METADATAN MERKITYS

TIEDON LAATUUN TIETOVARASTOINNISSA

Heli Junnula
Pro gradu -tutkielma
Tietojenkäsittelytiede
Kuopion yliopiston
tietojenkäsittelytieteen laitos
Kesäkuu 2008

KUOPION YLIOPISTO, informaatioteknologian ja kauppatieteiden tiedekunta
Tietojenkäsittelytieteen koulutusohjelma

HELI JUNNULA, T.: Metatiedon merkitys tiedon laatuun tietovarastoinnissa
Pro gradu -tutkielma, 64 s., 2 liitettä (2 s.)
Pro gradu -tutkielman ohjaajat: FT Virpi Hotti ja TkL Ahti Planman
Kesä 2008

Avainsanat: tietovarasto, tiedon laatu, metadata, metatietomalli, analysointi

Tämä tutkielma käsitteli tietovarastointiin liittyvän metatietomalliin tallennettavan metadatan merkitystä tiedon laatuun tietovaraston kannalta. Tutkielmassa käsiteltiin myös tiedon laatuun vaikuttavia tekijöitä ja sitä, kuinka tiedon jalostuminen informaatioksi ja tietämykseksi tapahtuu. Esimerkkitapauksena käytettiin Kuopion yliopistossa käytössä ja rakenteilla olevia tietovarastoja sekä tietojen analysointiin ja raportointiin käytössä olevia välineitä. Tutkielman yhtenä tarkoituksena oli antaa pohjatietoja metatietomallin valintaa ja metadatan tallentamista varten. Tutkielmassa käsiteltiin yleisimmin käytössä olevat metadatastandardit ja esitettiin metadatalle asetettavia kriteereitä metatietomallin rakentamisen pohjaksi tietovaraston rakentamisen yhteydessä.

Tietovarastoinnilla tarkoitetaan useasta operatiivisesta järjestelmästä kerättyä tietoa, jota voidaan käyttää tietojen raportointiin sekä organisaation toiminnan analysointiin ja seurantaan. Tietovarastossa voidaan säilyttää myös historiatietoja, summatietoja ja tietoja organisaation ulkopuolisista järjestelmistä. Tietovarastoa päivitetään yleensä automaattisesti ja raskaat ajot ajoitetaan yöaikaan, jolloin ei kuormiteta operatiivisten järjestelmien resursseja. Tässä tutkielmassa huomioitiin niitä tietovarastojen lataukseen liittyviä kysymyksiä, jotka vaikuttavat tiedon laadukkaaseen hyväksikäyttöön ja joita voidaan ratkaista hyvin suunnitellun ja toteutetun metadatastandardiin pohjautuvan metatietomallin avulla.

Tietovaraston käytön etuina ovat esimerkiksi tiedon nopea saatavuus, historiatietojen saatavuus, tietojen yhdisteltävyys ja tiedon tarkastelunäkökulman vaihtaminen. Tietovaraston tietojen sisällön ja tietovarastoon liittyvien lataustietojen kuvaamiseen käytetään metatietomallia. Tallennettua metadataa voidaan hyväksikäyttää tehtäessä analysointia ja raportointia operatiivisesta toiminnasta. Käytettäessä metatietomallia saadaan tietovaraston tiedoista päätöksenteon tueksi nopeasti oikeaa ja laadukasta tietoa.

Esipuhe

Tämä tutkielma on tehty Kuopion yliopiston tietojenkäsittelytieteen laitokselle keväällä 2008. Tutkielman ohjaajina toimivat Virpi Hotti ja Ahti Planman, joille haluan osoittaa kiitokseni. Kiitän myös Kuopion yliopiston tietotekniikkakeskuksessa tietohallinto- ja tietojärjestelmäryhmää, jossa olen saanut tehdä mielenkiintoista työtä. Työni kautta olen saanut myös gradu-tutkimukseeni syvyyttä.

Erityiskiitokset Pappalle, jota ilman yliopisto-opiskeluni eivät olisi alkaneet. Rakasta perhettäni haluan kiittää sydämestäni: Tapio - kiitos kannustamisesta ja perheemme hyvinvoinnin huolenpidosta; Milla, Annika ja Jami - kiitos, kun olette suhtautuneet ymmärtäväisesti ja olleet kärsivällisiä.

Kuopiossa 13.6.2008

_____

Heli Junnula

Sisällysluettelo

1 JOHDANTO

Yliopiston tietojärjestelmät ja tietokannat sisältävät runsaasti yksityiskohtaista tietoa muun muassa talous- ja henkilöstöhallinnosta, mutta tiedon käyttäminen päätöksenteon tueksi tai suunnitteluun on hankalaa ja hidasta. Toiminnan suunnittelua ja seurantaa varten tarvitaan tietoa useista tietojärjestelmistä pitkältä aikaväliltä. Myös tarvittavien tietojen yhdisteleminen, löytäminen, rajaaminen ja mahdollisesti vielä täydentäminen tai muokkaaminen toisiinsa integroimattomista tietolähteistä vie turhaan aikaa tai on joskus jopa mahdotonta.

Tämän tutkimuksen tavoitteena on määritellä tietovaraston metatietomallin rakentamisessa huomioon otettavat seikat. Lisäksi pyritään havainnoimaan ongelmakohdat, joilla voi olla merkitystä tiedon uudelleen jalostamisen kannalta. Tutkimuksessa käytetään tietovarastojen rakentamisen lähestymistapana Kuopion yliopistossa käytössä olevaa VATI-tietovarastoa ja tämän pohjalta rakenteilla olevaa Itä-Suomen yliopiston ISTO-tietovarastoa. Esimerkkinä raportointi- ja analysointityökalusta käytetään Kuopion yliopistossa käytössä olevia KASSi-sovelluksien työkaluja ja yliopiston seuranta- ja analysointitarpeita sekä laitos- että yliopiston johtotasolla. Pro gradu -tutkimuksen tuloksia voidaan käyttää rakennettaessa tietovarastoon liittyvää metatietomallia.

Tietovarastoa rakennettaessa on ensin kyettävä arvioimaan ja valitsemaan olennainen tieto, joka halutaan siirtää ja/tai jalostaa tietovarastoon [Nie02, s. 16]. Tietoa siirrettäessä on myös tuotettava tietoa siitä, mitä tietoa tietovarasto sisältää toisin sanoen metatietoa eli metadataa. *Metatieto* (*metadata*) on tietoa tiedosta, eli kuvailevaa ja määrittävää tietoa jostakin tietovarannosta tai sisältöyksiköstä [Wik08]. Metadataan on myös määriteltävä tieto siitä, missä muodossa tieto on tietovarastossa [KHL01]. Metadataan määritellään tietotyyppi mukaan lukien se, miten tieto johdetaan (esim. euromuunnoskerroin), tiedon vaihteluväli eli suurin ja pienin sallittu arvo, tieto siitä voidaanko yksilöivä data korvata toisella vai ei sekä tiedon tarkkuudelle asetettu arvoalue eli millä tarkkuudella tieto on esitettävä. Muita tyypillisiä metatietoja ovat tiedon nimi, sanallinen määritelmä,

tiedon omistaja, pituus, mistä tietojärjestelmästä tieto on peräisin sekä tiedon käyttöoikeudet. [Hov97].

Ensimmäisen kerran puhuttiin operatiivisten kantojen vastakohtana informaatiokannoista tai lyhyesti infokannoista 1980-luvulla. Data Warehouse -termin lanseerasi USA:ssa William Harvey Inmon 1994 ja häntä pidetäänkin tietovarastoinnin isänä [Hov97].

Inmonin määritelmän mukaan tietovarasto sisältää kokoelman tietoja, joille on määritelty tiettyjä ominaisuuksia [Inm05]. Professori Plattner on selittänyt tietovarastoon tallennetun tiedon ominaisuudet seuraavasti (Taulukko 1):

Taulukko 1. Tietovarastoon tallennetun tiedon ominaisuudet [Pla07]

Ominaisuus	**Selitys**
Kohdepainotteinen (subject-oriented)	Tarkoittaa, että tietovarasto on järjestetty tietoja käyttävän yhteisön kannalta merkityksellisiin käsitteisiin, kuten toimittajat, tilaukset, tuotteet ja asiakkaat.
Yhtenäinen (integrated)	Tarkoittaa tietovarastoon tallennetun tiedon fyysistä yhtenäistämistä ja koossa pitämistä. Yhtenäistäminen sisältää useita näkökantoja tietovarastoinnista, kuten nimiöimiskäytänteet ja tiedon esitysmuodon sopiminen.
Hitaasti muuttuva (non-volatile)	Tietovarastoon tallennettuja tietoja ei koskaan muuteta tai poisteta, vaan tietoja pidetään tallennettuina tulevaisuuden raportointia varten.
Aikasidonnainen (time-variant)	Tietovarastoon tallennettujen tietojen pohjalta voidaan luoda tilannekatsauksia, jotka kattavat tietoja pitkältä aikaväliltä. Tietovarastoissa pidetään yleisesti tallessa tietoja viidestä kymmeneen vuoteen.

Tietovarastot sisältävät sitä käyttävän organisaation tai virtuaalisen yhteisön tarvitsemat tiedot siinä muodossa, että tietojen perusteella voidaan tehdä johdon tarvitsemia päätelmiä ja analyysejä. Kun tieto on tallennettu oikein, voidaan metatiedoilla täydennettyä tietovarastoa käyttää korvaamaan kalliit ja aikaa vievät *johdon tietojärjestelmät (Business Intelligence tools)*. [Pla07]

Johdon tarvitsemat tiedot ovat luonteeltaan toimintaa ohjaavia ja analyyttisiä, ja ne pohjautuvat operatiivisista järjestelmistä saataviin tietoihin. Tietojen tarkasteluun käytetään raportteja, analyysejä ja graafisia kaavioita. Tiedon luonne operatiivisissa järjestelmissä on tapahtumakeskeinen eli järjestelmään tallennettu tieto liittyy johonkin tietyllä ajanhetkellä tapahtuvaan toimintaan. Esimerkiksi opintosuoritus kohdistuu opiskelijan suorittamaan opintojaksoon tai tenttiin.

Tiedon luonteen vertailu helpottaa ymmärtämään operatiivisten järjestelmien tietojen ja tietovaraston tietojen ominaispiirteitä. Tietovarastosta tuotettavat analyysit tarvitsevat valtavan määrän tietoja operatiivisen järjestelmän tapahtumista, jotka on suunniteltu operatiivisen toiminnan tueksi. Tietovarastojen kyselyillä pyritään tuottamaan johdon tarpeita vastaavia tietoja strategisen päätöksenteon tueksi, kun taas operatiiviset järjestelmät ovat tapahtumaorientoituneita.

Plattner esittää seuraavan taulukon (Taulukko 2) mukaisesti operatiivisen tiedon ja tietovaraston tiedon luonteen [Pla07].

Taulukko 2. Operatiivisen tiedon ja tietovaraston tiedon luonne [Pla07]

Operatiiviset järjestelmät		Tietovarasto
Tapahtuman liittyvät tiedot	vs.	Analyyttiset tiedot
Toimintaan liittyvät tiedot	vs.	Taktiset tai strategiset tiedot
Järjestelmäriippuvaiset tiedot	vs.	Kohdekohtaiset tiedot
Raportointitiedot	vs.	Satunnaiskyselyt
OLTP	vs.	OLAP
Yksityiskohtaiset tiedot	vs.	Summatasoiset tiedot
Nykyhetken tiedot	vs.	Historiatiedot (2-7 vuotta)
Normalisoidut tiedot	vs.	Denormalisoidut tiedot
Tietoja voidaan muuntaa	vs.	Muunnetun tiedon tallennus

Seuraavaksi esitellään taulukossa 2 kuvatut tiedon luonteiden käsitteet. Operatiivisesta järjestelmästä saatavat `tapahtumiin liittyvät tiedot` muuttuvat `analyyttisiksi`, kun `toimintaan liittyvät tiedot` kootaan raporteilla ja analyyseillä päätöksentekoa varten `taktisiksi tai strategisiksi` tiedoiksi tietovarastossa. Operatiiviset järjestelmät eivät tue `satunnaiskyselyitä` ja pitkiin aikasarjoihin perustuvia analyysejä, vaan operatiivisen järjestelmän perustehtävänä on auttaa käyttäjiä siinä perustehtävässä, johon ne on ensisijaisesti tarkoitettu. Mikäli operatiivisiin järjestelmiin halutaan luoda yrityksen johdon satunnaiseen tarpeeseen tehtyjä `raportointeja`, niistä jäisivät kuitenkin pois tietoon liittyvät muissa perusjärjestelmissä olevat tiedot.

Operatiivisissa järjestelmissä voidaan tutkia vain `järjestelmäkohtaisia tietoja`, kun taas tietovarastosta tietoa tuotettaessa voidaan usean järjestelmän tietoja yhdistellä ja tiedon tarkastelunäkökulmana käytetään `kohdekohtaisia tietoja`. Tietoja voidaan raportoida perusjärjestelmistä *`OLTP (on-line transaction processing)`* -menetelmiin kuuluvilla välineillä, kuten esimerkiksi järjestelmäkohtaiset raportit. Tietovarastojen tietoja tutkitaan yleensä analysointia ja raportointia varten hankituilla *`OLAP (on-line analytical processing)`* -välineillä, joilla tietojen yhdisteleminen ja graafinen raportointi on nopeaa ja helppoa. Operatiiviset jär-

jestelmät ovat tapahtumaorientoituneita tietokantoja, joissa tiedot ovat yksityiskohtaisia, kun organisaation johto on usein ensin kiinnostunut summatason tiedoista. Summatason tiedoista on voitava pureutua mielenkiintoisiin kohteisiin tarkemmin aina tapahtumatasolle saakka ja voitava palata summatasolle. Johdolle on usein toiminnan suunnittelussa myös tärkeää muuttaa tarkastelunäkökulmaa, mikä ei onnistu helposti operatiivisissa järjestelmissä.

Operatiivisissa järjestelmissä ylläpidetään nykyhetken tietoja ja tietojen muuttuessa tieto voidaan korvata muuttuneella tiedolla. Tietovarastossa ylläpidetään myös historiatietoja, joista voidaan selvittää kunkin ajanhetken kulloinkin voimassaoleva tiedon sisältö. Operatiivisissa järjestelmissä tieto on *normalisoitua* eli tiedon toistoa ei ole, koska tieto on haluttu pitää helposti hallittavana ja tiedon toisto aiheuttaa tietojen tallentamisen moneen kertaan, jolloin virheriski kasvaa. Tietovarastossa tieto on *denormalisoitua* eli tietoja voidaan monistaa dimensiotauluihin, jolloin kyselyt saadaan tehokkaammiksi. Tietojen monistaminen tapahtuu automaattisesti latauksen yhteydessä, jolloin tietojen yhteneväisyys säilyy.

Operatiivisissa järjestelmissä tietoa voidaan muuntaa, mutta tietovarastossa tallennetaan tiedon muutosajankohta ja siihen liittyvät muutosta koskevat tiedot ja tieto voidaan palauttaa haluttuun ajankohtaan tai tutkia tiedon muutokseen kohdistuvia tietoja.

Analysointivälineet mahdollistavat myös tiedon alaspäin *porautumisen (drill-down)* ja ylöspäin *karkeistamisen* tai *yleistämisen (roll-up)* ilman, että käyttäjä joutuu ajamaan uusia raportteja. Operatiivisten järjestelmien raportointi on yleensä vakioitua siten, että esimerkiksi taloushallinnon tiedot ajetaan raporteille tarvittaessa tai määräaikoina. Mikäli jotakin halutaan tarkentaa tai tutkia syitä tapahtumalle, on usein ajettava uusia raportteja tai tutkittava operatiivisen järjestelmän tapahtumia yksi kerrallaan.

Tiedon hyväksikäyttö edellyttää yksittäisen tiedon (data) tallentamista, muokkaamista ja yhdistelemistä (informaatio) ennen kuin tiedosta voidaan jalostaa tietämystä. Tietovarasto onkin informaatiokanava, joka antaa toiminnalle tukea ja suuntaviivoja. Tietojen

määrän valtava kasvu on generoinut tarpeen uusille tekniikoille ja automaattisille työkaluille, jotka voivat käsitellä ja opastaa käyttäjiä jalostamaan tietoa ja informaatiota tietämykseksi [HaK06].

Toisessa luvussa kuvataan tietovarastoinnin perusidea ja selitetään tiedon eri merkityksiä. Luvussa esitellään tiedon siirtoon liittyvät vaiheet (uuttaminen, puhdistus, lataus). Lisäksi luvussa pureudutaan tiedon olemukseen tarkemmin käsiksi, määrittelemällä tieto, informaatio ja tietämys. Tiedon laadulla on merkittävä rooli tiedoista koottavien analyysien ja raporttien pohjalta tehtävien johtopäätösten oikeellisuudelle. Luvussa käydään läpi tiedon laatuun vaikuttavia seikkoja.

Kolmannessa luvussa määritellään metadata ja esitetään metadata luokitteluja. Luvussa käsitellään myös yhteentoimivuutta. Tietokantojen tallennukseen liittyviä ja erityisesti Dublin Core- ja LOM-standardeja esitellään omassa aliluvussaan ja tarkemmin paneudutaan OMG:n CWM-standardiin. Lopuksi käsitellään metatietomallia ja standardien käyttämistä metatietomallin pohjalla.

2 TIETOVARASTO

Organisaatiot keräävät tietoja toimintansa suunnittelua ja seurantaa varten useisiin tietokantoihin. Joidenkin tietojen tallentaminen perustuu lakiin (kirjanpitovelvollisuus), kun toiset tiedot ovat oleellisia organisaation toiminnan kannalta (yliopistojen opiskelijatietojen tallentaminen). Nykyisellä tietotekniikalla on tietojen tallennus tullut vaivattomaksi ja informaation määrä on kasvanut suureksi. Tästä seuraa, ettei kaikkea tietoa saada helposti käyttöön tai viive tiedon saamiseksi kasvaa liian suureksi. Saadakseen toimintansa tueksi tarvitsemaansa tietoa organisaatiostaan, on tietojen käyttäjän kerättävä tietoja useista eri tietolähteistä, tietojärjestelmistä ja niitä on pystyttävä yhdistelemään, poimimaan ja tekemään erilaisia analyyseja saatavan tiedon pohjalta. Organisaation johdolle on tärkeää saada pitkän aikavälin analyysejä toiminnasta aikasarjoina ja toiminnan ongelmakohdat on pystyttävä helposti havaitsemaan. Tietojen on oltava sellaisessa muodossa, että ongelmakohtiin on pystyttävä pureutumaan tiedon tarkimmalle tasolle saakka helposti ja vaivattomasti. Operatiiviset tietokannat sisältävät yrityksen perustoimintaan tarkoitettua toiminnallista tietoa esimerkiksi osto- ja myyntitapahtumista, varastotilanteesta, suoritetuista tutkinnoista, julkaisuista ja niin edelleen. Operatiivisen tietokannan yksittäiset tiedot eivät ole sinällään riittäviä päätöksenteon tueksi, vaan tietoja on voitava yhdistää useasta tietokannasta, tiedolle on voitava tehdä erilaisia muunnoksia ja tietoa on ryhmiteltävä, että siitä saataisiin kiinnostavaa ja tarpeellista tietoa.

Tietovaraston isänä tunnetun W. H. Inmonin määrittely tietovarastosta [Inm05] on seuraava:

> ”*Tietovarasto (data warehouse)* on kohdepainotteinen, yhtenäinen, hitaastimuuttuva ja aikasidonnainen kokoelma tietoja, jotka tukevat päätöksen tekoa.”

Tiedolle asetetut ominaisuudet on esitetty luvussa 4, jossa käsitellään myös tiedon luonteen muuttumista tietovarastossa. Tietovarasto sisältää tietoja, joita tietohallinto, organisaation johto ja analyytikot voivat käyttää paremman ja nopeamman päätöksenteon tu-

kena [JLV02]. Tietovarastoon siis kerätään tietoa operatiivisista tietokannoista. Tietovarastosta saatavan tiedon avulla toimintaa voidaan paremmin suunnata ja kohdentaa, kun prosessien ongelmakohdat on helpompi havaita. Myös tiedon saannin nopeus auttaa ongelmatilanteisiin nopeaa reagoimista jo ennen kuin vahinkoa on ehtinyt syntyä toiminnan kannalta liikaa. Tietovarastossa eri operatiivisten järjestelmien tuottama tieto tallennetaan yhteen tietokantaan.

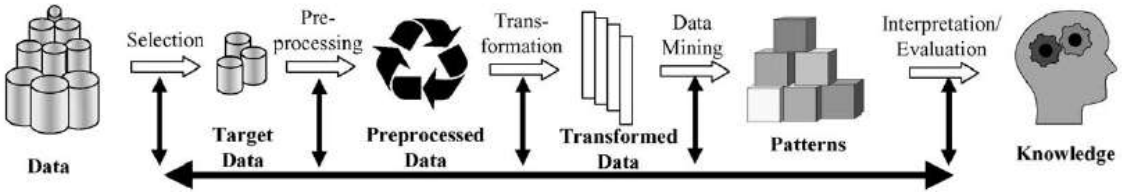
Tietovarastoon kerätyn tiedon luonne on pysyvää ja sitä täydennetään eräajotyyppisesti, yleensä aikaan, jolloin operatiiviset kannat eivät ole käytössä. Operatiivisia kantoja päivitetään ja luetaan jatkuvasti ja ne sisältävätkin tietoa lähes reaaliaikaisesti. Viivettä aiheuttaa tietenkin tallennuskapasiteetin hitaus. Tietovarastoon tietoa on koottu historiallisesti pitkältä aikaväliltä ja tiedot on koottu tiettyyn ajanjaksoon saakka. Tietovarastoon tehdään kyselyjä, joiden vasteaika muodostuu tärkeäksi mittariksi [Sor03].

Tietovarastossa tallennetaan tapahtumiin kohdistuvia yksittäisiä *tietoja (data*). Tietotekniikassa ohjelman käyttämät tiedostot tai muistialueet sisältävät dataa [Wik08], eli operatiivisten järjestelmien koodistot ja niihin liittyvät tapahtumat voidaan käsittää tiedoksi.

Tietovarastosta voidaan käyttää myös nimitystä informaatiotietokanta tai informatiivinen tietokanta [Gar98]. Tällöin on kyseessä perustiedon yhdistelemisestä syntyvä informaatio. *Informaatio (information)* on tietoa, joka esitetään oikeassa yhteydessä ja joka on merkityksellistä käyttäjän sen hetkiseen tarpeeseen [Hov97].

Tietämys (knowledge) koostuu informaatiosta analysoinnin tuloksena saadusta tiedosta. Tiedon hyväksikäyttö edellyttää yksittäisen tiedon (data) tallentamista, muokkaamista ja yhdistelemistä (informaatio) ennen kuin tiedosta voidaan jalostaa tietämystä. Tietovarasto onkin informaatiokanava, joka antaa toiminnalle tukea ja suuntaviivoja.

Tiedon muuntuminen informaatioksi ja informaatiosta tietämykseksi vaatii tiedon keruuta ja oikeanlaista tiedon yhdistelemistä. Törmäsen mukaan tietovarastointi onkin tärkeä tiedon ja tietämyksen hallitun johtamisen teknologinen osa-alue [Tör99].



Kuva 1. Tiedon muuntuminen tietämykseksi [HoH02]

Tietovarastossa tietoa jalostetaan, jolloin siitä saadaan informaatiota. Informaatiota hyväksikäyttävä ihminen, joka tulkitsee saamaansa informaatiota esimerkiksi toiminnan kuvaukseen tai suunnitteluun, muuntaa tiedon tietämykseksi. Kuva 1 esittää tiedon muuntumisprosessit [HoH02].

- *Tietojen valinta (selection)*. Valitaan operatiiviset järjestelmät ja niissä sijaitsevat tiedot, joita käytetään tietojen analysoimiseen.
- *Tietojen esiprosessointi (preprocessing)*. Tietoja puhdistetaan, täydennetään ja korjataan ennen analysoinnin tekemistä.
- *Tietojen muuntaminen (transformation)*. Tiedot muutetaan yhteismitallisiksi.
- *Tiedon louhinta (data mining)*. Pyritään löytämään mielenkiintoiset ja säännönmukaiset tiedot, jotka voivat olla hyödyllisiä tai tarpeellisia tietojen analysoinnin kannalta.
- *Tiedon tulkinta ja arviointi (interpretation, evaluation)*. Analyysien ja raporttien tulkinta muuntaa tiedosta koostetun informaation *tietämykseksi (knowledge)*.

Kuopion yliopistossa talouden suunnittelu- ja seurantajärjestelmiin siirretään tietoja tietovarastosta, jonne niitä on koottu paitsi talousjärjestelmästä myös muista operatiivisista järjestelmistä. Budjetointiin käytettävässä RaTaS-budjetointijärjestelmässä on mahdollisuus käyttää opintohallinnon järjestelmästä saatuja tietoja avuksi budjetoitaessa seuraavan kauden kuluja ja tuottoja. Budjetointijärjestelmässä on tuotettu suunnittelun avuksi tieto siitä, mitä kursseja kyseessä oleva vastuualue on järjestänyt edellisenä vuonna ja kuinka paljon tunteja siihen on sisältynyt. Tulevaa kautta suunniteltaessa voidaan syöttää budjetoidut tunnit kursseittain ja järjestelmä laskee syötettyjen tuntien ja

määritellyn tuntihinnan perusteella budjettiluvun. Vaikka budjetointijärjestelmä on tehty tukemaan talousjärjestelmää, on siinä kyetty hyödyntämään myös opintohallinnon järjestelmästä saatuja tietoja. Tiedon operatiivinen luonne on siis muuttunut strategiseksi, kun sitä on käytetty toiminnan suunnitteluun strategisen päätöksenteon tueksi.

Tietovaraston suunnittelu vaatii tiedon tuottajalta tallennettavan tiedon ja sen tuottaneen organisaation hyvää tuntemusta. Tietovaraston tarkoitus ja käyttö voidaan toteuttaa hyvin vain, jos tiedetään, mitä tietovarastolta tahdotaan ja mihin sitä tullaan käyttämään. Metadatan käyttö analysoinnissa ja raportoinnissa auttaa tiedon jalostamisessa ja mahdollistaa sen, että tietämys perustuu oikeaan ja laadukkaaseen tietoon.

Kappaleessa 2.1 käsitellään tietovaraston muodostusta, joka sisältää tiedon uuttamisen, puhdistamisen ja lataamisen vaiheet.

2.1 Tietovaraston muodostus

Tiedon keruusta ja karsinnasta käytetään nimitystä *uuttaminen* [RaD00], jonka jälkeen tieto on *puhdistettava* [Dev97] ja tämän jälkeen tieto voidaan *ladata* [WuB97] tietovarastoon. Englanninkieliset termit ovat *Extract, Transfer* ja *Load* ja näin tiedon siirrosta tietolähteestä muutosvaiheineen ja sen tallentamista tietovarastoon kutsutaan nimellä *ETL-vaihe.* Seuraavissa luvuissa on tietovarastoon siirron vaiheet tarkemmin.

Tietovarastoon siirrettävä tieto poimitaan erilaisista lähteistä ja tätä vaihetta kutsutaan tiedon *uuttamiseksi (extract).* Uuttamisesta on suomalaisessa kirjallisuudessa alettu käyttää myös nimitystä *poiminta* ja *tiedon eristys*. Keräämisvaiheessa operatiivisten tietojärjestelmien tiedoista kerätään olennainen ja tietovaraston kannalta tärkeä tieto ja tiedosta voidaan poistaa tarpeettomia yksityiskohtia.

Tiedon siirtämistä varten voi uutettu tieto olla joko *täydellinen tilannevedos (master file, snapshot)*, jolloin uutettu tieto sisältää tietovaraston päälle kirjoitettavan tiedon. Toinen mahdollisuus on, että tietovarastoon viedään vain muuttuneet ja uudet tiedot, jolloin kyseessä on *muutostiedosto (delta file)*. [Jok02].

Tiedon luonteesta riippuu, onko tieto syötettävä tietovarastoon välittömästi vai voidaanko tietovarastoa täydentää eräajotyyppisesti päivittämällä se esimerkiksi yöaikaan, jolloin operatiivisten järjestelmien käyttö ei häiriinny. Välitön tiedonsiirto voidaan toteuttaa *herättimillä (trigger)* tai sovellusavusteisella operatiivisen tietokannan lokiin perustuvalla uuttamistekniikalla. Viiveellä toteutetussa tiedonsiirrossa taas voidaan käyttää aikaleimoihin ja tiedostojen keskinäiseen vertailuun perustuvaa uuttamista [RaD00].

Ennen tiedon siirtämistä tietovarastoon jatkokäyttöä varten, on tietoa ehkä muokattava oikeaan muotoon. Operatiivisissa järjestelmissä tieto voi olla myös puutteellista tai virheellistä. Ennen tietojen latausta tietovarastoon tapahtuvaa tiedon muuntamista kutsutaan *tiedon puhdistamiseksi (cleansing data).* Tätä tarkoitusta varten luodaan *muunnosalue (staging area).* Muunnosalue voi olla tietovaraston osana tai erillisenä tietokantana tai muunnos voidaan suorittaa erillisellä tietojärjestelmällä, esimerkiksi tekstitiedostossa. Tiedon muunnokset voivat sisältää tietotyyppien muuntamista tietovarastokannan vaatimaan muotoon ja puutteellisten tai virheellisten tietojen merkitsemistä näille varatuilla tiedoilla. Tietoja voidaan myös summata halutulle tarkkuustasolle tai hakujen nopeuttamiseksi. Tiedon puhdistamisen jälkeen puhdistettu tieto luetaan tietovaraston varsinaisiin tauluihin, joita käytetään tiedon hakuun ja raportointiin.

Puhdistetut tiedot *ladataan (loading)* tietovarastoon, jonka jälkeen ne ovat käytettävissä. Tiedon lataamiseen on eri tietokantatoimittajilla olemassa omia välineitään. Joissakin tietovarastointiratkaisuissa puhutaan tiedon lataamisen sijaan tiedon "vetämisestä". Tällöin tietovarasto on suorassa kytkennässä operatiivisiin tietokantoihin. Tämä vaihtoehto takaa tiedon reaaliaikaisuuden, mutta rasittaa operatiivista kantaa ja saattaa aiheuttaa operatiivisen tietojärjestelmän käytön jähmeyttä ja hidastumista. Toinen seikka, joka tässä ratkaisussa on otettava huomioon, on se, että tiedot saattavat olla vielä keskeneräisiä ja näin tiedon hyväksikäyttö saattaa johtaa vääriin johtopäätöksiin.

Yleisemmin suositaan operatiivisten järjestelmien ja tietovaraston erillään pitämistä. Metadatan merkitys onkin olennaisen tärkeä. Tietovaraston tiedoista tuotettujen tietojen ajantasaisuus riippuu siis siitä, kuinka usein tietoja siirretään tietovarastoon. Metatie-

doista saadaan selville ajan jakso, jota tiedot koskevat. Myös koodistojen muuttumishetket voidaan selvittää metatiedoista.

Tietojen lataaminen eräajotyyppisesti on suositeltavaa, koska latauksen aikana tietovarasto ei ole käytettävissä. Usein tietovarastoihin lataukset suoritetaankin yöaikaan, etteivät operatiivisten järjestelmien käyttäjät häiriinny.

2.2 Tiedon laatu

Tiedon laatu (data quality) on tärkeä osatekijä tietovaraston onnistumisen kannalta. Vain oikeilla ja luotettavilla tiedoilla kootusta informaatiosta voidaan tehdä oikeita päätöksiä. Organisaation johto voi asettaa tietovaraston rakentamisen kyseenalaiseksi, ellei se saa sieltä luotettavaa tietoa. Tiedon laatuun onkin syytä kohdentaa resursseja, sillä menetetty luottamus on vaikeaa saada takaisin.

Tiedon laatuun voidaan määritellä kuuluvaksi tiedon eheys, kattavuus, yhtenäisyys ja aukottomuus [MaW06]. Lisäksi tiedoista keräävän analyysien tai raporttien tekijän on helpompaa kerätä juuri oikeaa tietoa, jos tiedolle on määritelty omistaja ja hän voi käyttää tietoon liittyvää hiljaista tietoa. Tiedon omistaja vastaa tietojen oikeellisuudesta ja päättää tiedon käyttöoikeuksista [HYK01]. Tiedon validiteetilla ja mittauksen tarkkuudellakin on merkitystä, esimerkiksi kun informaatio on tieteellisen tutkimuksen välineenä. Seuraavissa alaluvuissa on tiedon laatua tarkasteltu edellä kuvattujen ominaisuuksien kannalta.

2.2.1 Koodistojen yhtenäisyys

Koodistojen eheys on myös säilytettävä siten, että samaa koodia ei oteta uudelleen käyttöön eri merkityksessä. Koodille annettu nimi voi tietenkin muuttua esimerkiksi henkilön vaihtaessa sukunimeään, mutta alkuperäinen henkilökoodi viittaa kuitenkin aina samaan henkilöön. Esimerkkinä Kuopion yliopistossa tiedekuntakoodi otettiin uudelleen käyttöön, kun yksi tiedekunta oli lopettanut toimintansa ja perustettiin uusi tiedekunta. Raindance-järjestelmässä oli annettu hammaslääketieteelliselle tiedekunnalle

koodiarvo 2. Kun uusi informaatiotekniikan ja kauppatieteiden tiedekunta aloitti toimintansa, otettiin koodiarvo uusiokäyttöön tälle tiedekunnalle. Tämä tehtiin siitä syystä, ettei haluttu koodistoon jäävän tyhjiä välejä, kun raportoinnissa tulostetaan myös tiedekuntakoodit. Kun tiedekuntien tietoja analysoitiin RASSi-talouden suunnittelu- ja seurantajärjestelmällä, oli RASSissa käytössä viimeisin tieto tiedekunnista ja niiden koodiarvoista. Kun tietoja tarkasteltiin pitkältä aikaväliltä, tulivat hammaslääketieteellisen tiedekunnan taloustiedot koodin 2 eli informaatiotekniikan ja kauppatieteellisen tiedekunnan alaisuuteen. Asiantunteva analyysin laatija huomasi virheen ja korjasi tiedot ennen kuin ne ehtivät yliopiston johdolle. Virhe olisi vältetty, mikäli koodiarvoksi olisi annettu uudelle entiteetille oma koodinsa ja analyysin teossa olisi huomioitu tiedon metadataan kuuluva tiedon alkamis- ja päättymispäivämäärä.

Lisäksi joillekin tiedoille voidaan antaa myös koodiston muotoon liittyviä lisämerkityksiä ilman, että tätä olisi kirjattu mihinkään. Operatiivisissa järjestelmissä tiedon koodiarvoon voidaan tallentaa hiljaista tietoa. Näin on esimerkiksi Raindance-järjestelmässä tehty taloushallinnon projektikoodin yhteydessä, jolloin projektikoodin alkuosa kertoo projektin tyypin. Esimerkiksi 928-alkuiset projektit on varattu vuodelle 2008 oleville perusvoimavaraprojekteille, 1-alkuiset projektikoodit on varattu Suomen Akatemian rahoittamille projekteille ja 68-alkuiset projektinumerot ovat yliopiston sisäisiä projekteja. Tätä tietoa ei ole tallennettu mihinkään, vaan tieto on vain sovittu ylläpidettäväksi näin. Tämän kaltainen hiljainen tieto ei tule esiin, muutoin kuin talousjärjestelmän toimintaprosessien ja tiedon tallentamiseen liittyvien sääntöjen syvän tuntemuksen kautta. Koodien arvoihin ei pitäisi sisällyttää muuta tietoa kuin tiedon yksilöinti. Koodiarvo tulisi olla yksikäsitteinen ja kaikki muu tieto tulisi sijoittaa omiin kenttiinsä. Mikäli koodiarvon sisältämä tieto sisältää muutakin tietoa kuin pelkän tiedon yksilöinnin, voisi metadataa käyttää myös koodiarvojen sisältöön liittyvään kuvaukseen.

Tiedon metadata (Luku 3.1) ja master data (Luku 3.3) olisi huomioitava analyysejä ja raportteja tehtäessä. Kuten edellisessä kappaleessa kerrottiin, voi koodistojen uusiokäyttö aiheuttaa virheitä. Tietojen laatuun on alettu kiinnittää myös tietovarastoinnin yhteydessä yhä enemmän huomiota, kun on huomattu, ettei tiedon määrä ole riittävä tae tiedon hyväksikäytölle, vaan tiedon luotettavuus on pystyttävä takaamaan.

2.2.2 Tiedon omistajuus

Tiedon omistaja (data owner) on henkilö tai organisaatioyksikkö, joka luo tai tuottaa tiedon. Mikäli tieto syntyy useassa organisaatioyksikössä, on tehtävä yhteinen sopimus siitä, kuka on tiedon omistaja ja siten myös vastuussa tiedon sisällöstä ja sen oikeellisuudesta. Tiedon omistaja myös määrittelee tiedon käyttövaltuudet, eli kenellä tai keillä on oikeus saada tieto käsiinsä. Esimerkiksi henkilötietolaki [FIN99] ja organisaation tietoturvasäännökset voivat määritellä tiedon sisällön salaiseksi. Tiedon omistajan vastuulla on huolehtia operatiivisen järjestelmän tietojen oikeellisuudesta ja siitä, että koodistoissa käytetään samoja koodeja kuin tiedon omistajat ovat määritelleet. Tiedon omistaja voi myös päättää tiedon käsittelylle yhdenmukaiset käsittelysäännöt, jolloin tiedon käsittely antaa kaikille sitä käyttäville saman tuloksen.

Samaa koodiarvoa ei oteta uudelleen käyttöön. Tietovarastossa tietoja voidaan tarkastella eri näkökulmista, kun operatiivisissa järjestelmissä tietoja tarkastellaan tapahtumittain joko näytöltä tai raporteilta [HYK01]. Tietovarastotarkastelu paljastaakin helpommin puutteelliset tai virheelliset tiedot. Tietovarasto mahdollistaa myös useiden perusjärjestelmien tietojen yhdistelemisen. Tiedon omistajan vastuulla on tiedon oikeellisuus ja eheys.

Seuraavissa luvuissa tarkastellaan tiedon laatua tarkemmin pureutumalla tiedon eheyteen, tiedon historiointiin, hiljaiseen tietoon, aineiston kattavuuteen, tiedon validiteettiin ja mittauksen tarkkuuteen lähemmin. Luvuissa esitetään myös esimerkit kulloinkin käsiteltävään aihealueeseen käyttäen esimerkkitapauksena Kuopion yliopiston käytössä olevia VATI-tietovarastoa ja KASSi-sovelluksia sekä rakenteilla olevaa Itä-Suomen yliopiston ISTO-tietovarastoa.

2.2.3 Tiedon eheys

Wikipedian mukaan *Tiedon eheys (Data integrity)* tarkoittaa, että tiedot ovat keskinäisesti yhteensopivia ja tiedot ovat oikeita ennalta annettuihin ehtoihin nähden [Wik08]. Organisaation hallinnolle (tietojen hyväksikäyttäjille) tiedon eheys tarkoittaa, että voidakseen luottaa tietoon, tiedon on oltava oikeaa ja reaaliaikaista. Tieto on eheää vain

kun se on täsmällistä, täydellistä, ajatonta, voimassaolevaa ja tiedon siirto prosessit on varmennettu [FlS005]. Tiedon arvo perustuu sen oikeellisuuden luotettavuuteen.

Tietovarastoon latauksen yhteydessä tarkistetaan tiedot ja puuttuvat tiedot voidaan täydentää tai merkitä puuttuvan tai virheellisen tiedon koodilla. Kenttiä ei kuitenkaan jätetä tyhjiksi. Kuopion yliopiston VATI-tietovarastossa on puutteellisen tiedon koodiksi valittu yhteneväisesti kaikissa kentissä 999999999999998 ja virheellisen tiedon koodiksi 999999999999999. Tietojen siirron yhteydessä tyhjät tiedot korvataan näillä koodiarvoilla ja tietovarastosta voidaan ajaa raportit niistä tiedoista, joissa on virheellistä tai puuttuvaa tietoa. Esimerkiksi Raindance-talousjärjestelmässä on tallennettu PROJ-käsitelajiin kytketty tieto projektin rahoittajasta (RAH). Mikäli tieto on jostakin syystä virheellinen eikä tietosisältö vastaa tietovarastoon määriteltyä, kyseiseen kenttään täydennetään 999999999999999 ja analysoija voi tietovarastoraportin avulla pyytää talousyksikköä korjaamaan tiedon. Taloushallinnon projektitiedon käsittelystä on myöhemmin esimerkki luvussa 2.3, jossa on kuvattu tarkemmin PROJ-käsitelaji ja siihen liittyvät kytkennät.

2.2.4 Tietojen historiointi

Tietojen analysoinnissa ja trendien seuraamisessa myös historiatietojen merkitys on tärkeä. Usein tiedot kuitenkin muuttuvat ajan mukana. Joitakin tietoja tulee lisää, toisten tietojen sisältö tai koodisto voi muuttua. Historiatietojen seuraamista voidaan tarkastella joko niin kuin tiedot olivat jonakin tiettynä ajanhetkenä tai siten, että historiatiedot konvertoidaan vastaamaan nykyhetken tietorakenteita.

Metatietomalliin tallennettavan tiedon päivämäärän alkamis- ja päättymispäivämäärillä voidaan hallita historiatietoja siten, että tietovarastosta saatavat tiedot vastaavat todellisuutta siten kuin tietoja halutaan tutkia.

Kuopion yliopiston VATI-tietovarastossa on historiointi toteutettu tietovaraston dimensiotauluissa tallentamalla tiedon latauspäivämäärän (= tiedon alkamispäivämäärä) ja tiedon muuttumispäivämäärän (= tiedon päättymispäivämäärä) lisäksi operatiivisessa järjestelmässä oleva tiedon muuttumispäivämäärä sekä muuttaja. Koska tietovarastosta

ei poisteta tietoja, eikä tietoja päivitetä päällekirjoittamalla, vaan tietoja ladataan aina vanhojen tietojen perään, nykyisen voimassaolevan tiedon hallintaan on perustettu oma kenttänsä (YP_NYKYINEN). Kentän arvoksi asetetaan uusimmalle tiedolle arvon 1 ja vanhentuneet tiedot saavat tämän kentän arvoksi 0. Näin voidaan hakea kunkin tiedon arvo halutulla ajanhetkellä. Seuraavassa kuvassa on esitetty tietovaraston taloustiedolle vastuualue luotu tietokantataulu.

Kuva 2. VATI-tietovaraston vastuualuetaulu

Ylläolevassa kuvassa (Kuva 2) on VATI-tietovaraston vastuualuetaulu, johon tiedot siirretään joka yö Raindance-taloudenohjausjärjestelmästä tietovarastoon. Dimensiotaulu sisältää myös vastuualuetietoon liitetyt rakenteelliset tiedot, toisin sanoen ne tiedot, mihin hierarkiatasoihin vastuualue kuuluu. Tietovaraston tietojen hallintaa varten olevat kentät sisältävät YP_-etuliitteen ja ne täydennetään automaattisesti tietojen siirron yhteydessä. Nämä tiedot voidaan luokitella metadataksi.

Metatietoihin tallennetuilla koodien voimassaolopäivämäärillä voidaan taata tietojen paikkansapitävyys. Raporteilla ja analyyseissä voidaan käyttää kullakin ajan hetkellä voimassa ollutta tai voimassa olevaa tietoa.

2.2.5 Hiljainen tieto

Operatiivisten järjestelmien tiedot sisältävät runsaasti *hiljaista tietoa (tacit knowledge)*, jonka merkitys tietoja analysoitaessa on huomioitava. Operatiivisissa järjestelmissä voi tapahtumiin olla sidottuna toimintaan, menettelytapoihin tai rutiineihin liittyvää hiljaista tietoa. Hiljaisella tiedolla tarkoitetaan sitä kunkin osa-alueen tietämyksen tai tietosisällön sisältämää tietoa, joka tulee esiin vain osa-alueen asiantuntemuksen mukana. *Selkeä tieto (explicit knowledge)* voidaan määritellä helposti määriteltäväksi, sanalliseksi ja helposti koodattavaksi, kun taas *hiljainen tieto* liitetään enemmänkin tunteeseen, kokemukseen tai intuitioon ja on sen vuoksi vaikeampaa pukea sanoiksi ja ilmaista toisille [Mit03].

Esimerkiksi julkishallinnon kirjanpitoon liittyy paljon valtion ohjausta ja säännöksiä, joita kirjauksia tehtäessä on otettava huomioon. Tietosisällön analysointiin on siis sisällytettävä kunkin osa-alueen riittävä perustuntemus, ettei tultaisi vääriin johtopäätöksiin analyysien ja raporttien tulkinnassa. Olisi myös hyvä, mikäli tätä hiljaista tietoa voitaisiin siirtää tietovarastoon mahdollisimman paljon metatietoja tallentamalla. Raporttien ja analyysien laatijoiden on helpompaa hallita metatietoihin tallennettua hiljaista tietoa.

Kuopion yliopistossa kirjataan tuotot ja kulut omiin liikekirjanpidontiliryhmiinsä. Yhden rahoittajan (Suomen Akatemia) kohdalla on kirjaussääntö, että rahoittajan maksamat tuotot on kirjattava menon oikaisuina liikekirjanpidontileille, jotka kuuluvat kulujen kirjauksiin. Kaikki muut tuotot kirjataan tuottotileille. Tästä johtuen raportoinnissa on huomioitava tämä erityistapaus, ettei tältä rahoittajalta saatu rahoitus jää huomioimatta raportoinnissa ja analysoinnissa. Tämän kaltainen hiljainen tieto tulisi kuvata riittävällä tarkkuudella metatietoihin, että tietojen raportointi ja niiden perusteella tehtävät päätökset olisivat oikeita.

Edellä esitettyjen esimerkkien pohjalta voidaan myös todeta, että raporttien ja analyysien tekemiseen on käytettävä riittävää asiantuntemusta. Metatietoihin tallennetuilla tiedon kuvauksilla voidaan helpottaa oikean tiedon löytymistä ja paikkansapitävyyttä,

mutta tietojen luotettavuuden kannalta on tietojen analysoijan tunnettava myös tiedon syntyyn vaikuttaneet toimintaprosessit ja tiedon tallennustavat.

2.2.6 Aineiston kattavuus

Tietovaraston tietoja voidaan käyttää myös tieteelliseen tutkimukseen, jolloin aineistolle asetettavat vaatimukset paitsi tiedon oikeellisuuden myös aineiston kattavuuden suhteen ovat oleellisia. Kvantitatiivisen tutkimuksen pohjalla on oltava riittävästi tietoja tutkittavasta kohteesta. Kvalitatiivisessa tutkimuksessa aineiston kattavuudelle asetetaan kriteerit aineiston edustavuudesta ja yleistettävyydestä. [EsS98, s. 60-61]. Aineiston kattavuuteen on kiinnitettävä myös huomiota tehtäessä yleistyksiä tietovaraston tiedoista.

Esimerkiksi tallennettaessa yliopiston laitoksien julkaisutietoja JULKI-tietokantaan voivat puutteelliset tiedot aiheuttaa virheraportointia. Tästä johtuen julkaisujen määrä on raporteilla todellisuutta pienempi. Seurauksena pahimmassa tapauksessa on se, että laitokselle jaettava tuloksellisuuteen perustuva rahoitus pienenee.

2.2.7 Tiedon validiteetti

Analysointeja tehdessä on tarkoin harkittava, mitä tietoja halutaan mitata ja ovatko operatiivisista järjestelmistä saatavat tiedot valideja.

> ”Validiteetilla tarkoitetaan mittarin kykyä mitata juuri sitä, mitä on tarkoituskin mitata. Kun teoreettinen ja operationaalinen määritelmä ovat yhtäpitävät, on validiteetti täydellinen.” [Uus01, s. 84].

Mikäli halutaan analysoida esimerkiksi työntekijöiden ikäjakauman ja koulutukseen käytetyn ajan korrelaatiota, on tietovarastossa oltava tieto molemmista. Ikä voi olla syötettynä suoraan tietojärjestelmään tai se voidaan laskea suoraan syntymäajasta tai henkilöturvatunnuksen alkuosasta. Mikäli ikä on syötetty tietojärjestelmään sellaisenaan, on otettava huomioon analysoinnin aika eli onko ikä tällä hetkellä sama kuin tietovarastointiin tallentamisen aikaan. Koulutuksen käytetty aika on oltava myös tallennettuna tietovarastoon tai se on voitava johtaa jostakin tietovarastossa olevasta tiedosta.

Kuopion yliopistossa tietovarastoon syötetään henkilötunnus, jonka perusteella ikä voidaan laskea halutun ajankohdan mukaisesti. Koulutukseen käytetty aika tallennetaan työajankohdennusjärjestelmään, josta voidaan laskea toimintokoodille tallennettujen aikojen summa. Analysointia tehtäessä on vielä mietittävä, mitkä toiminnot halutaan laskea koulutukseen käytettyyn aikaan mukaan. Onko esimerkiksi toimintokoodille ”Jatko-opiskelu” kirjatut tunnit otettava mukaan vai halutaanko analysoida vain nykyisten työtehtävien hoitamiseen tarvittavaa koulutusta.

2.2.8 Tiedon esityksen tarkkuus

Mittauksen tarkkuudella voidaan kuvata mittaustuloksen hyvyyttä ja se ilmaistaan yleensä virherajojen avulla. Tarkkuus voidaan luokitella *sisäiseen (precision)* ja *ulkoiseen (accuracy)* tarkkuuteen. [Wik08]. *Tiedon esitystarkkuus (data precision)* on oleellista tietojen analysointeihin ja raporteille tehtävien laskennallisten kenttien osalta. Mikäli tiedon esitystarkkuus ei ole tarpeeksi tarkalla tasolla, myös siitä tehtävät laskennat tai mittaustuloksen perusteella tehdyt tietojen muunnokset tai johdetut tiedot antavat vääriä tuloksia.

Myös mittauksen tarkkuudella on huomattava merkitys analysointeja tehtäessä. Uusituvan mukaan esimerkiksi iän määrittelyn tarkkuustasoksi riittävä moniin tutkimustuloksiin on viiden tai kymmenen vuoden tarkkuus [Uus01]. Mikäli edellisessä luvussa esitetty iän ja koulutukseen käytetyn ajan korrelaatio halutaan analysoitavaksi, on mietittävä analysoidaanko vuoden vai pidemmän aikajakson tuloksia. Mitä tarkemmalle tasolle analysointi halutaan, sitä tarkemmalla tasolla myös tietovaraston tietojen on oltava tallennettuna. Analysointeja tehdessä tiedon tarkkuustaso voidaan tarkistaa metatiedoista, mikäli se on metatietoihin tallennettu.

2.3 Esimerkki tietovarastoprosessista Kuopion yliopistossa

Yliopiston johtamista ja laitosjohtamista palvelemaan on hankittu Kuopion yliopistoon HAT-ohjelmisto [Bus08], jota on käytetty Kuopion yliopiston tietotekniikkakeskuksessa kehitettyjen KASSi-sovellusten rakentamiseen. Operatiivisista tietojärjestelmistä

kootaan historia- ja summatietoa sisältävät tiedot tietovarastoon, josta edelleen toiminnan suunnittelu- ja seurantajärjestelmiin. Ensimmäisenä KASSi-sovelluksena on tietovarastosta rakennettu taloushallinnon seurantaa ja analysointia varten tarvittava raportointi- ja analysointityökalu (RASSi).

Kuopion yliopiston rahatalouden seurantajärjestelmä, RASSi, joka on tietovarastopohjainen, mahdollistaa tietojen analyyttisen tarkastelun siten, että tarkastelunäkökulma on helposti vaihdettavissa. Tietovaraston tietojen tarkastelua varten on Kuopion yliopiston tietotekniikkakeskuksessa rakennettu OLAP-välineisiin kuuluvalla HAT-ohjelmistolla KASSi-työkaluja helpottamaan tietojen tarkastelua. Aikasarjat ja graafiset kaaviot helpottavat yliopiston johtoa tekemään strategisia ratkaisuja.

Rahoitukseen ja rahojen käyttöön liittyvät tiedot siirretään Raindance-taloudenohjausjärjestelmästä, joka on ollut yliopistolla käytössä vuodesta 1994. Raindance-järjestelmän tietokantana toimii maxx-tietokanta, joka on kuitenkin siten suljettu, ettei yliopiston henkilökunnalla ole mahdollisuutta päästä itse tietokantaan muutoin kuin Raindance-järjestelmän parametritietojen ylläpidon kautta. Itse ohjelmisto koostuu ns. yrityksistä, joihin voidaan parametroida haluttujen tietojen perusteella talouden hallintaan liittyviä käsitteitä. Esimerkiksi projektirahoitus on toteutettu siten, että järjestelmään on syötetty käsitelaji PROJ, joka on kytketty muihin tähän arvoon kiinteästi yhteydessä oleviin käsitelajeihin ns. käsiterakenteen avulla. Täten esimerkiksi projektin vastuuhenkilö, käytettävä talousarviotili ja rahoittajakoodi löytyvät käsiterakennekytkennästä. Kytkentää voisi verrata verkkorakenteisessa tietokantamallissa osoitin viittaukseen tai relaatiotietokantamallissa taulujen osoittimiin.

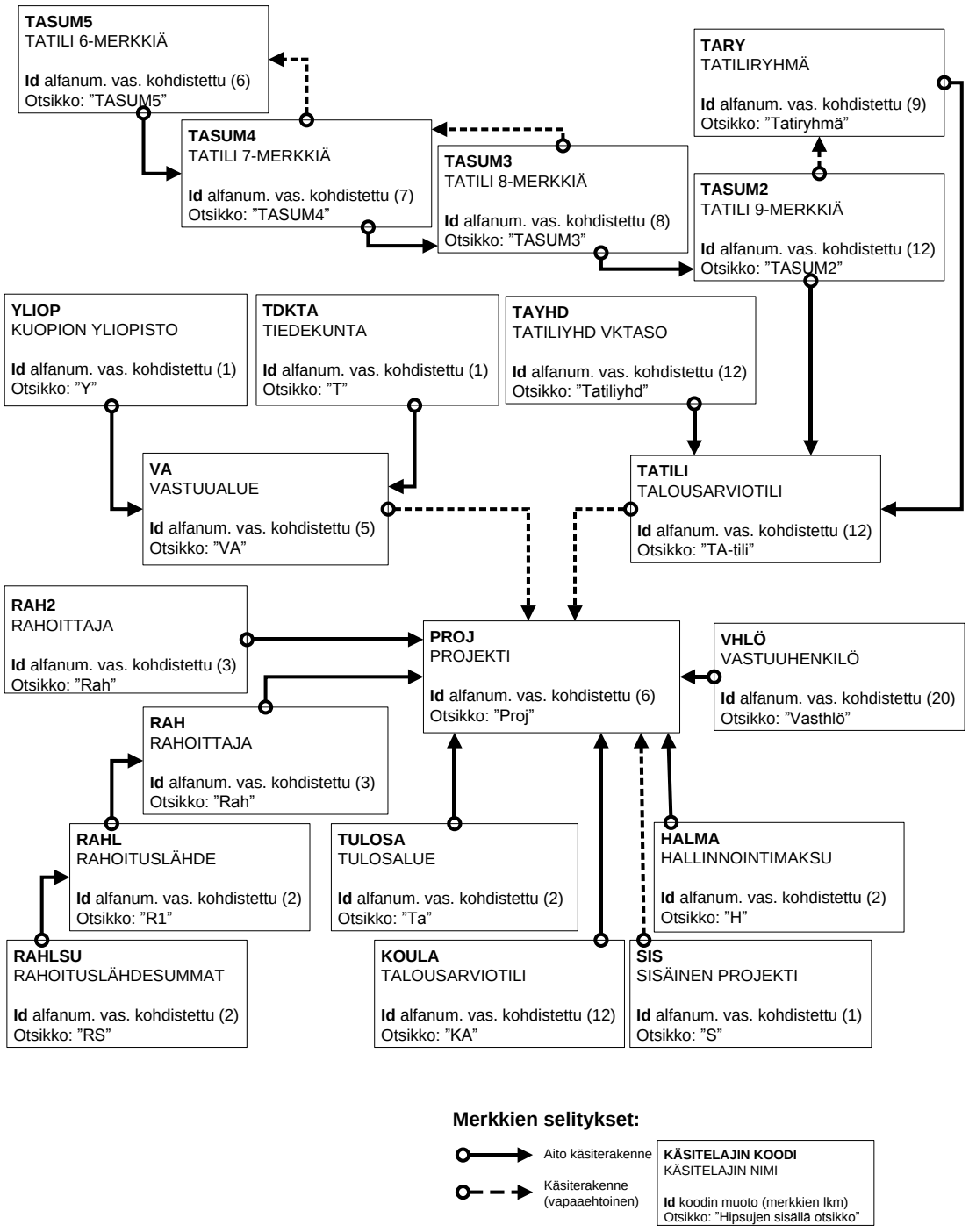
Tietovaraston käyttöönottovaiheessa on määriteltävä ne käyttäjät ja käyttötapaukset, joiden toimintaa tietovarasto palvelee. Tietojen käyttöä säätelee myös se, onko käyttäjällä oikeus tietoihin ja missä roolissa hän on käyttäessään tietoa. Tietovaraston käytön yhteydessä on siis myös mietittävä käyttövaltuuksien ja erilaisten roolien hallintaa. Myös tulevaisuuden tarpeet on huomioitava lähdettäessä miettimään tietovaraston jatkojalostamista. Ennen tietovarastoon siirtämistä on oleellista miettiä, mitä tietoja taloudenhallintajärjestelmästä on oleellista siirtää tietovarastoon myöhempää tarkastelua var-

ten. Analysointi- ja seurantatarpeet on siis kartoitettava ja koodistot mietittävä. Myös koodistojen yhtenäistäminen on otettava huomioon, kun halutaan ottaa mukaan tietoja myös muista operatiivisista järjestelmistä.

Joensuun ja Kuopion yliopistojen yhdistyminen vuoden 2010 alusta Itä-Suomen yliopistoksi asettavat tietovarastolle sekä tietojen analysoimiselle ja raportoinnille uudet haasteet. Tietovaraston rakentamisvaiheessa on selvitettävä myös ne ongelmakohdat, kuten koodistojen yhteensovittaminen, ajantasaisuus, tietojen siivous ja tiedon saatavuuden nopeus, joilla on vaikutusta tietovaraston käyttöön. Niin ikään tietovaraston tuottaman tiedon oikeellisuuden ja luotettavuuden kriteerit ovat korkealla, unohtamatta tiedon saannin reaaliaikaisuutta ja helppoutta.

Tietovarastoksi Kuopion yliopistolla hankkeen alkuvaiheessa valittiin MySql sen yleiskäyttöisyyden ja kokonaisedullisuuden vuoksi. Myöhemmin tietovarastoinnin edetessä laajempaan käyttöön harkittiin tietovarastoinnin tietokantaratkaisua uudelleen. Välineeksi valittiin Oracle 10g ja tietovaraston hallinnointiin Oracle Warehouse Builder. Tietokantaratkaisuksi valittiin tähtimalli, koska se mahdollistaa tehokkaat kyselyt, kun tietokantataulujen välisiä liitoksia ei ole paljon.

Seuraavassa kuvassa (Kuva 3) on esitetty Raindance-järjestelmässä käytössä olevat taloushallinnon projektiin liittyvät koodistot (*käsitelajit*) ja niiden väliset yhteydet (*käsiterakenteet*).



Kuva 3. Raindance-järjestelmän PROJ-käsitelaji ja sen kytkennät

Taloushallinnon järjestelmässä on paljon koodistoja ja yhteen `käsitelajiin` voi liittyä monia tietoja ja niiden ryhmittelytekijöitä. Mikäli koodistot eroavat toisten järjestelmien vastaavista koodeista tai niillä on eri merkitys eri järjestelmissä, on tietovaraston huolehdittava koodien lähdejärjestelmistä tulevien tietojen käsittely siten, etteivät

koodistot mene sekaisin. Myös koodistojen hierarkioihin liittyvät tiedot on tallennettava tietovarastoon.

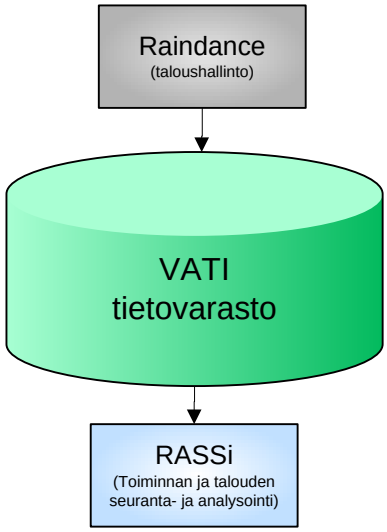
Metatietoihin voidaan tallentaa tiedot sekä lähdejärjestelmästä että tiedot koodistoihin liittyvistä rakenteista. Esimerkiksi taloushallinnon projektiin kytketty vastuualuetieto (VA) voi siirtyä tietovarastoon toisista järjestelmistä eri tavalla kuin Raindancesta. Tällöin tietovarastosta tuotettavassa analyysissä tai raportissa on otettava kantaa siihen, kumman järjestelmän vastuualuetietoa käytetään.

Metatietoihin voidaan tallentaa myös tiedon omistaja ja tallentaja. Kun tiedolle on määritelty omistaja, on tietojen raportointi ja analysointi helpompaa. Esimerkiksi, taloushallinnon kirjauskäsitetiedoille on itsestään selvää määritellä tiedon omistajaksi taloushallinto. Edelläkin kuvattu vastuualuetiedon omistaja on yliopiston hallitus, joka päättää yliopiston organisaatiosta ja siellä käytössä olevista hallinnon rakenteista. Vastuualueelle annettavan kooditiedon omistaja onkin vaikeampi määritellä.

Ensimmäisen kerran, kun uusi organisaatiotieto, esimerkiksi vastuualue, tallennetaan johonkin operatiiviseen järjestelmään, ei organisaatiotiedolle ole määritelty koodiarvoa, vaan tieto tallennetaan yleensä manuaalisesti järjestelmiin. Tästä johtuen samalla organisaatiotiedolla voi olla eri koodiarvo eri operatiivisissa järjestelmissä, kuten opintohallinnon ja taloushallinnon järjestelmissä. Tästä seuraa pahimmillaan se, etteivät tiedot ole yhteismitallisia, eikä niiden perusteella voida tehdä luotettavaa raportointia, vaikka tiedot olisikin viety tietovarastoon.

Kuopion yliopiston tietojärjestelmistä ja niiden liittymistä on tehty liitteenä oleva järjestelmäkuvaus (liite 1). Järjestelmäkuvauksesta voidaan nähdä, mistä järjestelmistä tietojen siirtoa on jo toteutettu. Tietovaraston toteuttaminen on lähdetty toteuttamaan yliopiston toiminnan ydinjärjestelmistä. Koska taloustiedot ovat täsmällisiä ja yliopiston johdolla sekä organisaatiotasojen johtajilla on talousseuranta tärkeää, on ollut luontevaa aloittaa siitä. Lisäksi tietovarastointisiirtoja on jo toteutettu henkilöstöhallinnon ydinjärjestelmistä sekä muutamista muista järjestelmistä. Tietovarastoon tallennettua tietoa on hyväksikäytetty tietojen välittämisessä ja talouden seuranta- ja suunnittelujärjestelmän

(RASSi) toteutuksessa. Seuraavassa kuvassa (Kuva 4) on erotettu järjestelmäkartasta tietojen siirto VATI-tietovaraston kautta RASSi-järjestelmään.



Kuva 4. Taloustietojen siirto seuranta- ja analysointijärjestelmään

RASSi on yliopiston laitosjohdon talouden suunnittelua ja seurantaa varten tietotekniikkakeskuksessa rakennettu HAT-ohjelmaan pohjautuva järjestelmä. Kuten on nähtävissä, taloustiedot siirretään taloushallinnon `Raindance`-järjestelmästä ensin tietovarastoon, josta ne siirtyvät `RASSi`in. Tietojen siirto on automatisoitu molemmissa päissä siten, että Raindance-järjestelmä tuottaa joka päivä tekstitiedoston kirjanpidon tapahtumista, jonka `VATI-tietovarasto` lukee eräajona talteen ja tuottaa edelleen RASSi-järjestelmää varten tiedot täydennettyinä verkkolevyalueelle tekstitiedostoiksi. Tietojen täydennyksen yhteydessä merkitään puuttuvat kooditiedot tietovarastoon varatulla puuttuvan tiedon koodilla. Tietovarasto tuottaa myös RASSi-järjestelmän tarvitsemat lajittelutiedot dimensiorakenteista. Dimensiotiedot on lajiteltu muun muassa tiedekunnittain vastuualueittain projekteittain, rahoittajittain projekteittain ja tulosalueittain projekteittain. RASSi-järjestelmään on automatisoitu tietojen sisään luku, jossa dimensiorakenteet luetaan ensin ja tapahtumatiedot tämän jälkeen. RASSissa on mahdollista tutkia tietoja kaikilla siihen määritellyillä dimensiotasoilla ja porautuminen on mahdollista tiedon tarkimmalle tositerivitasolle saakka. Näkökulmaa vaihtamalla voidaan tietojen lajittelua muuttaa helposti toisen dimensiorakenteen mukaisesti. Myös rivitietojen muuttaminen graafisiksi kaavioiksi on mahdollista.

Tietojen analysointia tukemaan on järjestetty Kuopion yliopistossa RASSi-tiimi, joka tuottaa tietojen analysoinnissa käytettävät raportit ja analyysit. Käyttäjillä on pääsääntöisesti käytössään vain katselulisenssillä toimiva RASSi-aineisto, joka on muodostettu kullekin vastuualueelle vain heidän käyttöoikeuksiinsa pohjautuvista tiedoista. Analyysien tuottamiseen käytetään ammattitaitoa, joka takaa sen, että taloushallinnon tiedot on koottu ja ryhmitelty oikein. Analyysien tekemiseen mahdollistavat lisenssit HAT-ohjelmiston ja RASSi-aineistojen käyttöön on käytössään vain taloushallinnon käyttäjillä, joilla on riittävästi tietoa taloustietojen tallentamisprosesseista ja -käytänteistä. Tällä pyritään varmistamaan se, ettei analyysien perusteella synny vääriä johtopäätöksiä, esimerkiksi organisaatioyksiköiden taloustilanteesta.

Mikäli analyysien ja raporttien käyttäjillä ei ole riittävästi tuntemusta tiedon sisältöön liittyvistä ryhmittelytekijöistä ja toimintatavoista tiedon tallentamisen yhteydessä, on tietovarastosta saatava tieto arvotonta. Koodistoihin liittyvien ryhmittelyjen rakenne voidaan kuvata metatiedoissa, jolloin analysoijat ja raportoijat voivat käyttää hyväkseen niitä. Kun tietovaraston koko kasvaa suureksi, on myös tietojen ja niihin sisältyvien master data ja metadata -tietojen hallinnointiin kiinnitettävä entistä enemmän huomiota.

3 METADATA

Metatiedon tuottaminen ja hallinta parantaa tietovarastosta saatavan tiedon laatua ja määrää [SVV99b]. Metadata auttaa tietovaraston käyttäjiä ja tiedon analysoijia valitsemaan oikean tiedon oikeaan paikkaan. Metadataa voidaan käyttää tietovaraston tietojen sisällysluettelona. Metadatalla onkin tietovarastoinnissa tärkeä rooli, koska sen avulla tietovaraston tietoja pystytään hakemaan ja käyttämään tehokkaammin ja luotettavammin. [Inm05]

Metadata tallennetaan yleensä omaan tietokantaansa tai se voi olla myös erillisenä dokumenttina vaikka tekstitiedostona. Metadatan käytettävyyden kannalta on kuitenkin luontevaa tallentaa metatiedot tietovaraston kanssa samaan paikkaan, omaksi tietokannakseen. Tällöin voidaan tietovarastoon latauksen yhteydessä täyttää metadatan tallennusta automaattisesti sellaisissa kuvauksissa, jotka voidaan helposti automatisoida.

Metadataa käytetään tietovaraston rakentamisessa, ylläpidossa ja käytössä. Metadata voidaan jakaa kahteen luokkaan sen sisältämän tiedon perusteella:

- *Tekninen metatieto (technical metadata),* joka sisältää tiedon teknisiä ominaisuuksia ja niiden edellyttämiä laitteisto- ja ohjelmistovaatimuksia kuvaavia piirteitä. Teknistä tietoa ovat myös tiedot operatiivisista tietolähteistä, tiedon latausvaiheista, summataulujen tekovaiheista, tiedon tietotyypistä ja pituudesta, eheystarkastuksista, tietojen muunnoksista, tietovaraston käyttöoikeuksista, varmuuskopiointihistoriasta ja niin edelleen. Tyypillisiä teknisiä ominaisuuksia ovat esimerkiksi tiedon *esitysmuoto (format), koko (size)* ja *tallennuspaikka (location).*

- *Liiketoiminta-metatieto (business metadata)* kertoo loppukäyttäjille liiketoiminnallisen näkemyksen tietovaraston sisältämään tietoon. Tietojen määritelmät kuvataan liiketoiminnan termein ja tietolähde, tiedon päivityshetki, tiedon mahdolliset laskentakaavat, käsitemalli ja moniulotteiset tietomallit (kuutiot) tukevat raporttien ja analyysien tekijää tuottamaan haluttuja tietoja tietovarastosta.

Tässä luvussa määritellään metadata ja pureudutaan tarkemmin käsitteen luokitteluun sekä esitetään metadatalle yllä olevan lisäksi muitakin luokitteluvaihtoehtoja. Tämän jälkeen käsitellään järjestelmien välistä tiedon siirtoon kuuluvaa yhteentoimivuutta. Lisäksi esitellään yleisesti käytössä olevia metadatastandardeja sekä pohditaan niiden heikkouksia ja vahvuuksia. Lopuksi käsitellään metatietomallia ja sitä kuinka metadata-standardit voivat olla metatietomallin pohjana. Määritelmiä ja käsitteitä verrataan Kuopion yliopistossa käytössä olevaan VATI-tietovarastoon ja sen tietokantakuvaukseen.

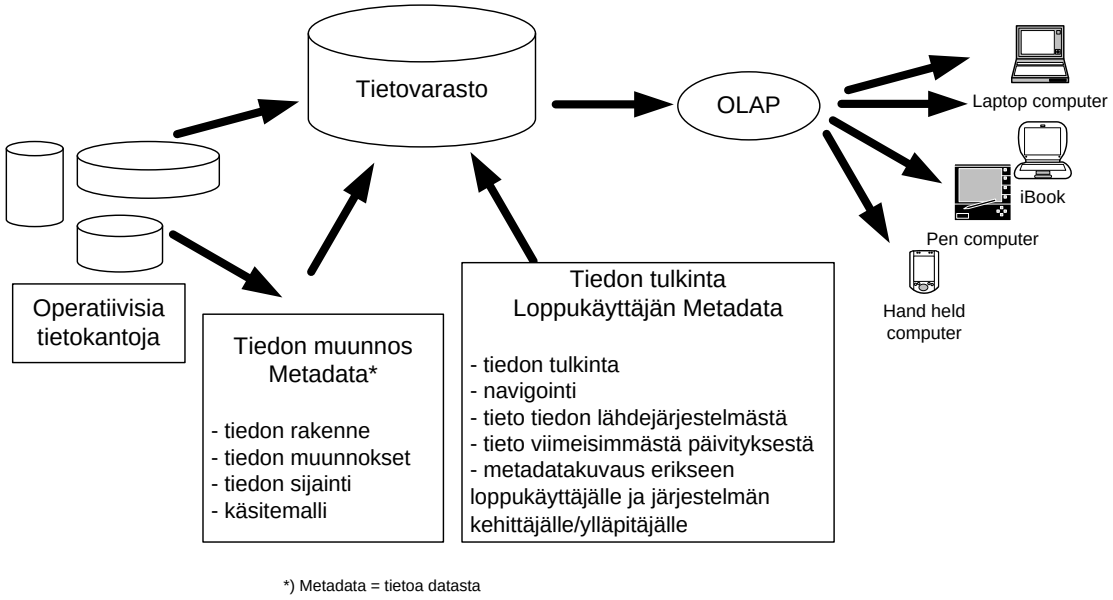
3.1 Metadata käsitteenä

Tietovarastoa rakennettaessa on ensin kyettävä arvioimaan ja valitsemaan olennainen tieto, joka halutaan siirtää ja/tai jalostaa tietovarastoon [Nie02 s. 16]. Tiedon siirtojen yhteydessä tuotetaan myös kuvailua tiedolle eli metadataa. *Metadata* on tietoa tiedosta (mm. [Gar98, s. 59], [Hac99, s. 165], [JaS98, s. 31]). Tietoja, jotka kuvailevat metadataa itseään, kutsutaan *meta-metadataksi*. Metadatasta käytetään usein myös termiä metatieto, mutta Yleinen suomalainen asiasanasto (YSA) ilmoittaa, ettei asiasanaa metatieto käytetä, vaan käytettävä asiasana on metadata [YSA99].

Metadatan hyvin kuvailevana rinnakkaisterminä YSA esittää *sisällönhallintaa (Content Management)*, mikä tarkoittaa toimintaa, jossa pyritään hallitsemaan digitaalista informaatiosisältöä mahdollisimman tarkoituksenmukaisesti [Wik08]. Digitaalinen informaatiosisältö muodostuu siirretyistä tiedoista ja sitä kuvailevasta metadatasta. Sisällönhallintaa varten on olemassa myös *sisällönhallintajärjestelmiä (content management system)*, joiden käyttö parantaa organisaation sisällä jaettavan informaation tarkkuutta ja laatua. Ne tukevat myös tehokkaita tiedon haku-, navigointi- ja suodatusmenetelmiä, joiden avulla käyttäjän on helppoa löytää oikea informaatio oikeaan aikaan. Sisällönhallintajärjestelmä yksinkertaistaa sekä tiedon tallennusta ja metadatan käsittelyä että auttaa informaation käsittelyssä. [PeS07]

Metadataan tallennetaan tyypillisesti seuraavat tiedot: tiedon nimi, sanallinen määritelmä, tiedon omistaja, tietotyyppi, pituus, lähdejärjestelmä, päivitysajankohta, laskennalliset kaavat, joihin tieto voi perustua ja tiedon käyttöoikeudet [Hov97]. Metadataan on

myös määriteltävä tieto siitä, missä muodossa tieto on tietovarastossa [KHL01]. Metadataan määritellään tiedon tietotyyppi sekä tietoa tuottavassa tietojärjestelmässä että tietovarastossa. Lisäksi metadataan on tallennettuna, miten tieto johdetaan (esim. euromuunnoskerroin), tiedon vaihteluväli eli suurin ja pienin sallittu arvo sekä tieto siitä voidaanko yksilöivä data korvata toisella vai ei. Myös tiedon tarkkuudelle asetettu arvoalue on metadataa eli millä tarkkuudella tieto on esitettävä.



Kuva 5. Metadata

Kuva 5 esittää kuinka, tietoja siirretään sekä suoraan että tietojen muokkauksen kautta tietovaraston tietokantatauluihin. Tietovaraston kannalta metatiedoissa on tietoon sisältyvät tekniset sekä siirtoon liittyvät tiedot, tiedon tulkintaan käytetään metatiedoista tiedon kuvausta, tietoja tiedon ajankohdasta ja tiedolle tehtyjä sanallisia kuvauksia. Tiedon hyväksikäyttäjälle metadata antaa tiedolle merkityksen ja hän voi käyttää metadataa tietojen tulkintaan ja helpottamaan tietojen keräämistä oikein.

Kuopion yliopistossa tehdään tietovarastoon siirron yhteydessä tietoon liittyvää metatietojen tallentamista. Liitteessä 2 on taloustietojen siirtoon käytettävä tietovarastokuvaus. Taloushallinnon tietojen lisäksi kuhunkin dimensiotauluun siirretään metatietoja tiedon syntymisestä: tiedon alkamispäivämäärä ja loppumispäivämäärä sekä nykyisen tiedon merkitsemiseksi kenttä YP_NYKYINEN. Metatiedoiksi tallennetaan siis lähinnä tiedon

sisällön voimassaolon ajankohta. Luvussa 2.2.2 kuvattiin näiden kenttien käyttöä tarkemmin. Muita metatietoja ei Kuopion yliopiston tietovarastossa ole otettu toistaiseksi käyttöön.

3.2 Metadatan luokittelu

Metadatassa määritellään varsinaisen tiedon käyttö, hallinta sekä käyttäytyminen. Metadatan tyyppejä voidaan luokitella yleisellä tasolla seuraavasti [CGG98].

- *Hallinnollinen metadata.* Resurssien hallitsemisessa ja ylläpidossa käytettävä metadata (hankintatiedot, omistajatiedot, sijaintitiedot, versionhallinta).
- *Kuvauksellinen metadata.* Resurssin sisällön kuvaamiseen käytettävä metadata, käyttäjien manuaalisesti täytettävä (luettelotiedot, hakemistot, tiedon kuvaukset).
- *Tiedon säilytyksen metadata.* Resurssin säilytyksen hallinnassa käytettävä metadata (fyysinen tallennuspaikka, säilytyksen, päivityksen ja siirron ohjeet).
- *Tekninen metadata.* Järjestelmän toimintoihin tai metadatan käyttöön liittyvä metadata (dokumentaatio, tiedostomuoto/pakkaus, käyttöoikeudet/salaus).
- *Tiedon käytön metadata.* Resurssin käytön hallinnassa käytettävä metadata (esitystiedot, käyttöloki, edelleenkäyttötiedot).

Metadataluokitteluja voidaan tehdä myös tiedon luonteen perusteella seuraavaksi esitettävällä tavalla [SBC03]:

- *Fyysinen metadata (physical metadata)* sisältää kuvauksen tiedon ominaisuuksista, jotka liittyvät tiedon rakenteeseen ja tallennusmuotoon, sekä jäljennöksen tiedon metadatan sijainnista. Edellisen luokituksen mukaisesti nämä tiedot sisältävät tiedon säilytyksen ja teknisen metadatan.
- *Ympäristöriippumaton metadata (domain-independent metadata)* kuvailee yleisiä tietoon liittyviä elementtejä, jotka eivät ole riippuvaisia järjestelmästä tai siitä missä tieto on syntynyt. Näitä yleisiä kategorioita voivat olla esimerkiksi tiedot julkaisijasta, tekijästä tai tiedon muokkaajasta sekä tietojen yhdistämisestä ja niihin käytettävissä olevista näkymistä.

- *Ympäristöriippuvainen metadata (domain-specific metadata)* sisältää tietoon liittyviä elementtejä, jotka kertovat tietoon liittyvät järjestelmäriippuvaiset tai tiedon ominaisuuteen liittyvät kuvailutiedot. Tiedon esittämiseen sovitut termit tai käytetyt mittaluvut voidaan sopia yhteneväisiksi.
- *Näennäisorganisaation metadata (virtual organization metadata)* sisältää kuvailun tiedoista, joiden määrittelyyn on yhteisesti sovittu tietokentät niiden toimijoiden kesken, jotka kuuluvat samaan tiedeyhteisöön tai yhteistyöelimeen. Esimerkiksi erilaiset tutkimusryhmät, joiden tietolähteinä käytetään eri järjestelmistä saatavia tietoja, voivat määritellä tietoelementit kuten huomautus-kentän sisältöön tarkoitetut tiedot.
- *Käyttäjäkohtainen metadata (user metadata)* sisältää ne tietoon kuuluvat metatiedot, jotka yksittäiset käyttäjät tallentavat tiedolle. Tähän kategoriaan kuuluvat edellisen ryhmittelyn mukaiset hallinnolliset ja kuvaukselliset metadatat.

Professori Plattner käyttää metadatan luokitteluun *liiketoimintatietämyksen (business intelligence)* näkökulmaa ja hän kuvaa metatiedon välttämättömäksi komponentiksi. Hänen mukaansa metadatan voidaan jakaa *tekniseen metadataan (technical metadata), liiketoiminnan metadataan (business metadata)* ja *toiminnalliseen metadataan (operational metadata).* Tämän luokittelun mukaisesti tekninen metadata kuvailee tiedon rakenteen ja sisällön, esimerkiksi tiedon tyyppi ja kentän pituus. Toiminnallinen metadata pitää sisällään tietovarastoon latauksen sisältämät operaatiotiedot, kuten esimerkiksi tietueiden määrän jossakin tietyssä tietovaraston taulussa. Liiketoiminnallisen metadatan kuvauksiin tallennetaan tiedon semanttiset tai liiketoimintaan liittyvät tiedot, joilla kuvataan tiedon käyttäytymistä järjestelmässä. [Pla07].

Metadata mahdollistaa tiedon tehokkaan analysoinnin ja käytön sekä uusien ratkaisujen aikaisempaa helpomman suunnittelun ja toteutuksen [Jok02]. Metadata antaa tietoa tietovaraston sisällöstä sekä tietovaraston suunnittelijalle, ylläpitäjälle että käyttäjälle. Metatiedon merkitys on avainasemassa suunniteltaessa tietovarastoa [SVV99a, SVV99b]. Yhtenäisten sääntöjen ja metatietomallin noudattaminen on ehdoton edellytys tietovarastoja rakennettaessa. Metadata antaa tietovarastoinnin tuottamalle tiedolle luotetta-

vuutta ja parantaa tiedon saavutettavuutta. Metadata sisältää tiedot paitsi itse tietovaraston sisältämästä tiedosta ja sen tietotyyppiin kohdistuvista muutoksista, myös tiedot lataustyönkulusta ja sen rakenteesta. Lataustyönrakenne sisältää tiedon raakadatan uuttamisesta lähteistä, tiedon muokkauksen toimenpiteistä ja tiedon lataamisesta. Metadata kertoo siis tiedon koko elinkaaren tietovarastossa [Jok02].

Metadataa voivat käyttää sekä tietovaraston ylläpitäjät tietovaraston hallintaan että loppukäyttäjät suunnitellessaan tietovarastosta saatavia raportteja ja näkökulmia. Metadata sisältää myös teknistä tietoa tiedon muodostumisesta kaikkine välivaiheineen aina tietovarastoon tallentamiseen saakka. Tietojen kuvaaminen metatietojärjestelmään on edellytys tietovaraston käytön leviämiselle yhä uusille käyttäjäryhmille [HYK01].

3.3 Master data

Master data on tietoa, jota jaetaan järjestelmien kesken, kuten esimerkiksi hierarkkiset listaukset asiakkaista, toimittajista, laskuista ja organisaatiotasoista, ja joita käytetään luokittelemaan ja määrittelemään tapahtumapohjaisia tietoja [MoV05]. Tietojen luokittelemisen pohjana käytetään organisaation raportointitarpeita. Esimerkiksi rahoittajat vaativat kulujen raportoinnin omien ryhmittelyjensä mukaisesti. Myös organisaation sisäisiä tietoja voi olla tarkoituksen mukaista ryhmitellä.

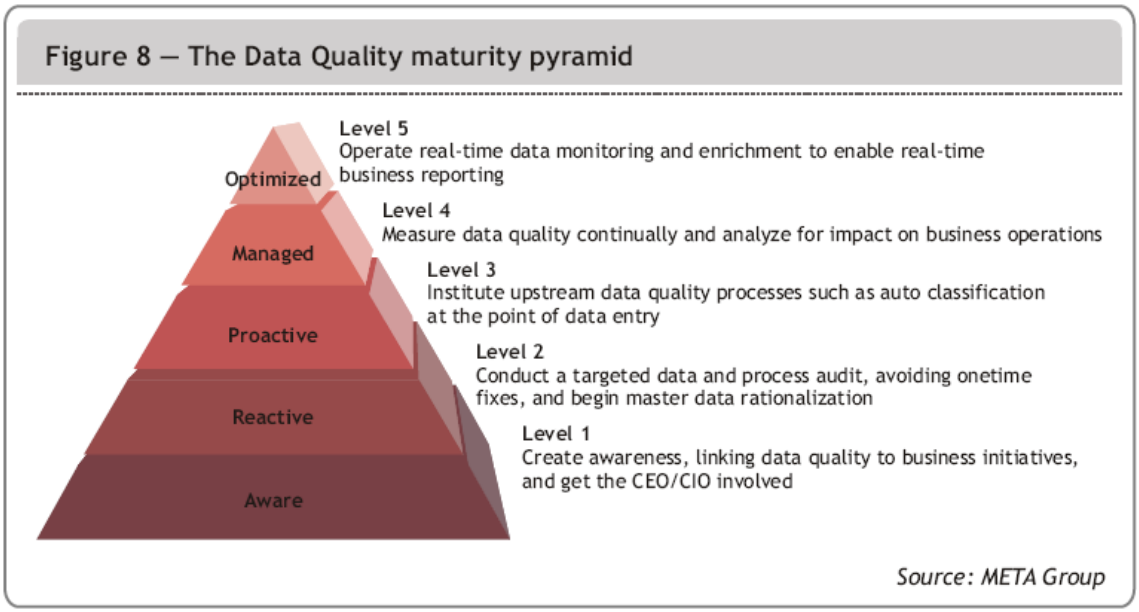
Master data ei tarkoita samaa kuin metadata. Master dataan sisällytetty informaatio palvelee paremminkin organisaation liiketoimintaa, kun taas metadata sisältää enemmän informaation teknisen tiedon tallennuksen näkökulmaa [MoV05]. Metadata voi kyllä sisältää master dataa, toisin sanoen tietojen ja koodistojen luokittelu- ja hierarkiatiedot voidaan tallentaa metadatan avulla.

Mikäli master data on huonosti hoidettua, eli se sisältää virheellistä, epätarkkaa, monistettua tai vanhentunutta tietoa, voi se tulla organisaatiolle kalliiksi ja luottamus tietojen paikkansa pitävyyteen, eli tietovarastoon, katoaa. Huono master data voi tuhota hyvin suunnitellun liiketoiminnan prosessit. Master data vaatii siis jatkuvaa ylläpitoa. [MET04]

META Group määrittelee master datan jaon kahteen kategoriaan [MET04]:

> *Ensisijaiset master data tiedot (Primary master data records):* Nämä tietueet ovat kuten alkulukuja, niitä ei voida muokata. Työntekijät, asiakkaat, toimittajat ja tuotteiden koodistot ovat tyypillisiä ensisijaisia master data tietueita.
>
> *Johdetut master data tiedot (Derived master data records):* Johdetut master data tiedot saadaan linkittämällä ensisijaisia master data tietoja yhteen. Linkittämällä asiakkaan tiedot tuotetietoihin, luodaan perusta hinnoittelutiedoille, jota käytetään myynnin järjestelmissä.

Mikäli asiakastiedot sisältävät virheellistä master dataa, toisin sanoen esimerkiksi asiakkaiden ryhmittely on virheellistä, tuottaa se hinnoittelun pohjaksi väärää tietoa. Tällä voi olla suuri merkitys liiketoiminnalle.



Kuva 6. Tiedon laadun kypsyyspyramidi [MET04]

Meta Group esittää `tiedon laadulle kypsyyspyramidin` (Kuva 6), jossa kuvataan sitä kuinka organisaation tulee kiinnittää huomiota tarkoituksenmukaisiin ohjelmistoihin, käytäntöihin, arkkitehtuuriin ja infrastruktuuriin voidakseen hallita ja käyttää tietoa paremmin [MET04].

Kypsyyspyramidin tasot ovat [MET04]:

- *Taso 1: Valveutuneisuus (Aware)*. Tällä tasolla olevat organisaatiot elävät tiedon kaaoksen keskellä. Heillä on tiedossaan, että tiedon laadulla on merkitystä myös liiketoiminnan suunnitteluun ja päätöksentekoon. Päätöksenteko perustuu satunnaiskyselyillä saatuihin raportteihin.
- *Taso 2: Reagointi (Reactive)*. Tällä tasolla päätökset ja toimeksiannot asetetaan usein kyseenalaiseksi, koska tiedon laadun merkitys on havaittu ja tästä johtuen organisaation johto tukeutuu päätöksenteossa vaistoonsa mieluummin kuin epäluotettaviin raportteihin.
- *Taso3: Ennakointi (Proactive)*. Tällä tasolla organisaatiossa huolehditaan tiedon laadusta ja tietovarastoon latauksen yhteydessä on huolehdittu tiedon puhdistamisesta. Organisaation johto luottaa tiedon laatuun ja käyttää tiedosta kerättyä raportointia strategiseen ja taktiseen päätöksentekoon. META Group arvioi, että tälle tasolle sijoittuu 15 - 20 % yrityksistä.
- *Taso 4: Hallinta (Managed)*. Tämän tason organisaatioissa tiedon laatu on asetettu tietohallinnon ensisijaiseksi tehtäväksi. Tiedon laatuun liittyvät toiminnot on rakennettu liiketoiminnan perusjärjestelmiin, mikä takaa menestyksellisen operatiivisen päätöksenteon. META Groupin mukaan vain 5 % yrityksistä yltää tälle tasolle.
- *Taso 5: Optimointi (Optimized)*. Tällä tasolla organisaatio ottaa kaiken irti tiedosta, ja silloin tiedon laadulle asetetut kriteerit ovat korkealla. Tiedon saannille on myös asetettu reaaliaikaisuuden vaatimus, tiedon laadusta tinkimättä. Organisaatio pystyy nopeaan ja tehokkaaseen päätöksentekoon. Myös toiminnan muutoksiin on mahdollisuus reagoida nopeasti.

3.4 Yhteentoimivuus

Tietotekniikan liitto ry:n sanastotoimikunnan toimittama ATK-sanakirja [ATK08] määrittelee käsitteen *yhteentoimivuus (interoperability)* seuraavasti:

> ”Järjestelmien kyky viestiä keskenään sellaisella tavalla tai siinä laajuudessa, että ne voivat rutiinimaisesti käyttää toistensa tuloksia omassa toiminnassaan. Tietotekninen yhteentoimivuus saavutetaan standardeja ja yhteisiä perusrakenteita käyttämällä.”

Yhteentoimivuudella tarkoitetaan kahden tai useamman järjestelmän mahdollisuutta vaihtaa ja käyttää vaihtamiaan tietoja [IEE90]. Tietojen vaihtaminen edellyttää järjestelmien syntaksin yhdenmukaisuutta (sisältäen tiedon rakenteen) ja tietojen käyttämisen kannalta myös kykyä ymmärtää tiedon semantiikka [Myk07]. Jotta järjestelmät olisivat yhteentoimivia, eli tietojärjestelmät ovat toistensa yhteydessä toimintakelpoisia, on järjestelmien käytettävä standardeja ja yhteisiä perusrakenteita [Tie06].

Yhteentoimivuutta voidaan vielä tarkentaa. *Yhteensopivuuden (compatibility)* vaatimus täyttyy, kun järjestelmä toimii ilman merkittäviä muutoksia sen yhteydessä toimivien toisten järjestelmien kanssa. Mikäli järjestelmä voidaan vaihtaa toiseksi ilman merkittäviä muutoksia, on kyseessä *vaihtokelpoinen (subsitutable, compatible)* järjestelmä. *Suoraan yhteensopiva (plug compatible)* järjestelmä on, kun se voidaan vaihtaa toiseen ilman muutoksia toisiin sen yhteydessä toimiviin järjestelmiin. [Tie06]

Kuten todettiin, yhteentoimivuuden takaavat parhaiten järjestelmät, jotka käyttävät standardeja. World-Wide Web on hyvä esimerkki siitä, kuinka käytetyt standardit (URL, HTTP ja HTML) luovat pohjan avoimille ja yhteentoimiville järjestelmille, jotka mahdollistavat loppukäyttäjälle valita palvelin- ja työasemaohjelmistot [Duv01].

Duval [Duv01] erottaa yhteentoimivuuden tasot, jotka on kuvattu seuraavassa taulukossa (Taulukko 3).

Taulukko 3. Yhteentoimivuuden tasot [Duv01]

1	Protocol	TCP/IP, http
2	Data binding	HTML, XML, RDF
3	Metadata scheme	LOM, Dublin Core
4	Semantic	Ontologies, classifications, vocabularies, taxonomies

Yhteentoimivuuden *protokolla (protocol)* -taso käsittää kaikkein teknisimpiä standardeja, kuten *TCP/IP (Transmission Control Protocol / Internet Protocol)* ja *HTTP (Hypertext Transfer Protocol).* Esimerkiksi HTTP-standardia käytetään sanoman välitykseen Web-selaimen ja palvelimen välillä eri toimittajien ja eri järjestelmien kesken. Toisella tasolla, *tiedon sitominen (data binding)*, käsittää standardit, jotka liittyvät tiedon sitominen tiedon esityksen formaattiin. Yleisiä standardeja tiedon sitomiseen ovat *HTML (Hypertext Markup Language), XML (eXtensible Markup Language)* ja *RDF (Resource Description Framework).* *Metadata skeema (metadata scheme)* taso määrittelee tietoelementtien metatietomallin laatimisessa käytetyt standardit, esimerkiksi LOM tai Dublin Core. *Semanttinen (semantic)* tasolla voidaan käsitellä tiedon ontologioihin, luokitteluihin, sanastoon ja taksonomiaan liittyvät tiedon käsittelyt. [Duv01].

Tässä tutkielmassa keskitytään tarkastelemaan metadata skeema tasoon liittyviä standardeja ja metadatastandardeista tarkemmin tutkitaan edellisen taulukon esittämät LOM- ja Dublin Core -standardit sekä OMG:n CWM-standardi.

3.5 Metadatastandardit

Metadatan merkitys tietovarastoinnin käytön ja hallittavuuden kannalta on tärkeä. Metadataa voidaan käyttää tietojärjestelmäkehityksen eri vaiheissa tukemaan vaatimusmäärittelyä, prosessien mallintamista, tietojen siirtoa, käyttäjähallintaa, tietovaraston ylläpitoa ja kehittämistä. Tiedon puhdistaminen on keskeinen osa-alue laadukkaan tiedon

tallentamisessa tietovarastoon. Tiedon likaisuus voi johtua esimerkiksi näppäilyvirheistä, mutta myös puutteellisista standardeista. Tiedon rikastamiseen eli tietojen yhdistelemiseen tarvitaan tietoja useista tiedoista.

Metadatan käyttö mahdollistaa sen, että tiedon eheys voidaan taata. Metatietoihin tallennettujen tietojen perusteella pystytään toteamaan tiedon *kirjausketju (audit trail)* eli tieto siitä, missä tieto on syntynyt, kuka sen on kirjannut ja koska tiedon kirjaaminen on tapahtunut. Metadata mahdollistaa myös tunnistamaan toisista tiedoista *johdetut (redundant), monenkertaiset (duplicate)* ja mahdollisesti *vanhentuneet (obsolete)* tiedot. Metadatan käyttö mahdollistaa myös luomaan linkkejä eri tietojärjestelmistä saatavien samojen tietojen, esimerkiksi vastuualuekoodien, välille. Metadatastandardeissa on kuvattu näiden tietojen hallintaa varten tarpeelliset kentät. Yhtenäistä metatietojen tallennusta varten on käytettävä yhtenäisiä käytäntöjä ja sopimuksia siitä, mitä metatietoja tallennetaan. Standardien käyttö auttaa löytämään ja hallitsemaan metatietoja.

Metadatastandardien käytön hyötyinä OMG (Object Management Group) esittää heterogeenisissä ympäristöissä hajautettujen oliopohjaisten järjestelmien uudelleen käytettävyyden, siirrettävyyden ja yhteentoimivuuden. Metatietomääritysten yhteneväisyys tekee mahdolliseksi kehittää heterogeenisiä sovellusympäristöjä kaikille yleisille laitealustoille ja operatiivisille järjestelmille. [Obj03].

Tässä tutkielmassa keskitytään metadata skeematason tutkimiseen ja seuraavissa kappaleissa esitetään yleisimmin käytetyistä metadatastandardeista Dublin Core ja LOM -standardit. Dublin Core on digitaalisten tallenteiden ja julkaisujen yleisimmin käytetty standardi ja LOM-standardia käytetään oppimateriaalien metatietojen kuvailuun. Lisäksi paneudutaan meta-metadatan tallentamiseen ja esitellään OMG:n CWM-standardi.

3.5.1 Dublin Core

Dublin Core on kirjastonhoitajien, tietotekniikka- ja SGML-asiantuntijoiden muodostama kansainvälinen yhteistyönä tuotettu metatiedon esittämisstandardi, joka pyrkii rakentamaan yhtenäistä kehystä digitaalisista tallenteista ja julkaisuista kerättävästä metadatasta. *SGML (Standard Generalized Markup Language)* on metakieli, jonka avulla

voidaan määritellä dokumenttien merkintäkieliä, esimerkiksi HTML ja XML ovat SGML:n johdannaisia [Wik08]. Dublin Core -työ on aloitettu vuonna 1995 ja kehitystyötä jatketaan vuosittain kokoontuvissa Dublin Core Metadata Workshop -kokouksissa, joihin osallistuvat useat eri maiden asiantuntijat [Dub08]. Määrittelyjä tehdään vapaaehtoistyönä. Ryhmä on määritellyt 15-osaisen kokoelman sähköisen julkaisun metadatan peruselementeistä (Dublin Core Metadata Element Set). Dublin Coren keskeisiä tavoitteita ovat kuvailutietojen luonnin ja ylläpidon helppous, yhteisesti sovittu semantiikka, yhteensopivuus muiden kuvailustandardien kanssa, formaatin kansainvälisyys, laajennettavuus sekä formaatin käyttäminen eri tiedonhakujärjestelmien kehittämisessä.

Suomenkielistä Dublin Core metadatamäärittelyjen tallennusalustaa ylläpitää Helsingin yliopisto ja se on julkaistu vuonna 1997. Metadataelementit määritellään viidentoista attribuutin avulla. Dublin Core on nyt hyväksytty ISO-standardiksi (ISO 15836) ja suomalaiseksi SFS 5895 -standardiksi. Suomen Standardoimisliitto (SFS) on standardisoinnin keskusjärjestö Suomessa, jonka kotisivujen kautta voi ostaa näitä standardeja [Suo08]. Jäseninä liitossa on elinkeinoelämän järjestöjä ja Suomen valtio.

Dublin Coren 15 kategoriaa ovat seuraavat:

- *Nimi tai Nimeke (title).* Tietoelementille annettu nimi.
- *Tekijä (author or creator).* Tallenteen sisällön tuottanut henkilö tai yhteisö.
- *Aihe (subject and keywords).* Tallenteen aihealueen kuvaus käyttäen asiasanoja.
- *Kuvaus (description).* Vapaamuotoisella tekstillä kuvattu tallenteen aihealueen kuvaus.
- *Julkaisija (publisher).* Tallenteen julkaisija, henkilö tai yhteisö, joka on julkaissut tallenteen.
- *Muu tekijä (other contributors).* Tekijä-kentässä mainitun henkilön tai yhteisön lisäksi, tallenteen tuottamiseen osallistunut henkilö tai yhteisö.
- *Päivämäärä (date).* Tallenteen julkaisupäivämäärä ISO 8601 -standardin mukaisesti tallennettuna, esim. 2005-05-18 (VVVV-KK-PP). Päivämäärälle voidaan lisäksi antaa tarkenteita, joista on kerrottu seuraavassa kappaleessa.

- *Laji (resource type).* Dublin Core -yhteisön ylläpitämän listan mukainen tallenteen kirjallisuuslaji.
- *Formaatti (format).* Tallenteen tiedostoformaatti.
- *Identifikaatiotunnus (resource identifier).* Tallenteen yksiselitteisesti identifioiva tunnus.
- *Lähde (source).* Tallenteen tunnus, johon kuvailtava tallenne perustuu.
- *Kieli (language).* Kieli, jolla tallenne on tehty.
- *Suhde (relation).* Tallenteeseen kiinteästi liittyvien toisten tallenteiden suhde.
- *Kate (coverage).* Tallenteella olevan tiedon ajanjaksoa tai maantieteellistä paikkaa kuvaava tieto.
- *Tekijänoikeudet (rights management).* Tallenteen käyttöoikeuksia hallinnoiva henkilö tai yhteisö. Vapaasti käytössä oleviin tallenteisiin merkitään *'Public domain'*.

Dublin Core -metadatastandardin pohjana on käytetty ISO 11179 -standardia, jossa metadataelementit on määritelty kymmenen seuraavan attribuutin avulla:

- *Nimi.* Tietoelementille asetettu nimi.
- *Tunniste.* Tietoelementille asetettu yksilöllinen tunniste.
- *Versio.* Tietoelementin versio.
- *Rekisteröijä.* Elementin rekisteröinyt taho.
- *Kieli.* Kieli, jolla elementti on määritelty.
- *Määritelmä.* Kuvaus elementin merkityksestä ja oleellisista käyttötavoista.
- *Pakollisuus.* Ilmaisee elementin pakollisuuden tai vapaaehtoisuuden.
- *Tietotyyppi.* Määrittelee elementin kuvaaman tiedon tyypin.
- *Toistuminen.* Määrittelee elementin suurimman toistumismäärän.
- *Kommentti.* Huomautus elementtiä käsittelevää sovellusta koskien.

Dublin Coren elementtejä voidaan toistaa ja ne ovat vapaaehtoisia, ja entiteettiryhmää voidaan laajentaa. [Ste01]. Dublin Core on osoittautunut toimivaksi eri yhteisöjen tietojen yhteensovittamiseen ja sille on saatavissa vakiintuneet sidonnat (bindings) eri ratkaisuille sekä lukuisa joukko vapaasti saatavia työkaluja [Nir02].

Suomessa Opetushallitus käynnisti vuonna 2005 koulusektorille soveltuvan metatietomallin määrittelyn verkko-oppimateriaalien kuvaukseen. Tämän määrittelytyön tuloksena valmistui SVY-yhteistyönä (Suomen virtuaaliyliopisto) *FinnEduMeta*, joka julkaistiin 2007. FinnEduMeta:n lähtökohtana on Dublin Core-metadataformaatti, koska se on RDF-yhteensopiva. Oppimateriaalien kuvauksessa käytetään paljon semanttisen webin tekniikoita ja käytänteitä, joissa RDF on yleinen. FinnEduMeta-metatietomallissa on pyritty löytämään tasapaino riittävän tarkan ja tarpeeksi vaivattoman kuvaustavan välille. Metatietojen voidaan katsoa olevan FinnEduMeta-mallin mukaisia, kun niissä on täytettynä metatietomallissa esitetyt pakolliset kentät ja niissä on noudatettu annettuja sovellusohjeita. Ohjeissa on metatietokenttiin sisällytetty sekä kontrolloitujen että vapaamuotoisten tekstien käyttömahdollisuus. [Ope07]

3.5.2 LOM (Learning Object Metadata)

Oppisisällön metatieto (LOM, Learning Object Metadata) on LTSC:n (Leaning Technology Standards Committee) luoma standardi, jolla pyritään kehittämään erilaisten oppisisältöolioiden löytämistä, hankkimista, yhteistyötä sekä käyttöä [IEE05]. LTSC on IEEE:n (the Institute of Electrical and Electronics Engineers) alaisuuteen perustettu komitea, joka julkaisi ensimmäisen LOM-standardin kesäkuussa 2002 yhdessä DCMI:in (Dublin Core Metadata Initiative) ja the IMS Global Learning Consortium:in kanssa [McC03].

Suomennoksen LOM-standardista on julkaissut TIEKE (Tietoyhteiskunnan kehittämiskeskus ry), joka määrittelee standardin seuraavasti [Tie02]:

> ”Moniosaisen standardin tarkoitus on edistää oppisisältöjen etsimistä, arviointia, hankkimista ja käyttöä, joita esimerkiksi oppijat tai opettajat tai tietokoneohjelmistot tekevät. Tämä moniosainen standardi edistää myös oppisisältöjen yhteiskäyttöä ja vaihdantaa sillä, että se mahdollistaa hakemistojen ja luetteloiden tekemisen ja samalla ottaa huomioon niiden kulttuuristen ja kielellisten asiayhteyksien moninaisuuden, joissa oppisisältöjä metatietoineen käytetään uusiin tarkoituksiin.”

LOM ohjaa opiskelijaa löytämään oppimateriaalia Internetistä käyttämällä hakusanoina otsikoita, kuvauksia tai tiedon tallennuspaikkaa samoin kuin kirjastoissa on totuttu etsimään mielenkiinnon kohteena olevaa kirjallisuutta [LTW08]. Standardi on kansainvälisesti vakiintunut erityisesti web-ympäristöön toteutettujen tietokoneavusteisten oppimisympäristöjen kuvaamisessa. Oppisisältöjen kuvaamiseen käytetään tieto-olioita, jotka voivat olla mitä tahansa, jotka kuvaavat oppimistapahtumaan liittyvää oliota. IEEE määrittelee oppimateriaalin seuraavasti [IEE02]:

> ”Mikä tahansa tieto-olio – digitaalinen tai ei-digitaalinen – jota voidaan käyttää oppimiseen, opettamiseen tai kouluttamiseen.”

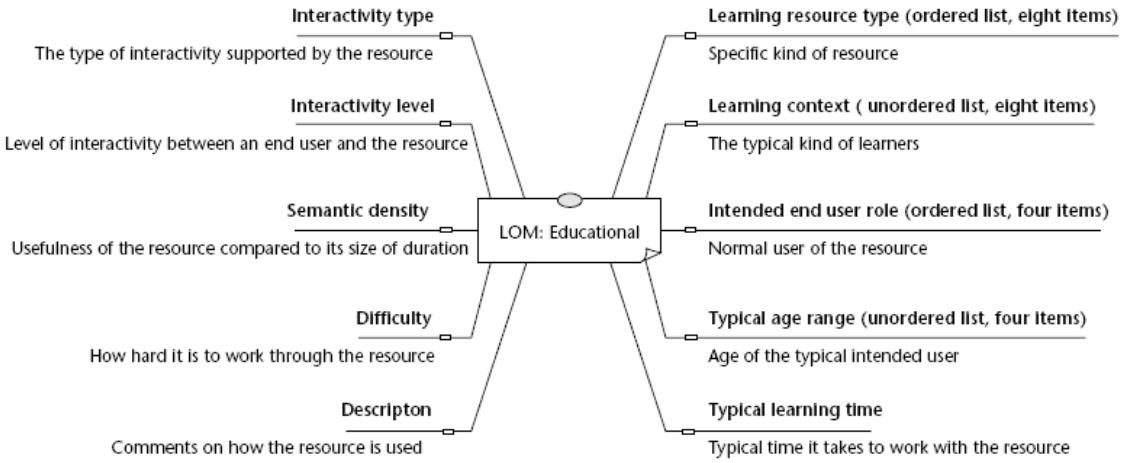
Oliot voivat olla fyysisiä tai abstrakteja [Tou07], digitaalisia tai ei-digitaalisia [Tie02]. LOM-standardia on alettu kehittää vuonna 1997 ja sitä kehitetään edelleen vastaamaan muuttuvan teknologian asettamia haasteita oppisisällön metadatalle.

LOM-standardin perusrakenne muodostuu yhdeksästä peruselementistä, joita voidaan kutsua kategorioiksi [Tou07] tai luokiksi [Tie02].

- *Yleinen (General).* Oppisisällön yksilöivä tunnus, joka kuvaa oppisisällön kokonaisuutta.
- *Elinkaari (Lifecycle).* Oppisisällön kehittymiseen vaikuttaneet historiaa ja nykytilaa kuvailevat tiedot.
- *Tieto metatiedosta (Meta-metadata).* Metadataa itseään kuvaavaa tietoa, ei oppisisältöön liittyvää tietoa.
- *Tekniset ominaisuudet (Technical).* Teknisiä vaatimuksia ja ominaisuuksia kuvaavaa tietoa.
- *Opetukselliset ominaisuudet (Educational).* Oppisisällön keskeisiä opetuksellisia ja pedagogisia ominaisuuksia kuvaavaa tietoa.
- *Oikeudet (Rights).* Oppisisältöön liittyvät tekijän- ja käyttöoikeudelliset tiedot.
- *Suhteet (Relation).* Oppisisällön ja muiden oppisisältöjen suhteita kuvaavaa tietoa.
- *Huomautukset (Annotations).* Oppisisältöön liittyvien kommenttien ja niiden antajan sekä kommenttipäivämäärän tiedot.

- *Luokitus (Classification).* Oppisisällön luokitusjärjestelmään liittyvät tiedot.

Jokainen peruselementti (luokka) sisältää tarkennuksia niihin liitetyillä *asiakohdilla (item)* [Ste01]. Seuraavassa kuvassa on kuvattu LOM-standardin ’Opetukselliset ominaisuudet’ -kategoria (Kuva 7).



Kuva 7. LOM-standardin Opetukselliset ominaisuudet -kategoria [Ste01]

Opetukselliset ominaisuudet -kategoria sisältää seuraavaksi esiteltävät tarkennukset [Tie02].

- `Vuorovaikutustyyppi (Interactivity Type)`. Tietoa oppisisällön ja sen käyttäjän välisestä vuorovaikutuksesta. *Esittävä (expositive)* oppisisältö on tyypillisesti lukemalla oppimista (kirjalliset dokumentit) tai hypertekstissä navigointia, *aktiivisessa (active)* oppisisällössä opiskelija osallistuu oppisisällön muokkaamiseen (simulaatiot, kyselyt, harjoitukset). Vuorovaikutus voi olla myös näiden *yhdistelmä (mixed)* tai *määrittelemätön (undefined).*
- `Oppimateriaalin laji (Learning Resource Type)`. Oppimisresurssin tyyppiä kuvaavaa tietoa. Valintalistalla on 14 vaihtoehtoa: *harjoitus (exercise), simulaatio (simulation), kysely (questionnaire), kaavio (diagram), kuva (figure), piirros (graph), luettelo (index), kalvo (slide), taulukko (table), kertova teksti (narrative text), koe (exam), kokeilu (experiment), ongelman asettelu (problem statement), itsearviointi (self assessment).* Kenttä on pakollinen.

- *Vuorovaikutuksen määrä (Interactivity Level)*. Arvio siitä, kuinka vuorovaikutuksellista materiaali on. Tasoja on viisi: *hyvin alhainen (very low), alhainen (low), keskiverto (medium), korkea (high), hyvin suuri (very high).*
- *Asiasisällön tiiviys (Semantic Density)*. Oppisisällön sisältämän informaation määrän verrannollisuus sen kokoon tai kestoon nähden. Valittavana viidestä tasovaihtoehdosta, jotka ovat samat kuin edellisessä kohdassa.
- *Kohderyhmän rooli (Intended End User Role)*. Tieto siitä, kenelle oppisisältö on suunniteltu ensisijaisesti käytettäväksi. Tasoja on neljä: *opettaja (teacher), oppimateriaalin tekijä (author), oppija (learner), johtaja (manager).* Esimerkiksi kun oppisisältö sisältää oppilaitokselle ohjeet oppimateriaalin jakeluun, merkitään tasoksi johtaja ja kun oppisisällössä kuvataan tämän oppimateriaalin kanssa työskentelemistä, on tasona oppija.
- *Konteksti (Context)*. Tieto siitä, missä ympäristössä oppisisältö on ensisijaisesti suunniteltu käytettäväksi. Tähän merkitään oppilaitostaso, jolla oppisisältöä voidaan käyttää. Esimerkiksi: *esiopetus (primary education), ala-aste (secondary education), yläaste (higher education), lukio (secondary school), yliopisto (university).* Tämä on tekstikenttä.
- *Ikäsuositus (Typical Age Range)*. Tieto siitä, minkä ikäisille käyttäjille oppisisältö on ensisijaisesti suunniteltu. Kenttä on tekstikenttä, johon kirjataan ikäryhmän väli tai minimi-ikä, esimerkiksi 0-5 tai vain aikuisille.
- *Vaikeustaso (Difficulty)*. Tieto siitä, mitä vaikeustasoa oppimateriaali sisältää sen tyypilliselle käyttäjälle. Kenttä on tekstikenttä ja siihen voidaan kuvata vaikeustasoa yleisesti käytössä olevilla termeillä, esimerkiksi hyvin helppo tai keskiverto.
- *Oppimisaika (Learning Time)*. Arvioitu aika tämän oppimateriaalin sisällön suorittamiseen. Kenttään merkitään esimerkiksi ”PT 1H30M” *(PT = Playing Time).*

- *Käyttötapakuvaus (Description)*. Oppisisällön käyttöön suositeltavia tapoja. Tietotyyppinä on tekstikenttä, johon voidaan kuvata esimerkiksi opiskelijalle tarkoitettu itseopiskelumateriaali.
- *Kohderyhmän kieli (Language)*. Oppisisällön kieli on kuvattu luokan *Yleinen* asiakohdassa, mutta tämä kenttä voidaan lisätä myös opetukselliset ominaisuudet -kategoriaan. Tällä kentällä kerrotaan, mitä kieltä oppimateriaalin käyttäjä käyttää. Esimerkiksi suomenkielisille opiskelijoille suunnattu englanninkielinen aineisto, saa arvokseen ”fi”, mutta luokassa *Yleinen*, kieleksi valitaankin ”en”.

LOM-standardissa jotkin kentät ovat pakollisia, mutta useat kentät sallivat vapaan tekstin tallentamisen. Esimerkiksi luokka ”Opetukselliset ominaisuudet” on pakollinen, jossa attribuutti ”Oppimateriaalin laji” on pakollinen, mutta useisiin attribuutteihin voidaan tallentaa tekstiä. Jotkin kentät voidaan täyttää automaattisesti, kuten tekijä ja luontipäivämäärä. Standardin käyttöä hankaloittaa kuitenkin tallennettavien kenttien suuri määrä. Valintalistat muodostavat oman ongelmansa ja joidenkin kenttien attribuutit sisältävät tallentajan subjektiivista tietoa, esimerkiksi oppimateriaalin vaikeustason määritteleminen. [Ste01]

Mikäli LOM-standardin kaikki kentät otetaan käyttöön, on tallennettava jopa 64 tietoa. Vertailun vuoksi mainittakoon, että Dublin Core -standardissa kenttiä on vain 15. Tästä johtuen opetusyhteisöt ovat olleet laiskoja ottamaan LOM:ia käyttöön ja kokeneet standardin käyttöönoton työlääksi, vaikka toisaalta metatiedolle asetetut vaatimukset tiedon saannin oikeellisuudelle ja tehokkaalle käyttämiselle vaativat tallennettavien tietojen suurta määrää [HaR02].

Kanadalaiset opettajat, tutkijat ja opiskelijat ovat edelleen kehittäneet LOM:ia omiin tarpeisiinsa ja luoneet sovellusprofiilin *the Canadian Core Learning Resource Metadata Protocol (CanCore)*, jossa on määritelty LOM:in pohjalta oppimateriaalin kuvaukseen suosituksia käytettävistä kentistä. CanCore antaa myös suosituksia elementtien kuvaukseen ja käytettäviin termeihin sekä määrittelyihin [Can05].

IEEE:n LOM-standardia käytetään opiskelijoiden, opettajien, tutkijoiden ja tutkimusryhmien oppimateriaalien ja tietämyksen jakelun helpottamiseksi [Ste01]. Jotta materiaalien jako ja hyväksikäyttäminen olisi mahdollisimman helppoa, metatietojen tallentamiseen olisi käytettävä resursseja, vaikka tallennusvaiheessa manuaalisten kenttien tallentaminen viekin aikaa ja voi tuntua turhauttavalta.

3.5.3 OMG CWM (common warehouse metamodel)

Common Warehouse Metamode (CWM) on OMG:n (Object Management Group) kehittämä standardi tietovarastojen metatietojen kuvaamiseen. Metadataa voidaan käyttää tietovarastojen kehittämiseen, ylläpitoon, hallintaan ja käyttöön. Koska tietojen hallintaan ja analysointiin tarvitaan tietoja useiden tietovarastojen metadatasta ja metatietomalleista, on oltava organisaation yhteinen *metadatavarasto (metadata repository).* CWM-standardi on siis *meta-metadataa.* CWM-standardi on luotu vastaamaan tarpeita, joilla tietovarastojen metadataa voidaan käyttää yhdessä. [Obj03]

CWM-standardi koostuu kolmen muun standardin pohjalle:

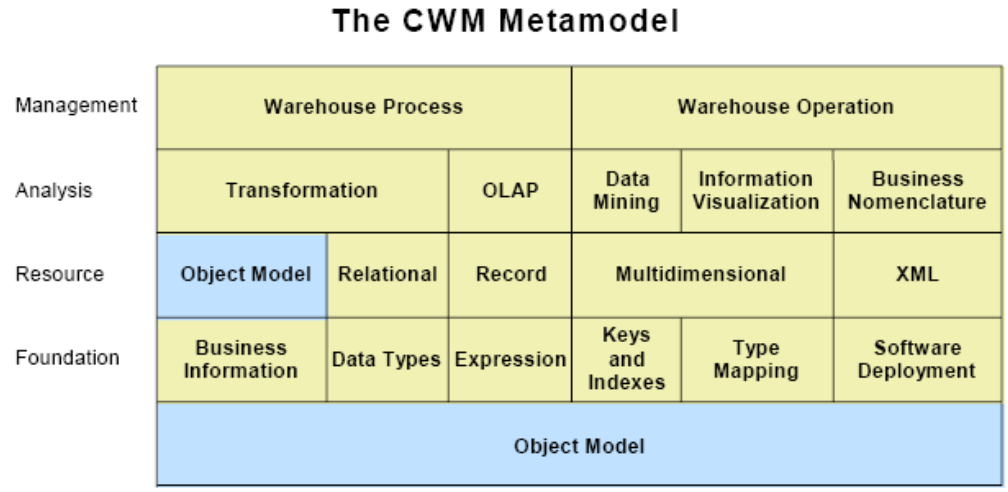
- UML (Unified Modeling Language)
- MOF (Meta Object Facility)
- XMI (XML Metadata Interchange)

UML (Unified Modeling Language) sisältää oman graafisen notaationsa olioiden ja niiden välisten suhteiden esittämiseen. *XMI (XML Metadata Interchange)* käsittää sanoman välitykseen käytetyn formaatin tiedot. *MOF (Meta Object Facility)* määrittelee metatietomallin ja tukee ohjelmallisen rajapinnan tallentaa ja käyttää metatietoja [Obj03].

MOF-standardi tukee kaikenlaista metadataa, joka on kuvailtu käyttäen *oliomallinnus (Object Modelling)* -tekniikoita. Metadatassa voidaan kuvailla järjestelmiä ja niihin liittyviä tietoja monelta kannalta katsottuna, ja tietoja voidaan kuvailla kuinka tarkalla tasolla tahansa riippuen metadatan vaatimuksista [Obj03].

CWM koostuu useista metatietomalleista, jotka voidaan luokiteltu tietosisällön pohjalta.

- *Tietolähteet (Data Resources).* Oliot, relaatiot, tiedostot, moniulotteiset tiedot ja XML-data tallennetaan tähän luokkaan.
- *Tietojen analysointi (Data Analysis).* Tähän luokkaan kuuluvat tiedot, jotka edustavat tiedon muunnoksia, OLAP-tietoja, tiedon louhintaan liittyviä tietoja, tiedon visualisoimiseen liittyviä tietoja ja terminologiaan liittyvää tietoa.
- *Tietovaraston hallinta (Warehouse Management).* Tähän luokkaan kuuluvat metatietomallit, jotka sisältävät tietoja tietovaraston prosesseista ja niiden tuloksista.



Kuva 8. CWM-standardin metatietomalli [Obj03]

CWM-standardin metatietomalli (Kuva 8) sisältää metadataa organisaation useista metatietomalleista, jotka edustavat organisaation tietovarastojen ja liiketalouden kiinnostuksen kohteita yllä olevan kuvan mukaisesti. CWM-metatietomalli käyttää *pakkauksia (packages)* ja *pakkausten hierarkkista rakennetta (hierarchical package structure)* kontrolloimaan tietojen monimutkaisuutta ja tukemaan ymmärtämistä sekä uudelleenkäytettävyyttä.

Pakkaukset (Kuva 8), jotka sisältävät metatietomallissa kuvattujen olioiden luokat ja niiden määrittelyt, on jaoteltu seuraavasti. [Obj03]

- *Oliomallin pakkaus (Object Model package)* sisältää tietoja CWM-pakkausten välisistä suhteista, käyttäytymissäännöistä, suhteista toisiinsa ja esimerkkejä CWM-luokitteluista.
- *Peruspakkaus (Foundation package)* sisältää tietoja metadatan sisältämistä tietotyypeistä, niiden liiketaloudellisesta luonteesta ja siitä, kuinka niitä on mahdollisuus käyttää sekä tiedon sovellusympäristöön liittyvistä tiedoista (avaimet, indeksointi, tyyppimuunnokset).
- *Resurssipakkaus (Resource package)* sisältää tietoja siitä, kuinka metadata on rakennettu ja niiden moniulotteisuudesta sekä XML-yhteensopivuudesta.
- *Analysointipakkaus (Analysis package)* sisältää metadatan tietojen muunnoksiin käytetyistä välineistä, analysointivälineistä, tiedon louhintavälineistä, tietojen visualisointivälineistä sekä systematiikasta ja käytetyistä sanastoista.
- *Hallintapakkaus (Management package)* sisältää tietoja, metadatan tallennusprosesseista ja tietovarastoprosessien tuloksista.

CWM-standardi on suunniteltu tietovarastojen suunnittelijolle ja tietovarastoalustojen toimittajille, palveluntuottajille, tietovaraston kehittäjille ja hallinnoijille, loppukäyttäjille ja informaatioteknologioista vastaaville. Metadata on näille kaikille toimijoille tärkeä työväline tietovaraston tietojen oikeellisuuden ja laadukkuuden takaamiseksi. Tietojen valtavan kasvun myötä myös tietovarastojen ja niihin tallennettavien tietojen määrä on kasvanut niin suureksi, että myös näiden tietojen hallintaan tarvitaan luotettavia työvälineitä. CWM-standardi kuvaillaan paikkaamaan tätä puutetta.

Metatietojen hallinnointi vaatii organisaatiolta sitoutumista tietovarastoprojektiin liittyvään metatietojen määrittelyyn. Ennen kuin pystytään kuvailemaan metatietoja, olisi organisaation toimintaprosesseihin liittyvät käsitteet ja toiminnot pystyttävä kuvailemaan riittävällä tasolla. Vasta tämän jälkeen voidaan kuvailla metatietoja ja meta-

metatietoja. Tämä onnistuu vain, mikäli organisaation johdolla on riittävästi halua ja ymmärrystä panostaa perustoiminnan lisäksi myös tiedon olemuksen ja käsitteiden viemiseksi organisaation toiminnan tasolla työskentelevien ihmisten keskuuteen. Perusjärjestelmiin tietoja tallentavan on helpompi ymmärtää työnsä laadukkuuden merkitys, kun tietojen virheettömyydelle asetettujen kriteerien merkitys tietoja käyttävien analysoijien toimesta on tallentajalle selvillä. Metatiedoista ja meta-metatiedoista on hyötyä vasta, kun on määritelty tietotarpeet ja tiedolle asetetut laatukriteerit.

3.6 Metatietomalli

Ennen tiedon tallentamista tietovarastoon on tiedonlähteet kartoitettava. Tieto voi olla eri tiedonlähteissä eri muodoissa, esimerkiksi tekstitiedostoina tai tietokantajärjestelminä ja niin edelleen [SVV99b]. Tietovaraston rakentamisen aikana on syytä tehdä tietotarveanalyysi, jossa kartoitetaan ne tietolähteet ja tiedot, joilla on tietovaraston käytön kannalta merkitystä. Tarveanalyysin pohjana voidaan käyttää käsitemallia. Tässä vaiheessa määritellään myös tietotyypit ja niiden kuvaukset. Nämä tiedot antavat pohjan metatietomallin rakentamiselle.

Metadata voidaan tallentaa fyysisenä dokumenttina, mutta sen tallennusmuoto riippuu kohteesta. *Metatietomalli (metamodel)* on dokumentti, jossa metadata ja sen sisältämät kentät kuvataan. Siinä kuvataan rakenne ja kuvauksissa käytetty abstraktiotaso. Useiden tietovarastojen erilaisia metatietoja voidaan kerätä yhteiseen *metadatavarastoon (metadata repository)* [Obj03]. Metatietomalli tallennetaan yleensä samaan paikkaan kuin siinä kuvatut metatiedotkin. Tietovaraston sisältöä kuvaava metadata voidaan liittää osaksi tietovarastoa. Metatietomalli auttaa tietovaraston ylläpitäjää tiedon siirron hallinnassa, kun tietoja siirretään useasta tietolähteestä tietovaraston käyttöön. Tietovaraston hallinnan kannalta on oleellista suunnitella metatietomalli huolellisesti.

Metatietomalli märittelee myös sen, mitä kenttiä täydennetään automaattisesti, mitkä kentät ovat pakollisia ja käytetäänkö joidenkin kenttien kuvaamiseen jotakin toista standardia. Esimerkiksi päivämäärien tallentamiseen voidaan käyttää ISO 8601 -standardin

[ISO04] mukaista merkintätapaa, jossa päivämäärä esitetään muodossa VVVV-KK-PP. Metatietomalli on syytä myös versioida.

> ”Ilman systemaattisesti koottuja ja tallennettuja metatietoja tietovarannot ränsistyvät ja tulevat vähitellen käyttökelvottomiksi tai ainakin vaikeasti käytettäviksi.” [Sal05]

Mielestäni Salminen kuvaa metatietomallin tarkoituksen yllä olevassa lainauksessa hyvin. Metatietomalli tulisi määritellä siten, ettei metadatan tallentaminen kuormita liikaa perusjärjestelmien käyttäjiä. Metatietomallin tulisi tukeutua olemassa oleviin standardeihin ja metatietomalliin tulisi kuvata täydennettävät kentät kuvauksineen siten, ettei kenttien täyttämisessä ole tulkinnanvaraisuuksia, vaan tallentajat täyttävät kentät samoin mahdollisimman paljon samalla tavalla.

4 YHTEENVETO

Virallisen standardin puuttuessa metadatamäärittelyt eroavat toistaiseksi vielä huomattavasti eri tieteen- tai taloudenalojen dokumenttienhallinnassa. Eri yhteisöjen määrittelyt rakentuvat hyvin samankaltaisista komponenteista, mutta pienet erot tekevät niiden yhdistämisestä vaikean. Yhteisesti hyväksytyn standardin löytyessä suuri osa näistä ongelmista saadaan korjattua. [Ala01]

Metadatan merkitys on tietovaraston hyväksi käytön kannalta olennaisen tärkeää, kuten on jo aiemminkin todettu. Kuopion yliopiston ja Joensuun yliopiston yhdistyessä yhteiseksi Itä-Suomen yliopistoksi, on työ tietojen viemiseksi yhteiseen tietovarastoon jo aloitettu. Myös tähän työhön osallistuessani, olen kuullut tietovarastoon ja metadataan liittyviä kommentteja yliopiston edustajilta. Seuraavassa muutama vapaasti siteeraamani kommentti:

> ”Dimensiorakenteiden määritteleminen yhteneväiseksi on ensiarvoisen tärkeää, että raportointia voidaan kehittää siten, että myös eri operatiivisista järjestelmistä saatavat tiedot käyttäisivät samoja hierarkkisia rakenteita ja olisivat näin ollen vertailukelpoisia.”

Tietovarastoon siirrettävä tieto on oltava yhteensopivaa ja tietoja pitää pystyä yhdistelemään. Metatietoihin on tallennettava riittävät tiedot raportoinnin ja analysoinnin tekijöiden tueksi, niin että raportointi on luotettavaa ja oikeaa.

> ”Tietovarastoinnin rakentamisessa olisi otettava substanssiryhmien tarpeet huomioon raportointi- ja analysointimahdollisuuksia määriteltäessä.”

Tiedolle on tallennettava myös siihen liittyvät rakenteelliset tiedot. Lisäksi tiedon mittaustarkkuus on oltava riittävä. Esimerkiksi analysoinnissa on pystyttävä porautumaan tiedon tarkimmalle tapahtumatasolle, mutta myös summatasot on oltava saatavilla. Raportointia ja analysointia on tehtävä riittävällä ammattitaidolla siten, että tiedon tulkintaan ei jää epäjohdonmukaisuuksia. Esimerkiksi tiedon ominaisuuksiin kuuluvat erityis-

piirteet ja tiedon tallennuksessa käytettävät poikkeustapaukset voidaan kuvata metatiedoissa.

Metadatastandardiin pohjautuva metatietomallin käyttö auttaa tietovaraston ylläpitäjää ja tietojen käyttäjiä hallinnoimaan tietoja. Tietojen sisältöjen kuvaukset löytyvät metatietomallin määrittelemistä kentistä aina oikean muotoisina ja kenttiä voidaan käyttää hakutekijöinä tai luokittelutekijöinä. Esimerkiksi tutkittaessa tiedon kirjaamiseen liittyviä tietoja antaa metatietomalli tietoja siitä, missä ja kenen toimesta analyyseihin ja raportointiin liittyvät tiedot ovat syntyneet. Lisäksi voidaan käyttää metatietomallin kuvauskenttiin tallennettuja tietoja kertomaan, mitä analyysien ja raporttien sisältämät tiedot pitävät sisällään.

Esimerkiksi taloushallinnon järjestelmässä käytettävä tapahtumien kirjaamiseen liittyvän liikekirjanpidontilin kuvaus voi olla hyödyllinen tieto sille, joka haluaa lajitella yksikkönsä menoja ja eritellä sieltä vielä aineisiin ja tarvikkeisiin käytetyt kuluerät. Jos raportoija ei ymmärrä talousjärjestelmän sisältämien koodien ja kirjanpidon kirjauskäytäntöjä, voi hän tarkistaa metatietoihin tallennetuista kuvauksista ne. Edellisessä tapauksessa liikekirjanpidon tilin kuvaus voisi sisältää tekstin:

> ”Liikekirjanpito käsittää vähintään tuotto- ja kululaskelman sekä taseen laatimiseksi tarvittavat tilit. Käytetään myös tapahtumien lajitteluun. Pakollinen tililuokka kirjanpidossa Kustannuslajilaskenta tapahtuu lkp-tilin kautta.”

Edellinen lainaus on valtionhallinnon yhteisen tietovarastohankkeen (YDW) käsitemäärittelystä [YDW08], joka on kirjattu käsitteelle LKP-tili.

Tiedon kasvun seurauksena alettiin rakentaa tietovarastoja, joiden tietoja käytetään päätöksenteon tukena, raportointiin ja liiketoiminnan suunnitteluun. Pian huomattiin, ettei tiedon määrä ole yksistään riittävä tae liiketoimintaympäristössä tapahtuvien muutosten nopeaan reagoimiseen, vaan tieto on jalostettava tietovarastossa informaatioksi ja edelleen tietämykseksi. Tietämykseen sisältyy paitsi luottaminen saatuihin raportteihin ja niitä sisältäviin tietoihin, myös kokemus- ja tunneperäistä tietojen tulkintaa. Mutta tietämys perustetaan kuitenkin pääsääntöisesti tietojen oikeellisuuden pohjalta. Tietämyk-

sen tueksi tarvitaan tietoja myös tiedon taustalta, ja tästä tarpeesta syntyivät *metadata management* ja *master data management* -ajattelutavat. Metadatan ja master datan tallennuksen peruslähtökohta on, että tieto voidaan osoittaa luotettavaksi ja laadukkaaksi.

LÄHTEET

[Ala01] Alanko M.: *Metadatan kokoaminen, hallinta ja käyttö sähköisillä julkaisualustoilla*. Helsingin yliopisto, Tietojenkäsittelytieteen laitos, Pro gradu-tutkielma, 2001.

[ATK08] *ATK-sanakirja 1*. Talentum, Helsinki 2008.

[Bus08] Business Navigator: http://www.macrosoft.fi/hat/, 2008. (11.6.2008)

[Can05] The CanCore Metadata Initiative: www.cancore.ca/en/, 2006. (25.5.2008).

[CGG98] Cromwll-Kessler W., Gilliland-Swetland A., Gill T.: *Introduction to Metadata – Pathways to Digital Information.* Getty Information Institute, 1998.

[Dev97] Devlin B.: *Data Warehouse – from Architecture to Inplementation.* Addison-Wesley, 1997.

[Dub08] Dublin Core Metadata Initiative: http://purl.oclc.org/dc, 2008. (22.5.2008).

[Duv01] Duval E.: Metadata Standards: What, Who & Why. *Journal of Universal Computer Science,* vol. 7, 2001, s. 591-601.

[EsS98] Eskola J., Suoranta J.: *Johdatus laadulliseen tutkimukseen*. Osuuskunta Vastapaino, Tampere, 1998.

[FIN99] FINLEX, Valtion säädöstietopankki: *Henkilötietolaki 523/1999*. Oikeusministeriö, 1999.
http://www.finlex.fi/fi/laki/alkup/1999/19990523 (11.6.2008).

[FlS005] Flowerday S., von Solms R.: Real-time information integrity = system integrity + data integrity + continuous assurances. *Computers & Security*, 24, 2005, s. 604-613.

[Gar98] Gardner S.: Building the data warehouse. *Communications of the ACM* 41, 9, 1998, s. 51-60.

[Hac99] Hackathorn R.: *Web Farming for the Data Warehouse*. Morgan Kaufmann, 1999.

[HaK06] Han J., Kamber M.: *Data Mining and Techniques, Second Edition.* Morgan Kaufmann Publishers, 2006.

[HaR02] Hatala M., Richards G.: *Global vs. Community Metadata Standards: Empowering Users for Knowledge Exchange*. Springer-Verlag, Berlin Heidelberg, 2002, s. 292-306.

[HoH02] Hong T., Han I.: Knowledge-based data mining of news information on the Internet using congnitive maps and neural networks. *Expert Systems with Applications*, 23, 2002, s. 1-8.

[Hov97] Hovi A.: *Data Warehousing - Tietovarastotekniikka*. Suomen Atk-kustannus Oy, 1997.

[HYK01] Hovi A., Ylinen J., Koistinen H.: *Tietovarastot liiketoiminnan tukena, Asiantuntija-sarja*. Gummerus Kirjapaino Oy, Jyväskylä, 2001.

[IEE02] IEEE, Learning Technology Standards Committee: *1484.12.1, IEEE Standard for Learning Object Metadata*. The Institute of Electrical and Electronics Engineers, Inc., New York, 2002.

[IEE05] IEEE, Learning Technology Standards Committee: *WG12 Learning Object Metamodel.* 2008.
http://ltsc.ieee.org/wg12/index.html (16.5.2008).

[IEE90] IEEE Standard Coordinating Committee of the Computer Society: *IEEE Standard Computer Dictionary, A Compilation of IEEE Standard Computer Glossaries*. The Institute of Electrical and Electronics Engineers, Inc., New York, 1990.

[Inm05] Inmon W. H.: *Building the data warehouse, Fourth Edition*. Wiley Publishing, Inc., Indianapolis, IN, 2005.

[ISO04] ISO: *International Standard, ISO 8601, Third edition. Data elements and interchange formats – Information interchange – Representation of dates and times.* ISO copyright office, Geneve, 2004.

[JLV02] Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P.: *Fundamentals of Data Warehouses, Second Edition*. Springer-Verlag, Berlin Heidelberg, 2003.

[JaS98] Jacob V., Sen A.: Industrial-strength data warehousing. *Communications of the ACM* 41, 9, 1998, s. 29-31.

[Jok02] Jokiniemi, J.: *Heterogeenisen tiedon lataaminen tietovarastoon.* Helsingin yliopisto, Tietojenkäsittelytieteen laitos, Pro gradu -tutkielma, 2002.

[KHL01] Kitchenham B., Hughes R., Linkman S.: Modeling Software Measurement Data. *IEEE Transactions on Software Engineering*, 27, 9, 2001. http://csdl.computer.org/dl/trans/ts/2001/09/e0788.pdf (28.8.2003).

[LTW08] Lee, M.-C., Tsai, K. H., Wang, T. I.: A practical ontology query expansion algorithm for semantic-aware learning objects retrieval. *Computers & Education*, 50, 2008, s. 1240-1257.

[MaW06] Mannino, M. V., Walter, Z.: A framework for data warehouse refresh policies. *Decision Support Systems*, 42, 2006, s. 121-143.

[McC03] McClelland, M.: Metadata Standards for Educational Resources. *Computer*. Volume 36, Issue 11, Nov. 2003, s. 107-109.

[MET04] META Group: *Item Master Data Rationalization. Laying the Foundation for Continuous Business Process Improvement.* A META Group White Paper, Sponsored by Zycus, October, 2004.

[Mit03] Mitri, M.: Applying tacit knowledge management techniques for performance assessment. *Computers & Education*, 41, 2003, s. 173-189.

[MoV05] Morris H. D., Vesset D.: *Managing Master Data for Business Performance Management: The Issues and Hyperion's Solution.* IDC, Framingham, White paper, April, 2005. http://www.oracle.com/technology/products/bi/epm/pdf/idc_whitepaper_managing_master_data_for_epm.pdf (11.6.2008).

[Myk07] Mykkänen J.: *Specification of Reusable Integration Solutions in Health Information Systems*. University of Kuopio, Department of Computer Science, Doctoral dissertation, 2007.

[Nie02] Niemijärvi V.: *Metatieto tietovarastoympäristössä*. Jyväskylän yliopisto, Tietojenkäsittelytieteen laitos, Pro gradu-tutkielma, 2002. http://selene.lib.jyu.fi:8080/gradu/v03/G0000077.pdf (28.8.2003).

[Nir02] Nirhamo L.: *Metatiedon käyttäytyminen, Selvitys metatiedon kuvaus- ja käyttötavoista sekä suositus niiden soveltamisesta Suomessa*. Turun yliopisto, Opetusteknologiayksikkö, 2002. http://arkisto.tieke.fi/standardointi.nsf/38e4483ea7238da4c225650f004a738d/f46eed6f4d5ae19ac2256bc8003ed81c/$FILE/_t9lin8obkd5im8rrebtlo8ubkei26qqbecln0_.doc (1.6.2008).

[Obj03] Object Management Group Inc.: *Common Warehouse Metamodel (CWM) Specification*. Object Management Group Inc., 2003. http://www.omg.org/docs/formal/03-03-02.pdf (16.5.2008).

[Ope07] Opetushallitus: *FinnEduMeta – suomalainen metatietomalli digitaalisten oppimateriaalien kuvaukseen*, 2007.

[PeS07] Pereira C. S., Soares A.L.: Improving the quality of collaboration requirements for information management through social networks analysis. *International Journal of Information Management*, 27, 2007, s. 86-103.

[Pla07] Plattner H.: *Trends and Concept in the Software Industry 2007*. Lecture notes, 2007. http://epic.hpi.uni-potsdam.de/pub/Home/TrendsAndConceptsI2007/12_-_Complete_List_of_Bibliography.pdf (11.6.2008).

[RaD00] Ram P., Do L.: Extracting delta for incremental data warehouse maintenance. *Proc. Of the 16th Internat. Conf. on Data Engineering*, San Diego, California, USA., 2000, s. 220-229.

[Sal05] Salminen, A.: *Metatiedot organisaation sisällönhallinnassa.* Eduskunnan kanslian julkaisu, 7, 2005. http://users.jyu.fi/~airi/papers/Metatietoartikkeli-2005.pdf (1.6.2008).

[SBC03] Singh, G., Bharathi, S., Chervenak, A., Deelman, E., Kesselman, C., Manohar, M., Patil, S., Pearlman, L.: A Metadata Catalog Service for Data Intensive Applications, *Proceedings of The ACM IEEE SC2003 Conference (SC'03)*, ACM, 2003.

[Sor03] Sorsa, M.: *Tietovaraston materiaalistettavien näkymien ja hakemistojen valinta.* Helsingin yliopisto, Tietojenkäsittelytieteen laitos, Pro gradu -tutkielma, 2003.

[Ste01] Steinacker, A., Ghavam, A., Steinmetz, R.: Metadata Standards for Web-Based Resources. *IEEE MultiMedia*, Volume 8, Issue 1, 2001, s. 70-76.

[Suo08] Suomen standardisoimisliitto SFS ry: www.sfs.fi, 2008. (25.5.2008).

[SVV99a] Staudt, M.; Vaduva, A., Vetterli, T.: Metadata management and data warehousing, *Swiss Life*, Information Systems Research, Technical Report number 21, 1999.

[SVV99b] Staudt, M.; Vaduva, A. Vetterli, T.: The role of metadata for data warehousing, *Swiss Life*, Information Systems Research, Technical Report number ifi-99.06, 1999.

[Tie02] Tietoyhteiskunnan kehittämiskeskus ry: *Standardin "Oppisisällön metatieto (LOM)" suomennos.* 2002. http://www.tieke.fi/mp/db/file_library/x/IMG/12148/file/LOM_fi.html (15.5.2008).

[Tie06] Tietotekniikan liitto ry: *ATK-sanasto,* 2006. http://www.ttlry.fi/yhdistykset/osaamisyhteisot/atk-sanasto/viikon_sana/?x20547=434022 (20.5.2008).

[Tou07] Toukola M.: *Oppisisällön metatieto.* Lappeenrannan teknillinen yliopisto, tietotekniikan osasto, Digitaalisen viestintätekniikan seminaari, 2007; www.it.lut.fi/kurssit/06-07/Ti5319200/esitykset/Toukkola%20Mari.doc (15.5.2008).

[Tör99] Törmänen A.: *Tietovarastointi – strategiasta toteutukseen*. Suomen Atk-kustannus Oy, 1999.

[Uus01] Uusitupa H.: *Tiede, tutkimus ja tutkielma – Johdatus tutkielman maailmaan,* WSOY, 2001.

[YSA99] *Yleinen suomalainen asiasanasto*. Helsingin yliopiston kirjasto, 1999. http://vesa.lib.helsinki.fi/ysa/index.html (16.5.2008).

[Wik08] Wikipedia-säätiö: *Wikipedia, Vapaa tietosanakirja:* http://fi.wikipedia.org/wiki, 2008. (25.5.2008).

[WuB97] Wu M., Buchman A.: Research issues in data warehousing. *Proc. of the BTW´97 (Datenbanksysteme in Büro, Technik und Wissenschaft),* Germany, 1997, s. 61-82.

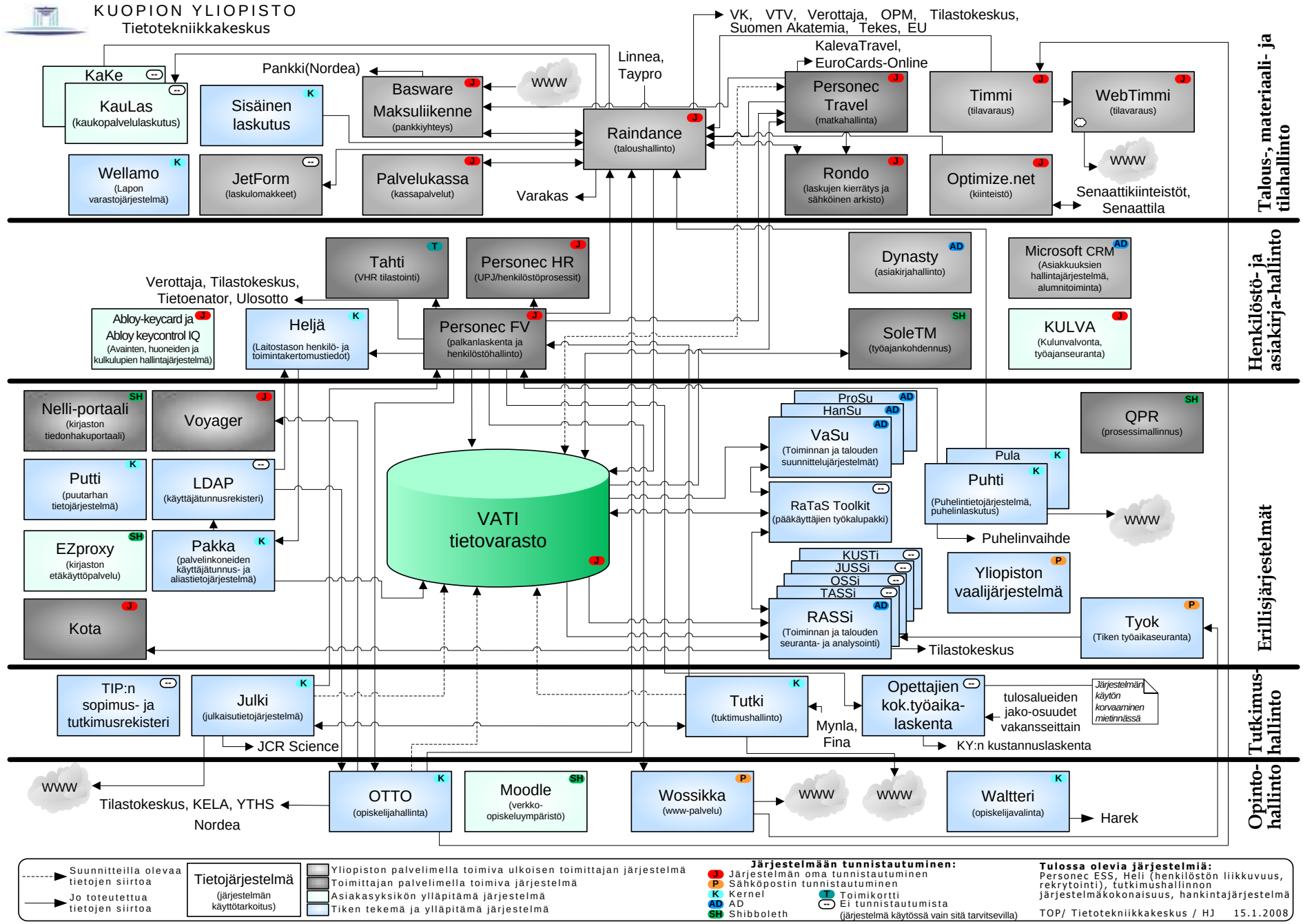
[YDW08] YDW, *Valtionhallinnon yhteinen tietovarasto -hanke,* 2008. http://www.csc.fi/sivut/ydw/kasitemaarittely/taloushallinto (1.6.2008).

LIITTEET

LIITE 1 – Kuopion yliopiston järjestelmäkartta

LIITE 2 – Itä-Suomen yliopiston (v. 2010 alusta) taloustietojen siirron tietokantakuvaus ISTO-tietovarastoon

LIITE 1 – Kuopion yliopiston järjestelmäkartta (http://www.uku.fi/tike/tj/Jarjestelmakartta.pdf)



LIITE 2 – Itä-Suomen yliopiston (v. 2010 alusta) taloustietojen siirron tietokantakuvaus ISTO-tietovarastoon

ISTO-TIETOVARASTO
KIRJANPIDON TAPAHTUMAT - TIETOKANTAKAAVIO
21.5.2008 Heli Junnula, Esa Kaarakainen, Seija Planman