tarvitaan
  - suurta talletusnopeutta
  - staattista talletusrakennetta (= ei poistoja)
  - tehokasta tilankäyttöä
  - peräkkäiskäsittelyä aina samassa järjestyksessä (tai järjestyksellä ei väliä)
- tyypillinen talletusrakenne raakatiedon talletuksessa

Hakupuut

- tietueet (tai avaimet) on talletettu puun lehtiin
- sisäsolmun lapset osittavat sen tietueavaruuden
- puun pitäminen tasapainossa takaa tehokkuuden
  - joka lehteen yhtä pitkä matka ($log_m(n)$)
- lukuisia muunnelmia: mm. B-, R-, LSD-, TV-puut
- erityisiä puurakenteita paikkatiedolle
  - tehokas haku 2–3-ulotteisessa datassa
  - talletuksen lokaalisuus ⇒ tyypillisten operaatioiden tehokkuus

Tietokanta

- suuren tietomäärän pitkäaikainen säilytys
- rakenteinen tieto
  - tietoalkioiden väliset suhteet
  - tiedon looginen rakenne
- tietoon kohdistuvat kyselyt ja operaatiot
- tiedon samanaikainen käyttö
- hallintaan liittyviä tehtäviä on siirretty käyttäjältä tietokannan hallintajärjestelmälle

Tiedostot vs. tietokanta

- säilytys: OK? (varmuuskopiointi)
- rakenteisuus, eheys: käyttäjä hallitsee ja ylläpitää
- kyselyt ja operaatiot: käyttäjä ohjelmoi ja toteuttaa
- samanaikainen käyttö: käyttäjät huolehtivat itse

Tiedonhallinnan peruskäsitteitä

- eritasoiset kaaviot (schema) joilla tietokantaa kuvataan
  - käyttäjän näkymä tietokantaan
  - tietokannan looginen sisältö
  - tietokannan talletusrakenteet
- käyttäjän operaatioita varten tarvitaan
  - määrittelykieli
  - kyselykieli
  - SQL sisältää molemmat komponentit

- työnjako
  - tietokannan käyttö
  - tietokannan ylläpito
  - tietokannan toteutus
- lähestymistapoja mallittamiseen
  - relaatiomallit (relaatiot, attribuutit, relaatiokaaviot)
  - oliomallit (oliot, luokat, tyypit)

Tietokantajärjestelmän osat

- käyttäjä
  - tietokantaan kohdistuvat kyselyt (kyselykieli ja sovellusohjelmat)
  - tietokannan sisällön muutokset
  - tietokantakaavion muunnokset
- tietokannan hallintajärjestelmä
  - kyselyiden käsittely
  - tapahtumien käsittely
  - talletusrakenteen käsittely
- tietojen talletus
  - data + metadata

Tietokannan suunnittelun vaiheet

- 1. laaditaan malli käyttäjän näkökulmasta
  - mitä tietokannassa pitäisi olla
  - miten osat suhtautuvat toisiinsa
  - miten tietokantaa on tarkoitus käyttää
  - käyttäjän käsitteet

- 2. kirjoitetaan malli tietokantajärjestelmän ymmärtämällä kielellä
  - relaatiorakenne, oliorakenne
- 3. muokataan rakenne tehokkaaksi (normalisointi)
  - toistot pois, turhat rajoitukset pois
  - hakemistot

Tietomallin laatiminen

- reaalimaailman asioita, niiden ominaisuuksia ja suhteita
  - kuvaavat nimet
  - havainnollinen esitystapa: kaaviot, teksti
- malli koostuu olioista, joihin voi liittyä
  - attribuutteja
  - relaatioita
  - operaatioita (menetelmiä)
- ominaisuuksiltaan samanlaiset oliot kuuluvat samaan luokkaan

- terminologia vaihtelee riippuen käytettävästä esitystavasta
  - ER-malli: yksilöt, ominaisuudet, ...
  - oliomalli: luokat, oliot, menetelmät, ...

Attribuutti (ominaisuus)

- olion ominaisuus, jolla on jokin arvo
- olioon sisältyvä tieto
- attribuuttiin liittyy tyyppi (arvojoukko), esim.
  - kokonaisluku
  - merkkijono
  - {punainen, keltainen, vihreä}
- attribuutti voi olla myös rakenteinen, esim.
  - osoite: katuosoite, postinumero
- attribuutilla voi olla myös puuttuva arvo

Relaatio

- ominaisuus joka liittää olion toiseen olioon
- olioiden välinen suhde
- relaatio voi liittää toisiinsa samaan tai eri luokkiin kuuluvia olioita
- relaatio voi olla
  - yksi yhteen
  - yksi moneen
  - moni moneen

Operaatio (menetelmä)

- toimenpide joka oliolle voidaan tehdä
- tapa jolla oliota voidaan käsitellä
- operaatiot liittyvät kiinteästi tiedon käyttötarkoitukseen
- operaatioiden kytkeminen olioihin on erityisesti oliosuunnitteluun liittyvä piirre
  - olio ja sen operaatiot muodostavat kokonaisuuden
  - olio "näkyy" muille olioille vain operaatioidensa välityksellä

Luokkahierarkia — periytyminen

- oliolla voi olla samat ominaisuudet kuin jonkin luokan olioilla, mutta lisäksi joukko erityisominaisuuksia ⇒ aliluokka
- aliluokka perii yliluokkansa ominaisuudet
  - attribuutit
  - relaatiot
  - operaatiot
- luokat muodostavat hierarkian
- moniperintä: luokalla on monta yliluokkaa ja se perii kaikkien niiden ominaisuudet

Avain

- attribuutti tai joukko attribuutteja
  - avaimen oltava yksikäsitteinen, eli kahden eri olion avaimet eivät voi olla samat

  $\Rightarrow$ avain yksilöi olion
- usein jokin attribuutti on erityisesti valittu avaimeksi
  - esim. henkilötunnus
- avaineheys: avaimella ei voi olla puuttuvaa arvoa
- oliomallissa jokaisella oliolla on yksikäsitteinen identiteetti: oliotunniste (OID)
- oliomallissa kahden eri olion kaikki attribuutit voivat olla samat

Suunnitteluperiaatteita

- mallin pitää vastata todellisuutta
- mallin pitää olla (mahdollisimman) yksinkertainen

$\Rightarrow$ sovellus- ja tapauskohtainen kompromissi

- minimaalisuus
  - samoja asioita ei pidä toistaa
  - luokkahierarkian hyödyntäminen
- rajoitukset
  - avaineheys, viite-eheys
  - attribuutin arvojoukkoa koskevat rajoitukset

Relaatiomalli

- relaatiotietokantojen perustana oleva yksinkertaiseen malli
- kaikki tiedot esitetään taulukkoina (relaatioina)
- attribuutit = relaatiotaulun sarakkeet
- relaatiokaavio = taulun otsikkorivi, esim.
  - mittaus(järvi, laji, lkm)
- relaatiokaaviot muodostavat yhdessä relaatiotietokantakaavion
- taulun rivi = monikko (tietue): sisältää kutakin attribuuttia vastaavan komponentin, esim.
  - (Tsuolbmajavri, Heterodissocladius maeari, 43)

- kaavion attribuutit esitetään aina samassa järjestyksessä
- attribuutit ovat yksinkertaisia eivätkä koosteisia
- relaation monikot muodostavat joukon
  - kaksi eri monikkoa eivät voi olla identtisiä
  - monikoiden käsittely perustuu niiden attribuuttien arvoihin, ei niiden "sijaintiin taulussa"
- relaatiokaavio määrittelee rakenteen, joka on (yleensä melko) staattinen
- relaatiotaulun sisältö (= relaation ilmentymä) voi muuttua
  - monikoiden lisäykset, poistot, päivitykset

Tietomallista relaatiomalliin

- olioluokka ⇒ relaatiotaulu
- yksinkertainen attribuutti ⇒ relaatiotaulun attribuutti
- rakenteinen attribuutti
  - esim. joukko relaatiotaulun attribuutteja
- olioiden välinen relaatio (yhteys)
  - attribuutti, joka sisältää viitattavan olion avaimen
    - hyvä ratkaisu, jos viitattavia olioita on vain yksi
  - relaatiotaulu, jonka attribuutteina ovat molempien olioiden avaimet
    - hyvä ratkaisu, jos relaatio on monesta moneen

Aliluokka relaatiomallissa

- kutakin aliluokkaa vastaa relaatiotaulu, joka sisältää kaikki attribuutit (sekä omat että peritwritten):
    - kaikki olion ominaisuudet löytyvät yhdestä taulusta
    - yhden attribuutin arvoja monessa taulussa

tai

- kunkin aliluokan relaatiotaulussa on vain ko. aliluokan omat ominaisuudet (ja avain)
    - olion ominaisuuksia monessa taulussa
    - samoja attribuutteja ei esitetä eri tauluissa

tai

- kaikki aliluokat esitetään yhtenä relaatiotauluna
  - komponentti saa puuttuvan arvon, jos attribuutti ei koske tätä monikkoa

Tiedon mallitus: yhteenveto

- talletusvaihtoehtoja: tiedostot, hakupuut ja muut erikoisrakenteet, tietokanta
- tietomallin käsitteitä:
  - attribuutti, relaatio, operaatio, luokkahierarkia
- tavoitteena yksikertainen mutta riittävän todenmukainen malli
- relaatiomalli, oliomalli