Juha Sinkkonen

Pitkittäistutkimuksessa kerätyn tutkimusaineiston dokumentointi

DDI-spesifikaation tarjoamat mahdollisuudet

Tietojärjestelmätieteen
pro gradu -tutkielma
20.12.2005

Jyväskylän yliopisto
Tietojenkäsittelytieteiden laitos
Jyväskylä

TIIVISTELMÄ

Sinkkonen, Juha Antero
Pitkittäistutkimuksessa kerätyn tutkimusaineiston dokumentointi.
DDI -spesifikaation tarjoamat mahdollisuudet / Juha Sinkkonen
Jyväskylä: Jyväskylän yliopisto, 2005.
129 s.
Pro gradu -tutkielma

Tieteellisen tutkimuksen yhteydessä syntyy yleensä suuret määrät erimuotoista tutkimusaineistoa. Aineiston hallinta aiheuttaa kuitenkin tutkimusta tekeville organisaatiolle usein vaikeuksia. Tutkimusaineisto tulisikin dokumentoida metadatan avulla siten, että dokumentaatio sisältäisi tarkat kuvaukset kaikista aineiston keräykseen käytetyistä menetelmistä sekä jokaisesta tietokokoelmaan tallennetusta dokumentista ja muuttujasta. DDI (Data Documentation Initiative) -spesifikaatio tarjoaa yhteiskuntatieteellistä tutkimusta tekevien tutkimusorganisaatioiden käyttöön XML (Extensible Markup Language) -pohjaisen standardin, jonka avulla tutkimusaineistojen hallintaa ja käytettävyyttä voidaan parantaa. Tutkielmassa tarkastellaan konstruktiivisen tutkimuksen kautta sitä, kuinka pitkittäistutkimuksen dokumentointi voidaan toteuttaa DDI-spesifikaatioon pohjautuen.

Tutkielmassa esitellään tieteellisen tutkimuksen yhteydessä syntyvän tutkimusaineiston ominaisuuksia ja erityispiirteitä, ja lisäksi sitä, kuinka tutkimusaineiston hallintaa voidaan parantaa metadatan avulla. DDI-spesifikaatiosta kuvataan sen tausta, osa-alueet, sen tarjoamat dokumentointimahdollisuudet sekä aikaisempia kokemuksia sen käytöstä. DDI-spesifikaatiota sovellettiin Jyväskylän yliopiston Lapsen Kielen Kehitys ja Geneettinen Dysleksiariski (LKK) -tutkimusaineiston dokumentointiin. Sovelluksen rakentamista varten analysoitiin LKK-tutkimuksen sisällönhallinnan osa-alueet sekä niiden liittymät toisiinsa. Erityistä huomiota kiinnitettiin niiden tarpeiden selvittämiseen, joita LKK-tutkimuksen aineiston kuvailuun kohdistuu.

Tutkielman tuloksena on LKK-tutkimuksen dokumentointitarpeiden perusteella kehitetty DDI-pohjainen metadataskeema, joka sisältää kaikki tutkimuksen ja sen aineiston kuvailulle oleelliset dokumentointikohdat. Lisäksi tutkielman tuloksena saatiin metadatan syöttämistä sekä hakemista varten kehitetyt käyttöliittymät. Tutkielman tuloksena on myös arviointi siitä, kuinka hyvin spesifikaation voidaan katsoa soveltuvan pitkittäistutkimusten dokumentointiin yleisemmin.

AVAINSANAT: tutkimustieto, metadata, XML, DDI-spesifikaatio

ABSTRACT

Sinkkonen, Juha Antero
Documentation of the Research Material Collected in a Longitudinal Study. Possibilities of the DDI specification / Juha Sinkkonen
Jyväskylä: University of Jyväskylä, 2005.
129 p.
Master's Thesis

Scientific studies produce a vast amount of various research material. However, the management of that material often causes problems to the organisations conducting the studies. Research material should be documented using metadata so that it would describe accurately all the methods used to collect the material and also every file and variable saved in the collection. The DDI (Data Documentation Initiative) specification offers an international XML-based standard for the documentation of social and behavioural studies. The DDI improves the controllability and usability of the research material collected in those studies. This thesis examines a constructive study that was carried out to determine how a scientific longitudinal study can be documented using the DDI specification.

This thesis presents the qualities and special features of the research material generated in scientific studies. It also outlines how metadata can be used to improve the management of the material. The DDI specification is unfolded by explaining its background, levels, the documentation possibilities it offers, and also by giving examples of its prior use.

The DDI specification was applied to the documentation of the research material collected in the Jyväskylä Longitudinal Study of Dyslexia (JLD). In order to develop a metadata application, different areas of the JLD study's content management were analysed. Special attention was paid to determine the requirements the JLD study sets for the documentation of its research material.

As an outcome of this thesis a DDI metadata schema was created based on the documentation needs of the JLD study. The schema includes all the necessary elements for describing the study and its material. In addition, as a result of the thesis interfaces for the input and the output of the DDI metadata were developed. Finally, this thesis also valuates the DDI specification's applicability to the documentation of longitudinal studies in general.

KEYWORDS: Research material, metadata, XML, DDI specification

SISÄLLYSLUETTELO

1 JOHDANTO

1.1 Tutkimuksen tausta

Tieteellisen tutkimuksen yhteydessä syntyy usein valtavat määrät erimuotoista tutkimusaineistoa. Tämä aineisto voi koostua esimerkiksi haastattelumuistiinpanoista, tutkimusraporteista, mittaustuloksista tai vaikkapa videotallenteista ja ääninauhoista. Monimuotoisuutensa vuoksi tutkimusaineiston sisältämien tietokokonaisuuksien ominaisuudet poikkeavat toisistaan usein suuresti. Poikkeavia ominaisuuksia ovat muun muassa tallennusformaatti, analysointiaste sekä tietokokonaisuuden saatavuus. Tämän lisäksi tietoa virtaa tutkimustietokantaan usein monista eri lähteistä, koska aineiston tallennuksesta vastaa lukuisat tutkimushenkilöt. Tällaisen massiivisen ja monimuotoisen tietokokoelman hallinta aiheuttaa tutkimusta tekeville organisaatioille usein vaikeuksia. Vaikeudet voivat johtua esimerkiksi siitä, ettei tutkimustietokantaan tallennetun aineiston ominaisuuksia tai erityispiirteitä tunneta tarpeeksi hyvin. Tämän lisäksi yksi yleisimmistä ongelmia aiheuttavista tekijöistä on tietokokoelman puutteellinen dokumentointi. Tutkimusaineisto tulisikin dokumentoida metadatan avulla siten, että dokumentaatio sisältäisi muun muassa tarkat kuvaukset käytetystä tutkimusmenetelmästä, jokaisesta tietokokoelmaan tallennetusta dokumentista ja muuttujasta, sekä tiedot yhteyksistä eri tutkimusaineiston osien välillä. Metadatan lisääminen tutkimusaineistoon parantaa sen hallittavuutta mikä puolestaan johtaa myös sen käytettävyyden paranemiseen, koska tällöin organisaatio pystyy hyödyntämään sen hallussa olevaa tutkimustietoa kontrolloidusti ja tehokkaasti. Kattavien metatietojen lisääminen tutkimusaineistoon olisi tärkeää myös tulevia sukupolvia silmällä pitäen.

Tutkielmassa käsiteltävän DDI (Data Documentation Initiative) -spesifikaation tarkoituksena on tarjota tutkimusta tekeville tutkimusorganisaatiolle työkalut tutkimusaineistojen tehokkaaseen ja kokonaisvaltaiseen hyödyntämiseen. DDI on vuonna 1995 aloitettu kansainvälinen hanke, jonka avulla pyritään kehittä-

mään XML (Extensible Markup Language) -pohjainen standardi yhteiskuntatieteellisen tutkimusaineiston dokumentointia varten. DDI-spesifikaatio mahdollistaa metadatan tallennuksen useassa eri tasossa nykyiset tekniset vaatimukset täyttävällä tavalla. DDI:n avulla tutkimusorganisaatio voi lisätä sen hallussa olevan aineiston hallittavuutta, käytettävyyttä ja ylläpidettävyyttä kuvailemalla kaikki aineiston syntyyn ja sen sisältöön liittyvät tekijät.

1.2 Tutkimusongelma ja -menetelmä

Tässä tutkielmassa paneudutaan tieteellisessä tutkimuksessa syntyvän tutkimusaineiston luonteeseen ja erityispiirteisiin sekä metadatan rooliin tutkimusaineiston hallinnassa. Tutkimusaineistojen kuvailun osalta tutkielman aihepiiriä rajataan edelleen koskemaan nimenomaisesti useammasta tutkimuskerrasta koostuvien pitkittäistutkimusten käsittelyä. Tutkielman avulla pyritään vastaamaan seuraaviin tutkimusongelmiin:

1) Miten pitkittäistutkimukseen liittyvää metadataa voitaisiin syöttää ja hakea DDI-spesifikaatiota hyödyntäen?

2) Kuinka hyvin DDI-spesifikaatio pystyy vastaamaan niihin vaatimuksiin, joita pitkittäistutkimuksessa syntyvän aineiston dokumentointi sille asettaa?

Tehty tutkielma on luonteeltaan konstruktiivinen, ja siinä DDI-spesifikaation soveltuvuutta pitkittäistutkimuksen kuvailukieleksi selvitetään käytännön tutkimusympäristössä. Edellä mainittuihin tutkimusongelmiin haetaan ratkaisua tarkastelemalla DDI-spesifikaation soveltamista varten tehdyn puolivuotisen työjakson vaiheita ja sen kautta saatuja tuloksia. Soveltamisen tarkoituksena oli selvittää mahdollisuuksia hyödyntää DDI-spesifikaation mukaista dokumentaatiota LKK-tutkimuksen aineiston kuvailussa. Tutkielman tulokset johdetaan

konstruktiivisen tutkimuksen kautta saaduista tuloksista, ja lisäksi tutkimuksen pohjatietoina käytetään ennen sitä tehtyä kirjallisuuskatsausta, jonka avulla tutustutaan tieteellisessä tutkimuksessa syntyvään tutkimusaineistoon kartoittamalla muun muassa sen alkuperää, koostumusta ja erityispiirteitä. Kirjallisuuskatsauksen toisena osana tarkastellaan tutkimusaineistojen hallinnassa käytettävän metadatan roolia ja merkitystä, metadatan eri tasoja sekä sen tallennukseen liittyviä seikkoja. Kirjallisuuskatsauksen lisäksi konstruktiivisen tutkimuksen taustatietona toimii DDI-spesifikaatiosta tehty yksityiskohtainen selvitys, jossa tarkastellaan muun muassa DDI-hankkeen tavoitteita ja historiaa, spesifikaation eri osa-alueita sekä DDI:n liittyviä, tiedossa olevia haasteita.

1.3 Tutkielman tavoitteet

Konstruktiivisen tutkimusosuuden kautta pyritään selvittämään sitä, kuinka LKK-tutkimuksen hallussa olevaa tutkimusaineistoa voitaisiin kuvailla DDI-spesifikaation mukaisesti, ja edelleen kuinka syötettyä dokumentaatiota voitaisiin hyödyntää mahdollisimman tehokkaasti. Spesifikaation soveltaminen aloitettiin tarkastelemalla LKK-tutkimuksen sisällönhallintaa siinä esiintyvien toimijoiden, toimintojen, sisältöyksiköiden, järjestelmien sekä niiden välisten tietovirtojen näkökulmasta. Erityistä huomiota kiinnitettiin myös niihin tarpeisiin, joita tutkimusaineiston kuvailuun kohdistuu LKK-tutkimuksen piirissä. Näiden tarpeiden pohjalta kehitettiin tutkimuksen käyttöön DDI-pohjainen metadataskeema, joka sisältää kaikki tutkimuksen ja sen aineiston dokumentoinnin kannalta oleelliset kohdat. Metadataskeema kehitettiin iteratiivisesti arvioimalla kunkin kehitysversion toimintaa LKK-tutkimuksen yhteyshenkilön kanssa. Lopullista metadataskeemaa hyödyntäen osa LKK-tutkimuksen laajasta aineistosta kuvailtiin DDI-muotoisen metadatan avulla. Lisäksi LKK-tutkimuksen dokumentointitarpeiden pohjalta kehitettiin metadatan syöttöä sekä olemassa olevan metadatan hakemista ja selaamista varten omat käyttöliittymät, joiden tar-

koituksena on mahdollistaa DDI-spesifikaation mukaisen metadatan hyödyntäminen kohdeorganisaatiossa.

Toiseen tutkimusongelman vastaukset pyritään johtamaan LKK-tutkimuksen DDI-pohjaisesta dokumentoinnista saatujen tietojen perusteella. Käytännön dokumentointityö osoitti minkälaisia erityisvaatimuksia vaatimuksia pitkittäistutkimus asettaa sitä kuvaavalle dokumentointikielelle, ja kuinka hyvin DDI pystyy vastaamaan näihin tarpeisiin. Saatujen kokemusten pohjalta pyritään arvioimaan DDI-spesifikaation soveltuvuutta pitkittäistutkimusten dokumentointiin yleisemmin puntaroimalla niitä seikkoja, jotka tukevat spesifikaation valintaa organisaation dokumentointikieleksi, ja mitkä seikat puolestaan tekevät spesifikaation valinnasta vähemmän haluttavan.

Yhtenä tutkielman keskeisenä tavoitteena voidaan pitää tutkimustietoon ja sen dokumentoitiin liittyvien seikkojen selvittämistä sekä niiden avaamista lukijalle. Lisäksi tutkielma antaa kattavan tietopaketin yhteiskuntatieteellisten tutkimusten dokumentointiin tarkoitetusta DDI-spesifikaatiosta, selvittämällä muun muassa sen tarjoamat dokumentointimahdollisuudet.

1.4 Tutkielman sisältö

Tutkielman aiheiden käsittely etenee siten, että luvussa 2 selvitetään tutkimustietoon (2.1) ja metadataan (2.2) liittyviä seikkoja. Tutkimustiedon osalta käsiteltäviä asioita ovat sen alkuperä ja koostumus (2.1.1), erityispiirteet (2.1.2) sekä tutkimustiedon jakamiseen liittyvät kysymykset (2.1.3). Metadatan osalta pohditaan sen käyttöä tutkimustiedon hallinnassa käsittelemällä sen roolia ja merkitystä (2.2.1), metadatan tallennusta ja tasoja (2.2.2) sekä sen käyttöön liittyviä vaikeuksia (2.2.3).

Tutkielman kolmannessa luvussa perehdytään puolestaan yhteiskuntatieteellisten tutkimusten sekä niissä syntyvän aineiston dokumentointiin tarkoitettuun DDI-spesifikaatioon. Luvun käsittely aloitetaan esittelemällä DDI:tä yleisesti (3.1.), minkä jälkeen siirrytään kertomaan spesifikaation historiasta ja hallinnoinnista (3.2). Kohdassa 3.3 tarkastelun kohteena on DDI-spesifikaation ja XML-kielen välinen suhde, ja kohdassa 3.4 annetaan tarkka kuvaus spesifikaation eri osa-alueista. Tätä kuvausta on myös jatkettu liitteenä 1 olevassa luettelossa. Kolmannen luvun lopuksi lukijalle kerrotaan DDI:tä koskevista rajoituksista ja ongelmista (3.4).

Tutkielman neljännessä luvussa esitellään DDI-spesifikaation soveltamista LKK-tutkimuksen dokumentointiin. Kohdassa 4.1 esitellään kohdeympäristö, tarkastelemalla sen sisällönhallintaa (4.1.2) Salmisen (2005b, 5-6) esittämän viitekehyksen (4.1.1) avulla. Lisäksi kohdeympäristön esittelyssä paneudutaan niihin tarpeisiin, joita sen tutkimusaineiston dokumentointiin kohdistuu (4.1.3). Kohdassa 4.2 siirrytään tarkastelemaan tutkimuksen kulkua sekä sen tiedonkeruutapoja. Tutkimuksen aikana toteutetut sovellusratkaisut on puolestaan avattu kohdassa 4.3 esittelemällä valittuja toteutustekniikoita sekä LKK-tutkimuksen käyttöön kehiteltyä DDI-pohjaista metadataskeemaa (4.3.1), sovelluksen yleistä toimintaa (4.3.2) sekä metadatan syöttöön (4.3.3) ja hakemiseen (4.3.4) kehitettyjä käyttöliittymiä. Kohdassa 4.4 toteutettua sovellusratkaisua arvioidaan suhteessa LKK-tutkimuksen dokumentointitarpeisiin, minkä jälkeen tarkastelu siirrettään konstruktiivisen tutkimuksen aikana kohdattujen vaikeuksien käsittelyyn (4.5). Ennen luvusta 6 löytyvää yhteenvetoa käsitellään vielä luvussa 5 tutkielman yleisiä johtopäätöksiä tarkastelemalla DDI-spesifikaation soveltuvuutta pitkittäistutkimusten dokumentointiin.

2 TUTKIMUSTIETO JA METADATA

Johdannossa esitettyihin tutkimusongelmiin paneudutaan aluksi tutustumalla kirjallisuuskatsauksen avulla tieteellisessä tutkimuksessa tallennettavan tiedon luonteeseen ja erityispiirteisiin sekä metadatan käyttöön tutkimustiedon hallinnassa. Tämä luku tutustuttaa lukijan tutkielman aihepiiriin ja luo teoriapohjan tutkielmassa myöhemmin esitettävän konstruktiivisen tutkimuksen tarkastelulle. Luvun käsittely on jaettu kahteen osaan siten, että aluksi käsitellään tieteellisten tutkimusten yhteydessä syntyvää tutkimustietoa (2.1) ja sen jälkeen käsittely siirrettään metadataan ja sen käyttöön tutkimustiedon hallinnassa (2.2).

2.1 Tutkimustieto

Tässä kohdassa käsittelen tutkimustietoon liittyviä tekijöitä. Aluksi käsitellään sitä miten tutkimustieto syntyy ja millainen sen koostumus on. Tämän jälkeen otetaan kantaa tutkimustiedon erityispiirteisiin ja luvun lopuksi käsitellään tutkimustiedon jakamista tutkijoiden kesken.

2.1.1 Tutkimustiedon alkuperä ja koostumus

Toivonen, Salmenkivi ja Verkamo (2004) esittävät, että tutkimustietoa syntyy pääasiallisesti kahdella eri tavalla:

1. Tutkimustietoa syntyy tieteellisen tutkimuksen kautta. Tietoa saadaan siis ennen kaikkea tehtyjen tutkimusten pohjalta. Erilaisia tiedon lähteitä ovat ilmiöistä tehdyt havainnot, kokeelliset tutkimukset sekä suoritetut simulointikokeet.

2. Tietokokoelmia syntyy toiminnan sivutuotteena. Tällöin tallennettu tieto on kerätty alun perin johonkin muuhun tarkoitukseen. Tästä syystä

myös tiedon rakenne riippuu yleensä vahvasti alkuperäisestä käyttötarkoituksesta.

Lisäksi tallennetun tiedon tallennusajankohta ja -muoto vaihtelevat eri tutkijoiden mukaan. Jotkut tutkijat tallentavat tutkimustietonsa tietokantaan vasta siinä vaiheessa, kun he ovat saaneet tutkimuksensa päätökseen, kun taas jotkut tutkijat tallentavat tietoa kaikissa tutkimuksensa vaiheissa, jolloin tietokantaan tallentuu analysointiasteeltaan erimuotoista dataa. (Thomson, Adams, Cowley & Walker 2003, 27)

Birnholtz ja Bietz (2003, 341–342) käsittelevät tekstissään puolestaan sitä, miten ja milloin tutkimustieto syntyy. He esittävät, että tieto kerätään joko tietovirtojen (data streams) tai tietotapahtumien (data events) kautta. Tutkimustiedon keräämisellä tietovirroista tarkoitetaan sitä, kun tietoa virtaa tutkijoille suhteellisen tasaisena pysyvistä tietovirroista pitkän ajan kuluessa. Muun muassa lääketieteellinen tai psykologinen pitkittäistutkimus hyödyntää tällaisia tietovirtoja. Esimerkkinä voidaan mainita vaikka AIDS-lääkkeiden tutkimus, jossa tutkijat saavat potilaiden verinäytteitä tutkittavikseen tasaisin väliajoin. Ongelmana tietovirtoja hyväksi käyttävässä tiedonkeruumenetelmässä on tutkimuksessa tapahtuvat muutokset. Itse tutkittavassa ilmiökentässä tapahtuva kehitys voi heikentää aikaisemmin kerätyn tiedon arvoa, kun taas esimerkiksi tutkimushenkilökunnassa tai tutkimustiloissa tapahtuvat muutokset voivat vaikeuttaa tiedon keräämistä ja tutkimuksen suorittamista jollain muulla tavalla. Tietotapahtumista kerättävällä tutkimustiedolla tarkoitetaan puolestaan sitä, kun tietoa kerätään erillisistä tapahtumista. Tällaisessa tiedonkeruutekniikassa joudutaan tapahtuman mittaamista usein suunnittelemaan monia kuukausia ja itse tapahtuma saattaa kestää vain muutamia sekunteja. Esimerkiksi maanjäristysten tutkimisessa tietoa kerätään tietotapahtumien kautta. Joskus myös mainittujen tiedonkeruumenetelmien yhdistely voi olla paikallaan. Näin toimitaan esimerkiksi silloin kun mitattava ilmiö kestää useita viikkoja. Ilmiöstä saadaan siis tietoa tasaisena virtana mutta toisaalta tiedonkeruun intensiteetti muistuttaa

tietotapahtumien mittaamisessa ilmenevää intensiteettiä. Esimerkiksi aurinkotuulien tutkimisessa yhdistellään kahta edellä mainittua tiedonkeruumenetelmää. (Birnholtz & Bietz 2003, 341–342)

Sekä Thomson, Adams, Cowley & Walker (2003, 27) että Jacobs & Humphrey (2004, 27–29) ottavat kantaa siihen kuinka tärkeää tutkimustiedon tallennus ja arkistointi on. Tiedon tallennus ja tarkka kuvaaminen metadatan (kts. Luku 2.2) avulla on välttämätöntä, koska kerätyllä tiedolla on usein pitkäaikaista tieteellistä arvoa. Tieto on usein myös luonteeltaan sellaista, ettei se ole uudelleen generoitavissa tai uudelleen generointi olisi ainakin kohtuuttoman kallista. Myös Birnholtz ja Bietz (2003, 339) korostavat sitä, että tutkimusten kautta saatu ja tallennettu tieto sekä sen jakaminen ovat perustavaa laatua oleva osa tieteen tekemisessä.

Toivosen, Salmenkiven ja Verkamon (2004) mukaan tutkimustieto koostuu kolmesta erilaisesta datasta:

1. Mikrodata eli raakadata. Tällä tarkoitetaan havaintojen, tutkimusten ja simulointien pohjalta kerättyä dataa. Raakadata koostuu tutkittavaan ilmiöön liittyvistä yksittäisistä tietoalkioista ja sen määrä riippuu pitkälti havaintojen säännöllisyydestä (kuinka usein) ja tiheydestä (hylätyt arvot ja datan tiivistys)

2. Makrodata eli analysoitu data. Tällä tarkoitetaan puolestaan raakadatalle tehdyn käsittelyn tuloksena syntynyttä dataa. Käsittely voi kohdistua myös analysoituun dataan, jolloin kyseessä voi olla esimerkiksi siitä tehtävät yhteenvedot. Makrodata on periaatteessa uudelleen konstruoitavissa mikrodatasta, mutta se voi olla erittäin työlästä.

3. Metadata. Metadata tarkoitetaan mikro- ja makrodataa kuvailevaa dataa, eli tietoa tiedosta. Tarkempi selitys metadatalle ja sen käytölle löytyy kohdassa 2.2.

2.1.2 Erityispiirteet

Yhtenä tutkimusaineiston erityispiirteenä voidaan pitää sen pitkäikäisyyttä. Kuten aikaisemmin todettiin, tallennetulla tutkimustiedolla on pitkäaikaista tieteellistä arvoa (Thomson, Adams, Cowley & Walker 2003, 27). Kaupallisen tiedon arvo puolestaan vanhenee usein nopeasti muun muassa teknisen kehityksen myötä. Tutkimustiedon arvon säilyminen riippuu tosin pitkälti tutkimusalasta. Esimerkiksi tietojärjestelmätieteen tutkimus ei näe samanlaista arvoa kymmeniä vuosia vanhalla tutkimustiedolla kuin esimerkiksi psykologian tutkimus.

Tutkimustieto on myös rakenteeltaan moniulotteista. Aineistoon sisältyvät tietokokonaisuudet, kuten yksittäiset haastattelutulokset sisältävät usein kymmeniä tai jopa satoja muuttujia, joita käytetään kuvailemaan kyseisen haastattelun tuloksia. Moniulotteisuus ilmenee siten, että tietokokoelman eri osissa saatetaan käyttää samoja muuttujien nimiä samojen tai samantapaisten tietokokonaisuuksien yhteydessä. Esimerkiksi muuttuja, jolla määritellään jotain haastattelussa tai testissä saatua vastausta, saattaa esiintyä täysin samannimisenä jonkin toisen haastattelun tai testin yhteydessä. Tämä vaikeuttaa esimerkiksi muuttujantason hakujen tekemistä, koska hakua ei voida pelkän muuttujan avulla rajata johonkin tiettyyn haastatteluun tai testiin. Muuttujien merkitykset saattavat myös vaihdella ajan kuluessa ja lisäksi muuttujajoukosta saattaa olla samanaikaisesti käytössä useita eri-ikäisiä versioita, jotka poikkeavat toisistaan esimerkiksi nimeämiskäytännön osalta. Tutkimustieto on myös rakenteellisesti monimutkaista. Tämä ilmenee tietokokonaisuuksien ja niiden osien välisinä monenlaisina ja monimutkaisina riippuvuuksina. (Toivonen, Salmenkivi & Verkamo 2004)

Tutkimusdatan erityispiirteenä voidaan pitää myös sen maksuttomuutta ja jaettavuutta. Tällä tarkoitetaan sitä, että suuri osa tieteellisten tutkimusten kautta saaduista tutkimustuloksista on vapaasti kaikkien tutkijoiden käytettävissä. Jacobs ja Humphrey (2004, 27) pitävätkin tutkimustiedon maksuttomuutta ensisi-

jaisena olettamuksena sille, että tieteellinen toiminta on ylipäätään mahdollista. He myös vastustavat tutkimusdatan pitämistä tutkimuksen tehneiden tutkijoiden yksityisenä omaisuutena, koska tällainen lähestymistapa vain rajoittaa tutkijoiden tasa-arvoista mahdollisuutta hyödyntää olemassa olevaa tieteellistä dataa, ja näin ollen se myös rajoittaa tieteellisen tutkimuksen tekemistä yleisemmällä tasolla. Jacobsin ja Humphreyn (2004, 28) mielestä tutkimusdatan tulisi toimia tieteellisen yhteisön yhteisenä pääomana, jota kaikki tutkijat voivat hyödyntää. Tämän kaltaista tiedon maksuttomuutta ja jaettavuutta kaikille halukkaille tahoille ei puolestaan voida pitää ominaisena piirteenä kaupallisten organisaatioiden hallussa olevalle datalla. Päinvastoin, liiketoimintaa suorittavissa yrityksissä, yrityksen hallussa oleva tieto nähdään nimenomaan kyseisen yrityksen pääomana, jolloin sillä nähdään olevan myös taloudellista arvoa ja kilpailuetua.

2.1.3 Tutkimustiedon jakaminen

Tämän kohdan tarkastelu perustuu Birnholtzin & Bietzin (2003, 339–346) sekä Bosen (2002, 15) esittämiin ajatuksiin.

Tieteellisen tutkimuksen yhteydessä kerättävän tiedon jakaminen toisten tutkijoiden kesken on välttämätöntä tieteen tekemiselle. Usein oman tutkimuksen suorittaminen on mahdollista vain toisilta tutkijoilta saadun tiedon avulla, ja myös omien tutkimustulosten tieteellinen validointi tapahtuu tarjoamalla ne muiden tutkijoiden käytettäviksi ja arvioitaviksi. Tiedon jakamisessa ja sitä kautta tieteellisessä toiminnassa voidaan onnistua vain jos saman alan tutkijoiden välillä tapahtuu yhteistyötä ja sosiaalista vuorovaikutusta. Perinteisesti tämä yhteistyö on ilmennyt työskentelynä fyysisessä läheisyydessä toisten tutkijoiden kanssa jossain tutkimusta tekevässä laitoksessa, kuten laboratoriossa. Tällöin tiivis tutkijayhteisö on tarjonnut sen tutkijoille sosiaalisen organisaation tiedon jakamista ja arviointia silmällä pitäen. Nykyään tutkimusta tekevät ryhmät ovat usein levittäytyneet maantieteellisesti laajalle alueelle, jolloin yhteis-

työtä pidetään yllä teknologian avulla. Jotta tässä voitaisiin onnistua, tulee yhteistyössä mukana olevien tutkimusryhmien toiminta ja tarpeet ymmärtää hyvin.

Tieteellisen tutkimustiedon jakamisella voidaan katsoa olevan kaksi tärkeää roolia:

1. Tiedon jakamista on perinteisesti pidetty modernin tieteellisen toiminnan tunnusmerkkinä. Avoimuus tieteellisessä prosessissa mahdollistaa tutkimuksen tulosten varmennuksen. Tiedon jakamisen kautta tutkijoiden on myös mahdollista viedä toisten tutkijoiden tekemää tutkimusta pidemmälle ottamalla aikaisemman tutkimuksen tulokset oman tutkimuksensa lähtökohdaksi. Näin tekemällä pystytään myös laajentamaan ja syventämään kyseisen aihepiirin tietämystä.

2. Uusien erittäin suurten tieteellisten projektien myötä tiedon jakamiselle on muotoutunut uusi tärkeä rooli. Näissä projekteissa käytetään hyväksi tietoa, jonka keräämisen analysointiin ovat osallistuneet useat ihmiset, instituutiot ja tutkimuslaitokset. Tämän kaltaisessa tiedon jakamisessa on kyse paljon muustakin kuin vain yksittäisten datakokoelmien vaihtamisesta. Tällöin kyse on myös suuremman mittakaavan yhteistyön tukemisesta sekä mittavien kansainvälisten tietokokoelmien synnystä.

Käytännössä tutkimustulosten jakaminen ja uudelleenkäyttö on kuitenkin osoittautunut ongelmalliseksi. Yksi usein vastaan tulevista ongelmista on tutkijoiden haluttomuus tiedon jakamiseen. Usein nähdään, että tutkijoiden mahdollisuus hyötyä oman tutkimuksen kautta saadusta tiedosta riippuu siitä onko heillä yksinoikeus hallita kyseistä tietoa. Mahdollisia tutkimustiedosta saatavia hyötyjä ovat puhtaasti rahalliset hyödyt, kuten esimerkiksi apurahat, mutta myös maineesta kilpaillaan tutkijapiireissä kiihkeästi. Haluttomuus tutkimustiedon jakamiseen voi johtua myös siitä, että tutkimuksen rahoittajana toimiva taho haluaa kontrolloida tutkimuksen tulosten käyttöä ja jakamista, koska se katsoo

niiden olevan taloudellinen resurssi. Toinen jakamista vaikeuttava tekijä on se, että tutkijat eivät usein tiedä keiden tutkijoiden hallussa heidän tarvitsemansa tieto on ja missä tämä tieto sijaitsee. Kun oikeanlainen tieto löydetään, joudutaan usein vielä käymään neuvotteluja tiedon omistajan kanssa, jotta luottamus ja sitä kautta pääsy tietoon voidaan saavuttaa. Kolmantena vaikeutena tutkimustiedon jakamisessa on toisilta tutkijoilta saadun tiedon ymmärtäminen. Jotta tietoa voitaisiin ymmärtää, tarvitaan taustatietoa sen synnystä, eli siitä, miten data on kerätty ja miten sitä on analysoitu. Usein tarvitaan tietoa myös siitä onko saatu tutkimustieto tarpeeksi laadukasta ja onko sen alkuperäinen käyttötarkoitus yhteensopiva suunnitellun käyttötarkoituksen kanssa. Jaetun tutkimustiedon ymmärtämistä voidaan oleellisesti parantaa metadatan avulla. Metadata mahdollistaa tiedon eri osien tarkan kuvailun ja sen avulla voidaan myös kuvata esimerkiksi tiedon keräämiseen käytetyt metodit.

Edellä mainittuja tiedon jakamiseen liittyviä ongelmia voidaan helpottaa erilaisten tiedon jakamista tukevien tietojärjestelmien avulla. Esimerkiksi kattavan ja monipuolisen metatiedon syöttäminen voidaan varmistaa käyttämällä standardoitua kuvausmenetelmää, ja lisäksi esimerkiksi käyttäjätunnusten ja salasanojen avulla voidaan määrittää kenellä on oikeus päästä käsiksi tiettyyn tietokokoelmaan. Teknisten menetelmien lisäksi on tärkeää ottaa huomioon myös sosiaaliset tekijät, jotka mahdollisesti vaikuttavat tutkimustiedon jakamiseen. Birnholtzin ja Bietzin (2003, 340) mukaan pitäisikin keskittyä pelkän tiedon jakamiseen sijasta kehittämään teknologioita, joiden avulla voidaan todella tukea tieteen tekemistä. He jatkavat, että näiden teknologioiden kehittäjien tulisi ymmärtää sekä tutkimustiedon tieteellinen rooli, joka sillä on tietämyksen synnyssä että tutkimustiedon sosiaalinen rooli, joka sillä on tieteellisessä toiminnassa.

2.2 Metadatan käyttö tutkimustiedon hallinnassa

Yksinkertaisimman määritelmän mukaan metadatalla tarkoitetaan dataa, jonka avulla kuvaillaan toista dataa, jotta sen käytettävyyttä voitaisiin parantaa (Marshall 1998, 162). Hieman tarkemman määritelmän mukaan voitaisiin sanoa, että metadata on tietoa, jota tarvitaan datan (raakadata ja analysoitu data) hakemiseen, käsittelemiseen, jäsentämiseen ja tulkitsemiseen, ja että sen tavoitteena on varsinaisen datan ymmärrettävyyden, käytettävyyden ja ylläpidettävyyden parantaminen (Toivonen, Salmenkivi & Verkamo 2004).

Metadatalla on tutkimustiedon hallinnassa erittäin tärkeä rooli. Suurin osa hallinnallisista toimenpiteistä tapahtuu juuri metadatan avulla. Myös monet niin sanotut tutkimustiedonhallintajärjestelmät keskittyvät tarjoamaan tutkimustiedolle kattavan ja toimivan metadatarakenteen, ja siten tiedon hallinta tapahtuu pääosin metadatan hallinnan kautta. Metadata tarjoaa myös työkalut tutkimustiedon käsittelyyn ja auttaa tutkimustiedon sisällön ymmärtämisessä.

2.2.1 Metadatan rooli ja merkitys

Kuten aikaisemmin todettiin, metadatalla on tärkeä rooli tutkimustiedon kuvaamisessa. Metadatan keräämisen ja tallennuksen kautta voidaan parantaa tutkimustiedon käytettävyyttä. Marshallin (1998, 162) mukaan metadatan kerääminen ja ylläpito on yksi organisaation avaintoiminnoista. Metadatan avulla organisaatio pystyy kuvailemaan ja hallinnoimaan sen hallussa olevia tietokokoelmia, ja lisäksi metadata mahdollistaa myös tiedon jakamisen siitä kiinnostuneille tahoille. Metadatan merkitys on kasvanut entisestään sen myötä kun tutkimusta tekevät organisaatiot ovat levittäytyneet maantieteellisesti laajemmalle alueelle, jolloin tiedon jakaminen tapahtuu pitkälti Internetin välityksellä. (Marshall 1998, 162) Tutkijat eivät välttämättä ole suoranaisessa yhteydessä toisiinsa vaan erilaisten tutkimusaineistojen etsiminen tapahtuu esimerkiksi hakukoneiden avulla. Tällöin tulee tietoa kuvailevan metadatan olla mahdolli-

simman informoivaa, jotta kyseinen tieto löydettäisiin, ja jotta voitaisiin myös varmistua tiedon laadusta ja sopivuudesta.

Toivosen, Salmenkiven ja Verkamon (2004) mukaan metadatan avulla voidaan lisäksi mahdollistaa se, että tutkimusaineistoa voidaan hyödyntää myös tulevaisuudessa. Mikäli tutkimusaineisto kuvataan riittävän tarkasti tutkimusmenetelmiä myöten, voidaan sen tulokset ymmärtää vielä silloinkin kun ketään tutkimuksen suorittaneista tutkijoista ei voida enää konsultoida tutkimukseen liittyvistä asioista. Näin tarkan metadatan avulla tutkimus on niin ikään toistettavissa, jolloin myös sen kautta saatujen tulosten paikkansapitävyys voidaan varmistaa. (Toivonen, Salmenkivi & Verkamo 2004)

Bose (2002, 15) ottaa metadatan merkityksestä puhuessaan puolestaan kantaa tutkimustiedon alkuperän selvittämiseen. Tutkijoiden on usein pystyttävä selvittämään mistä lähteistä jonkun tutkimuksen tiedot on johdettu ja edelleen mistä nämä tiedot on puolestaan saatu jne. Näin ollen jonkun yksittäisen tutkimustiedon alkulähteen ja nykyisen muodon välillä voidaan nähdä olevan ketju tai virta (data flow), joka muodostuu kaikista niistä pisteistä, joissa tietoa muokataan. Alkulähteen ja nykyisen muodon väliin saattaa siis lukeutua useita muokkauspisteitä (muut tutkijat), joissa alkuperäistä tietoa prosessoidaan muistuttamaan enemmän tiedon nykyistä muotoa. Kaikkien tietoon tehtyjen muunnosten selvittäminen ja sitä kautta tiedon alkulähteen selvittäminen auttaa tutkijoita arvioimaan lähdetiedon laatua ja käytettävyyttä. Tiedolle tehtyjen prosessointivaiheiden tunteminen ja niistä saatu metadata mahdollistaa myös alkuperäiselle datalle tehtyjen muunnosten toistettavuuden. Näin ollen voidaan varmistaa eri tutkijoiden tekemien muunnosten oikeellisuus. (Bose 2002, 15–16)

2.2.2 Metadatan tallennus ja tasot

Tutkimustietoon liittyvää metadataa kerätään siitä hetkestä lähtien jolloin tietty datajoukko hyväksytään osaksi tietokokoelmaa ja metadatan kerääminen jat-

kuu aina siihen asti kun kyseinen datajoukko voidaan arkistoida (Beedham, 2004). Beedham (2004) jatkaa, että tutkimustietoon liittyvää metadataa tallennetaan yleensä useammalla tasolla. Seuraava metadatatasojen jaottelu yhdistelee Beedhamin (2004) ja Ryssevikin (2001) esittämiä tasoja. Ryssevikin (2001) esittämät metadatatasot ovat käytössä DDI (Data Documentation Initiative) -spesifikaatiossa. DDI on XML (Extensible Markup Language) -pohjainen metadatakieli ja sen on suunniteltu palvelemaan yhteiskuntatieteellisen tutkimusaineiston dokumentointia. DDI-spesifikaatiossa dokumentaatiolle on määritelty erittäin tarkka rakenne, jossa on kuvattu kaikki ne tutkimuksiin liittyvät elementit, joista metadataa tulisi tallentaa. DDI:n mukainen dokumentaatio koostuu siis metadatasta, joka mahdollistaa aineiston tehokkaan ja tarkan hyödyntämisen. (Ryssevik 2001) Tarkempi kuvaus DDI-spesifikaatiosta löytyy luvusta 3.

1. **Tutkimustaso**. Tutkimustason metadatassa kuvaillaan tutkimuksen alkuperä, sisältö, laajuus ja metodologia. Tällä tasolla kuvataan siis muun muassa tutkimuksessa käytetyt tiedonhankintamenetelmät sekä ne analysointimenetelmät, joita kerättyyn tietoon on sovellettu.

2. **Tiedostotaso**. Tiedostotason metadataan tallennetaan jokaisen tutkimusaineistoon kuuluvan tiedoston ominaisuudet. Metadatassa kuvataan muun muassa tiedostojen fyysinen tallennusmuoto, niiden rakenne ja ulottuvuudet. Lisäksi voidaan tallentaa tietoa esimerkiksi tiedostoista puuttuvista tiedoista.

3. **Muuttujataso**. Muuttujatason metadata kuvailee tarkasti kaikki tiedostojen sisältämät muuttujat. Näitä ovat esimerkiksi kyselytutkimuksen yhteydessä jokaisen kyselylomakkeessa olevan kysymyksen tehtävä ja painoarvo. Siinä missä tiedostotason metadatan avulla voidaan tietokokoelma kuvailla fyysisesti, voidaan se muuttujatason metadatan avulla puolestaan kuvailla loogisesti.

4. **Hallinnollinen metadata**. Hallinnollinen metadata jaetaan edelleen kahteen osaan: ylläpitotietoon ja hallinnolliseen tietoon. Ylläpitotiedon avulla varmistetaan, että kaikki oleellinen informaatio tietokokoelmasta on tallennettu tulevaisuuden tutkijoita ja arkistohenkilökuntaa silmällä pitäen. Näin voidaan taata tutkimuksen pitkäikäisyys. Hallinnolliseen tiedon avulla puolestaan tallennetaan tilastollista informaatiota tietokokoelman käytöstä.

5. **Meta-metadata.** Meta-metadatalla tarkoitetaan tietoa metadatasta ja siihen tallennetaan tietoa itse metadatadokumentista, esimerkiksi sen tekijästä, käytetyistä standardeista sekä sen tilasta dokumentointiprosessin eri vaiheissa. Mikäli metadata luodaan käyttämällä hyväksi jotain paperimuodossa olevaa dokumenttia, niin meta-metadatassa kuvataan lisäksi tämä metadatan lähdedokumentti. Myös Birnholtz ja Bietz (2003, 340) ovat ottaneet kantaa meta-metadatan tärkeyteen. Heidän mielestään myös metadatan, siinä missä itse tutkimusaineistonkin, ymmärrys perustuu pitkälti tutkijayhteisön sisällä olevaan tietoon. Näin ollen on oleellista selittää mitä tutkimusaineistoa kuvailevalla metadatalla tarkoitetaan.

Yaun ja Hawkerin (2004, 387–388) mukaan metadataa voidaan tallentaa kolmella eri tavalla. Yksi tapa on upottaa metadata kuvattavaan dokumenttiin. Tämä voidaan tehdä esimerkiksi sijoittamalla metadata erillisten <META> -elementtien sisään HTML-, XML- ja SGML-merkatuissa dokumenteissa. Sisäinen metadata tekee dokumenttien etsimisestä ja metadatan lukemisesta kuitenkin vaikeaa, koska metadata ja kuvattavan dokumentin sisältö on sekoitettu keskenään. Toinen tapa tallentaa metadataa on sijoittaa se kuvattavasta dokumentista erillään olevaan ulkoiseen metadatadokumenttiin, josta tehdään viittauksia kuvattavaan dokumenttiin. Ulkoisen metadatan käyttö helpottaa sen koneellista prosessointia, jolloin esimerkiksi metadatan muokkaus ja dokumenttien etsiminen helpottuu. Yksittäisten metadatadokumenttien käyttö on helppo tapa tallentaa metadataa, mutta siinä metadata yhtenäinen hallinta ei

ole mahdollista. (Yau & Hawker 2004, 387–388) Kolmantena metadatan tallennustapana Yau ja Hawker (2004, 387–388) pitävät sen sijoittamista erilliseen metadatatietokantaan, joka sisältää kaiken metadatan sekä viittaukset kuvattaviin dokumentteihin. Metadatan tallennus tietokantaan parantaa sen hallittavuutta, mutta vaatii toisaalta enemmän rahallisia investointeja sekä henkilökunnan sitouttamista ainakin tietokannan käyttöönottovaiheessa. (Yau & Hawker 2004, 387–388)

Marshall (1998, 162–163) puolestaan erottelee automaattisesti generoidun ja ihmisen synnyttämän metadatan. Automaattisesti generoitu metadata muodostuu tietokoneohjelmien luomista yhteenvedoista, indekseistä sekä muista dokumentteja kuvailevista tekijöistä. Marshall (1998, 162–163) pitää ihmisen synnyttämää metadataa kuitenkin tärkeämpänä, koska sen avulla voidaan dokumentit kuvata huomattavasti tarkemmin. Ihmisen synnyttämä metadata täydentää automaattisesti generoitua ja sen avulla pystytään myös arvioimaan eri tietokokonaisuuksien tarkoitusta, laatua ja painoarvoa. (Marshall 1998, 162–163)

2.2.3 Metadataan liittyvät vaikeudet

Voidaan varmasti sanoa, että metadatan käyttö on tutkimusorganisaatiolle arvokasta ja hyödyllistä. Valitettavasti sen luominen ja ylläpitäminen on useimmissa tapauksissa organisaatiolle kuitenkin vaikeaa ja kallista (Marshall 1998, 163). Ensinnäkin toimivan metadatarakenteen löytäminen on vaikeaa. Siitä huolimatta, että on olemassa useita erilaisiin käyttöyhteyksiin tarkoitettuja pitkälle kehitettyjä ja standardoituja metadatarakenteita, joudutaan erilaisten tietokokoelmien yhteydessä usein kuitenkin suorittamaan eri standardien yhdistelyä ja muokkaamista. Standardien sopimattomuus ja sitä kautta tarve tehdä niihin muutoksia johtuu usein kyseisen tietokokoelman erikoislaatuisesta luonteesta tai käyttötavasta. Ongelmia metadatarakenteen valinnassa ja luomisessa aiheuttaa myös kuvattavan aineiston moninaisuus. Saman rakenteen pitäisi pystyä

esimerkiksi kuvaamaan sekä fyysisessä muodossa olevaa dataa, kuten paperit ja videonauhat että digitaalisessa muodossa olevaa dataa. (Marshall 1998, 163)

Kun tietokokoelmaa kuvaavalle metadatalle on saatu rakenne, joko ottamalla jokin standardi käyttöön tai muokkaamalla jotain standardia omiin tarkoituksiin sopivaksi, joudutaan usein vaikeuksiin kun valittujen arvojen kuvaamista pannaan käytäntöön. Metadatan syöttäminen valitun metadatarakenteen mukaisiksi dokumenteiksi vaatii organisaatiolta huomattavia investointeja esimerkiksi hankittavien järjestelmien tai palkatun työvoiman muodossa. Usein metadatan koodaaminen hoidetaankin pikku hiljaa pitemmän ajan kuluessa ja siihen tarvittava työpanos jaetaan useiden työyksiköiden ja työntekijöiden tehtäväksi. Tämä puolestaan saattaa johtaa ongelmiin metadatan johdonmukaisuuden ja yhtenäisyyden kanssa. Myös metadatan päivittäminen tietokokoelmissa tapahtuvien muutosten mukaisesti aiheuttaa organisaatiolle lisätyötä. (Marshall 1998, 163)

Metadatan merkitys ja tärkeys pitäisi myös selvittää organisaation tutkijoille. Usein tutkijat eivät ymmärrä metadatan avulla saavutettavia hyötyjä ja näin ollen eivät näe mitään syytä investoida aikaansa metadatan luomiseen. Päinvastoin, metadatan luominen nähdään tällöin ylimääräisenä rasitteena, joka häiritsee tutkijoiden varsinaisen työn tekemistä. Jotta metadatan arvo voitaisiin ymmärtää, pitää tutkijoiden käytössä olla asianmukaiset työkalut, joiden avulla metadatan syöttäminen ja hyödyntäminen on mahdollisimman tehokasta ja helppoa. (Thomson, Adams, Cowley & Walker 2003, 32)

Ongelmia aiheuttaa myös metadatan ymmärrettävyys. Metadata tulisi tallentaa sellaiseen muotoon, että ihmiset pystyisivät suhteellisen helposti tulkitsemaan metadatan sisältöä ja että metadata olisi myös koneellisesti luettavissa. XML (Extensible Markup Language) on osaltaan tuonut helpotusta tähän ongelmaan, koska sitä pidetään yleisesti suhteellisen helposti ymmärrettävänä metadatakielenä. (Yang, Kafatos & Wang 2002, 53; W3C 2004a)

3 DATA DOCUMENTATION INITIATIVE

Tässä luvussa esitellään yhteiskuntatieteellisen tutkimusaineiston dokumentointia varten kehitelty metadataspesifikaatio, joka kantaa nimeä Data Documentation Initiative (DDI). Aluksi spesifikaatiota käsitellään yleisesti kertoen esimerkiksi spesifikaation kehittäjien tavoitteista sekä spesifikaation merkityksestä sitä käyttävälle organisaatiolle (3.1). Tämän jälkeen tarkastellaan lyhyesti DDI-spesifikaation historiaa ja hallinnointia (3.2), sekä tutustutaan lyhyesti spesifikaation käyttämään XML (Extensible Markup Language) -kieleen (3.3). Kohdassa 3.4 esitellään puolestaan DDI-spesifikaation viisi osa-aluetta, joiden sisältämien kohtien avulla dokumentointi toteutetaan. Luvun lopuksi esitellään vielä spesifikaatiota koskevia haasteita (3.5) sekä tutustutaan DDI-dokumentaation käyttämiseen esittelemällä kaksi projektia, joissa aineiston dokumentointi on toteutettu spesifikaation mukaisesti (3.6).

3.1 Data Documentation Initiative yleisesti

The Data Documentation Initiative (DDI) on hanke, jonka avulla pyritään kehittämään kansainvälinen XML (Extensible Markup Language) -pohjainen standardi yhteiskuntatieteellisten tutkimusten ja niissä syntyvän tutkimusaineiston dokumentointia varten. DDI-spesifikaation avulla tuotetun dokumentaation tarkoituksena on mahdollistaa tutkimusaineiston tehokas ja kokonaisvaltainen hyödyntäminen. Spesifikaation avulla pystytään vaikuttamaan tietokokoelmia kuvaavan dokumentaation sisältöön, esitystapaan, siirtämiseen sekä säilytykseen. DDI-spesifikaation on tarkoitus korvata olemassa olevat vanhentuneet metadatastandardit ja lisäksi spesifikaation kehittäjät pyrkivät siihen, että siitä tulisi perusta yhteiskuntatieteellisten tietokokoelmien kokoamiselle, jakelulle ja käytölle. Koska DDI ei ole vielä saavuttanut yleisesti hyväksytyn standardin

asemaa, käytetään tässä tutkielmassa standardin sijasta termiä spesifikaatio. (Data Documentation Initiative 2005a)

Kuten aikaisemmassa tutkimustiedon jakamista käsittelevässä luvussa todettiin, tieteellisen tutkimuksen yhteydessä kerättävän tiedon jakaminen tutkijoiden kesken on välttämätöntä tieteen tekemiselle (Birnholtzin & Bietzin 2003, 339). Tämä tosiasia koskee luonnollisesti myös yhteiskuntatieteellisessä tutkimuksessa syntyvää aineistoa. Muun muassa Ryssevik ja Musgrave (2001, 165) toteavat, että yhteiskuntatieteellinen tutkimusaineisto voidaan saattaa muiden tutkijoiden käytettäväksi vain metadatan avulla. Metadata toimii siis siltana tiedon tuottajien ja käyttäjien välillä tarjoten informaatiota, jota tarvitaan tiedon etsimiseen, ymmärtämiseen ja uudelleenanalysointiin. Ilman metadataa tietokokoelmat ovat niiden käyttäjille enemmän tai vähemmän merkityksettömiä joukkoja numeroita ja kirjaimia. (Norwegian Social Science Data Services 1999, 23)

Myös DDI-spesifikaation kehittäjät pitävät elintärkeänä sellaisen dokumentaation olemassaoloa, joka mahdollistaa tietokokoelmien sisältämän tiedon täyden ymmärtämisen ja uudelleenanalysoinnin ilman, että tiedon alkuperäistä kerääjää/tuottajaa tarvitsee konsultoida. Yhteiskuntatieteellisiä tietokokoelmia ylläpitävissä arkistoissa ollaan myös yleisesti sitä mieltä, että olemassa olevat dokumentaatiomuodot eivät ole riittäviä edellä mainitun kaltaisen dokumentaatiotason saavuttamiseen. (Data Documentation Initiative 2005a) Olemassa oleva dokumentaatio on myös usein teknisesti vanhettunutta. Tällä tarkoitetaan pääasiallisesti sitä, ettei käytetty dokumentaatiokieli ole rakenteista, ja näin ollen se ei siis myöskään ole tietokoneohjelmien ymmärrettävissä. Muun muassa semanttista webiä silmällä pitäen rakenteisen kielen, kuten XML:n käyttö dokumentoinnissa on erittäin tärkeää. (Data Documentation Initiative 2005a)

Tarvetta nykyaikaiset vaatimukset täyttävälle- metadatastandardille on siis olemassa. Ryssevikin (2001) mukaan tarvetta lisää entisestään yhteiskuntatie-

teellisten arkistojen hallussa olevan aineiston moninaisuus ja siihen kohdistuva runsas kysyntä. Moninaisuudella tarkoitetaan sitä, että arkistot sisältävät usein dataa julkiselta sektorilta (tilastokeskukset), kaupalliselta sektorilta (mielipidemittaukset) sekä akateemiselta sektorilta (akateeminen tutkimus). Runsaalla kysynnällä puolestaan tarkoitetaan sitä, että arkistojen hallussa oleviin tietokokoelmiin kohdistuu jatkuvasti pyyntöjä akateemisten tutkijoiden lisäksi myös julkiselta ja kaupalliselta sektorilta. Moninaisuus ja runsas kysyntä aiheuttavat metadatastandardiin kohdistuvien vaatimusten tiukentumista, koska tällöin 1) arkistoidun tietokokoelman käyttäjät ovat harvoin osallistuneet sen tuottamiseen, 2) arkistoitua dataa käytetään jatkuvasti sellaisiin käyttötarkoituksiin, joita sen tuottajat eivät ole ottaneet huomioon (uudelleenanalysointi), 3) arkistoitua dataa käytetään useita vuosia sen tuottamisen jälkeen ja 4) akateemiset käyttäjät vertailevat ja yhdistelevät usein dataa monista eri lähteistä. (Ryssevik 2001) DDI-hankkeen tarkoituksena onkin vastata juuri näihin tarpeisiin. DDI-spesifikaation mukaisen dokumentaation avulla yhteiskuntatieteellistä tutkimusta tekevät tutkijat voivat kirjata selkeästi tutkimuksiinsa liittyvän empiirisen datan kaikki keskeiset ominaispiirteet. Metadatan kirjaamisen jälkeen nämä ominaispiirteet välittyvät varsinaisen tutkimusdatan mukana kaikille niille tahoille, jotka sitä käyttävät. DDI-spesifikaation mukainen metadata tarjoaa siis välttämätöntä tietoa varsinaisesta tutkimusdatasta sekä tutkimuksesta, jossa se kerättiin. (Data Documentation Initiative 2005a)

3.2 Historia ja hallinnointi

DDI-hanke aloitettiin vuonna 1995, jolloin perustettiin asiantuntijoista koostuva DDI-komitea. Komitean toimintaa ohjaa ICPSR (Inter-University Consortium for Political and Social Research) -niminen organisaatio, ja sen tavoitteena on kehittää uusi metadatastandardi, joka voitaisiin hyväksyä maailmanlaajuisesti yhteiskuntatieteellistä tutkimusta tekevässä yhteisössä. Tarkoituksena on myös

korvata käytössä olevat, vanhentuneet metadatastandardit modernilla ja Internet-kelpoisella standardilla. Ensimmäiset luonnokset DDI-spesifikaatiosta pohjautuivat SGML (Standard Generalized Markup Language) -kieleen, mutta vuonna 1997 siirryttiin käyttämään vuotta aikaisemmin julkaistua XML-kieltä. DDI-spesifikaation beta-testaus suoritettiin maalis- ja marraskuun välisenä aikana vuonna 1999. Testaukseen osallistui yhteensä 13 organisaatiota sekä Euroopasta että Pohjois-Amerikasta. Beta-testauksen tarkoituksena oli muun muassa kokeilla DDI-spesifikaation mukaisen metadatan tuottamista erilaiselle tutkimusdatalle, kehittää ohjelmia, joilla metadatan tuottamista voidaan helpottaa sekä vertailla DDI-spesifikaatiota olemassa olevien metadatastandardien kanssa. Testaajilta saatujen raporttien pohjalta DDI-spesifikaatioon tehtiin tarvittavia muutoksia, joilla pyrittiin muun muassa korjaamaan havaittuja teknisiä ongelmia. Ensimmäinen virallinen versio DDI-spesifikaatiosta julkaistiin maaliskuussa 2000. (Data Documentation Initiative 2005a; Norwegian Social Science Data Services 1999)

Vuoteen 2003 asti DDI-komitea julkaisi uuden version spesifikaatiosta muutaman kuukauden välin. Jokainen uusi versio sisälsi pieniä parannuksia edeltävään versioon. DDI-komitea on teettänyt spesifikaatiolle myös muutamia laajempia ulkoisia arviointeja, joissa esiin nousseiden kehitysideoiden pohjalta spesifikaatioon on tehty joitain suurempia muutoksia. Uusin DDI-spesifikaation virallinen versio (2.1) julkaistiin elokuussa 2005, ja se sisälsi vain pieniä muutoksia aikaisemmin käytössä olevaan, vuonna 2003 julkaistuun versioon (2.0). (Data Documentation Initiative 2005a)

DDI-hankkeen hallinnointirakenteessa on tapahtunut muutoksia vuoden 1995 jälkeen. Hankkeessa mukana olevien organisaatioiden määrän kasvaessa ja spesifikaation laajenemisesta johtuvien kustannusten lisääntyessä perustettiin DDI:n hallinnointia varten helmikuussa 2003 DDI-allianssi, joka siis korvasi DDI-komitean. Allianssi jakautuu kahteen eri komiteaan: valmistelukomiteaan (Steering Committee) ja asiantuntijakomiteaan (Expert Committee). Valmistelu-

komiteaan lukeutuu tällä hetkellä kahdeksan henkilöä, jotka tulevat allianssiin kuuluvista organisaatioista. Valmistelukomitean tehtävänä on tehdä DDI-spesifikaatiota koskevia suurempia päätöksiä, jotka ohjaavat spesifikaation kehittymistä haluttuun suuntaan. Asiantuntijakomitea muodostuu puolestaan nimensä mukaisesti eri alojen asiantuntijoista. Komiteaan lukeutuu muun muassa XML-osaajia, yhteiskuntatieteellistä tutkimusta tekeviä tutkijoita sekä tiedon arkistoinnista vastaavia henkilöitä. Asiantuntijakomitean tehtävä on tarjota valmistelukomitean sekä koko DDI-allianssin käyttöön tietoa, josta olisi hyötyä spesifikaatiota kehittäessä. Itse allianssiin kuuluu kymmeniä organisaatioita Pohjois-Amerikasta ja Euroopasta. Kaikki allianssiin kuuluvat organisaatiot pääsevät vaikuttamaan DDI-spesifikaation kehitykseen. (Data Documentation Initiative 2005a; Norwegian Social Science Data Services 1999)

3.3 DDI-spesifikaatio ja XML

Jollei tekstissä toisin mainita tämän kohdan tarkastelu perustuu W3C:n (2004a, 2004b) ylläpitämissä XML ja XML-skeema -spesifikaatioissa esitettyihin tietoihin sekä Salmisen (2005a) kirjoittamaan suomenkieliseen XML-opetusmateriaaliin.

DDI-spesifikaation mukainen dokumentaatio on toteutettu XML-kielellä. XML on dokumenttirakenteiden määrittely- ja esitystapakieli, joka on tarkoitettu rakenteisen informaation tallennukseen ja jakeluun verkossa. XML-kielen avulla tallennettu tieto välitetään ohjelmistosovellukselta toiselle rakenteisten XML-dokumenttien avulla. Jokaisella XML-dokumentilla on sekä fyysinen että loogi-nen rakenne. Fyysinen rakenne koostuu yksiköistä, joita kutsutaan entiteeteiksi. Kaikilla XML-dokumenteilla on yksi tekstimuotoinen entiteetti, jota kutsutaan dokumenttientiteetiksi (document entity) tai juurientiteetiksi (root entity). Tämän lisäksi XML-dokumentit voivat sisältää myös muita entiteettejä, joiden muoto voi olla tekstin lisäksi esimerkiksi ääntä tai videokuvaa. Viittausten avul-

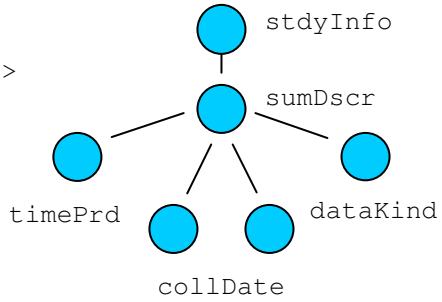
la yhteen liitettyjen entiteettien muodostama yksittäinen XML-dokumentti voi siis sisältää hyvin monimuotoista informaatiota.

XML-dokumentin loogisen rakenteen keskeisimpiä osia ovat elementit, jotka tuodaan esiin merkkauksen (markup) avulla. Elementtien merkkaus tapahtuu käyttämällä tunnisteita (tags), jolloin elementin sisältö kirjoitetaan alkutunnisteen (begin-tag) ja lopputunnisteen (end-tag) väliin (esimerkiksi <osoite>Keskikatu 13</osoite>). Merkkauksen avulla dokumenttien sisältämät asiat voidaan nostaa selvemmin esille erottamalla sisällön eri osat toisistaan. Merkkauksen ansiosta XML-dokumenttien sisällön osat ovat myös tietokoneohjelmien eroteltavissa, mikä puolestaan mahdollistaa esimerkiksi sen, että tutkimusaineiston metatietoon tehtäviä hakuja voidaan kohdistaan tarkasti tiettyihin dokumentaation osiin, kuten tutkimusmetodia koskeviin tietoihin.

Yksittäinen XML-dokumentti voi sisältää lukemattoman määrän elementtejä, joista kukin vastaa tietyn dokumenttiin kuuluvan tietosisällön tallentamisesta. Lisäksi jokainen elementti voi sisältää attribuutteja, joilla elementin ominaisuuksia voidaan kuvailla tarkemmin tai tuoda esiin elementin sisältöön liittyvää ylimääräistä tietoa. XML-dokumentin elementit sisältävät usein toisia elementtejä, joita kutsutaan lapsielementeiksi. Tällöin sisäkkäiset elementit muodostavat hierarkkisen puurakenteen. Seuraavalla sivulla olevassa kuviossa 1 annetaan esimerkki XML-dokumentin merkkaustavasta.

Kuviossa 1 *stdyInfo* -elementillä on lapsielementti *sumDscr*, jolla puolestaan on lapsielementit *timePrd*, *collDate*, ja *dataKind*. Esimerkistä voidaan huomata kuinka elementin sisältämät lapsielementit sekä varsinainen tekstisisältö on kirjoitettu kyseisen elementin alku- ja lopputunnisteen väliin. Esimerkistä ilmenee myös se, kuinka attribuutit esitetään XML-dokumenteissa (*collDate* -elementin attribuutit *event* ja *date*). Lisäksi elementtien muodostama puurakenne on esitetty graafisesti kuvion oikeassa reunassa.

```
<stdyInfo>
  <sumDscr>
    <timePrd>
      1. luokan kevätlukukauden
      loppu vuosina 2001-2004
    </timePrd>
    <collDate event="start" date="2001-05-10">
      10. toukokuuta 2001
    </collDate>
    <collDate event="end" date="2004-06-16">
      16. kesäkuuta 2004
    </collDate>
    <dataKind>
      Testitulos
    </dataKind>
  </sumDscr>
</stdyInfo>
```



KUVIO 1. Esimerkki XML-dokumentin merkkaustavasta sekä elementtien muodostamasta puurakenteesta.

Rakenteisuutensa vuoksi XML soveltuu erinomaisesti DDI-spesifikaation mukaisen dokumentaation tuottamiseen. Kun tutkimustietoon liittyvää metadataa tallennetaan XML-kielen avulla, voidaan jokainen yksittäinen metatiedon osa kuvailla oman elementin tai elementtiryhmän sekä attribuuttien avulla.

XML-dokumentissa sallittujen elementtien ja attribuuttien nimet sekä sen looginen rakenne on mahdollista määritellä erillisessä dokumenttityypin määrittelyssä (Document Type Definition, DTD). Loogisen rakenteen lisäksi DTD-määrityksen avulla voidaan määritellä myös XML-dokumentin fyysinen rakenne esittämällä minkälaisia entiteettejä siinä saa esiintyä. Loogisen ja fyysisen rakenteen lisäksi määrityksen avulla voidaan ottaa kantaa myös siihen minkälaista sisältöä XML-dokumentissa esiintyvillä elementeillä saa olla (toisia elementtejä vai tekstiä).

Myös DDI-spesifikaatiossa käytetään DTD-määritystä, joka on tallennettu erilliseen DTD-dokumenttiin. Itse asiassa DDI-spesifikaatio konkretisoituu juuri DTD-määrityksen muodossa. Spesifikaatiota kehittävän ICPSR:n Internet-sivuilta ilmaiseksi ladattavissa olevan DTD-dokumentin avulla voidaan tarkistaa onko toteutettu DDI-dokumentaatio spesifikaation mukaista. Tätä XML-dokumenttien oikeellisuuden tarkastamista suhteessa DTD:hen kutsutaan vali-

doinniksi, ja se voidaan suorittaa lukuisilla XML-kielen editointiin tarkoitetuilla ohjelmilla. Validoinnissa ohjelma tarkistaa noudattavatko dokumentissa esiintyvät elementtien ja attribuuttien nimet ja sisällöt annettua DTD-määritystä. Lisäksi validoinnissa tarkistetaan dokumentin loogisen ja fyysisen rakenteen oikeellisuus. Dokumentaatio ei läpäise validointia esimerkiksi silloin jos siinä esiintyvät elementit eivät ole sallittuja tai ne on kirjoitettu väärin, joidenkin elementtien sisältämät lapsielementit ovat väärässä järjestyksessä tai jos jokin pakollinen elementti tai attribuutti puuttuu kokonaan. XML-dokumentin määrittely voidaan DTD:n lisäksi suorittaa myös useilla muilla määrittelykielillä. Tiettyä määrittelykieltä tukevat ohjelmistot osaavat tarkistaa XML-dokumentin validiteetin suhteessa tähän määrittelykieleen. Eräs laajalti käytössä oleva ja myös DDI-allianssin tukema määrittelykieli on W3C:n XML Schema. Kyseinen kieli tarjoaa DTD:tä huomattavasti monipuolisemmat mahdollisuudet elementteihin ja attribuutteihin liittyvien rajoitteiden määrittelyyn. ICPSR:n Internet-sivuilta on ladattavissa myös XML Schema -muodossa oleva DDI-määritys.

Aivan DDI-spesifikaation kehityksen alkuaikoina käytettiin dokumentaation tuottamiseen ISO (International Organisation for Stardadization) -organisaation standardoimaa SGML (Standard Generalized Markup Language) -merkkauskieltä, mutta heti XML-standardin ilmestyessä DDI-allianssi siirtyi käyttämään XML-kieltä tutkimusaineiston kuvailuun. XML-kielen valinta näyttää osoittautuneen oikeaksi, sillä heti XML-standardin julkaisemisen jälkeen on sen suosio ja käyttö kasvanut räjähdysmäisesti kaikkialla maailmassa. XML-standardista on muovautunut yleismaailmallinen tekniikka tiedon tallentamiseen ja jakeluun verkossa. Suosionsa takia XML-kielelle ja sitä kautta myös DDI-spesifikaatiolle on olemassa erittäin laaja ohjelmistovalikoima. (Norwegian Social Science Data Services 1999) XML:n käyttö takaa myös sen, että DDI-spesifikaatio on laitteisto- ja ohjelmistoriippumaton. Toisin sanoen, XML:n ja sitä kautta DDI:n käyttöä ei ole sidottu mihinkään tiettyyn laitteistoon, käyttö-

järjestelmään tai ohjelmistoon, mikä puolestaan mahdollistaa spesifikaation vapaamman käytön. (Data Documentation Initiative 2005a; W3C 2004a)

3.4 DDI-spesifikaation osa-alueet

DDI-spesifikaation mukainen dokumentaatio jakautuu viiteen osa-alueeseen, jotka sisältävät yhteensä noin 300 eri kohtaa, joihin metatietoa voidaan tallentaa. Näistä kohdista 178 on elementtejä ja loput elementtien sisältämiä attribuutteja. Spesifikaation osa-alueita ovat: 1) Metadatadokumentin kuvailu (Document Description), 2) Tutkimuksen kuvailu (Study Description), 3) Tiedostojen kuvailu (Data Files Description), 4) Muuttujien kuvailu (Variables Description) ja 5) Muu tutkimukseen liittyvä materiaali (Other Study-Related Materials). Jokaisen DDI-dokumentaation osa-alueen tehtävänä on tallentaa siihen kuuluvaa metatietoa, jota kertyy yhteiskuntatieteellistä tutkimusta tehdessä. DDI-spesifikaation mukainen DTD-dokumentti määrittelee tarkasti mitä XML-kielen elementtejä ja attribuutteja kussakin dokumentaation kohdassa saa esiintyä ja minkälaista rakennetta näiden tulee noudattaa.

Seuraavissa alaluvuissa DDI-spesifikaation osa-alueet esitellään tarkemmin antaen tietoa muun muassa siitä, mitä kussakin kohdassa on tarkoitus dokumentoida ja minkälainen merkitys kohdilla on kokonaisdokumentaation kannalta. Lisäksi jokaisen osa-alueen kohdalla on annettu graafinen esitys siinä käytettävästä rakenteesta, sekä viimeistä osa-aluetta lukuun ottamatta myös esimerkki kunkin osa-alueen mukaisesta XML-dokumentaatiosta. Osa-alueiden tarkastelu perustuu DDI-allianssin julkaisemaan sanastokirjastoon (Tag Library; Data Documentation Initiative 2005b), jossa on selitetty jokaisen DDI-spesifikaation osa-alueen sekä niihin liittyvien elementtien ja attribuuttien tarkoitus ja käyttötapa. (Data Documentation Initiative 2005a; Ryssevik 2001)

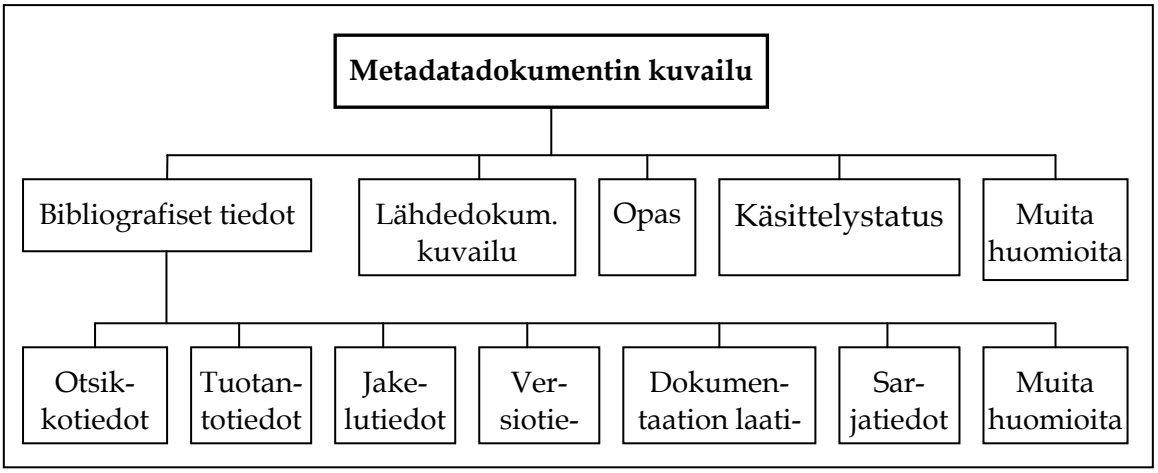
Liitteenä 1 olevassa luettelossa on esitelty DDI-spesifikaation mukaisessa dokumentaatiossa käytettävät elementit sekä niihin mahdollisesti liittyvät lapsielementit ja attribuutit. Lisäksi luettelosta voidaan myös nähdä puurakenne, jota kunkin DDI-dokumentaation osa-alueen tulee noudattaa. Liitteen 1 luettelossa esiintyvien ensimmäisen tason elementtien (1.0, 2.0, 3.0, 4.0 ja 5.0) numerointi vastaa seuraavien alalukujen numerointia, minkä tarkoituksena on helpottaa kunkin osa-alueen tarkempaa tarkastelua.

3.4.1 Metadatadokumentin kuvailu

DDI-spesifikaation mukainen dokumentaatio aloitetaan usein itse metadatadokumentin kuvailulla. Tämä kohta sisältää siis tietoa metadatasta, eli metametadataa. Jokaisen DDI-muotoisen metadatadokumentin kuvailu on tärkeää, jotta eri tutkimuksiin tai tutkimuskertoihin liittyvät dokumentit voitaisiin erottaa toisistaan.

Kuviossa 2 on annettu graafinen esitys metadatadokumentin kuvailussa käytettävästä rakenteesta. Kuviosta voidaan huomata, että osa-alueen tärkeimmän kokonaisuuden muodostaa metadatadokumentin bibliografinen kuvailu, joka koostuu metadatadokumentin otsikko-, tuotanto-, jakelu- ja versiotiedoista sekä tiedoista DDI-spesifikaation mukaisen dokumentaation laatineista henkilöistä ja useampien metadatadokumenttien muodostamista dokumenttisarjoista. Varsinaisen DDI-metadatadokumentin lisäksi tässä kohdassa on mahdollista kuvailla aikaisemmat, mahdollisesti paperimuodossa olevat metadatadokumentit, jotka ovat toimineet DDI-dokumentaation lähteinä. Lisäksi kuviosta 2 ilmenee, että osa-alue voi sisältää myös oppaan DDI-dokumentin käytölle, tiedot sen käsittelystatuksesta sekä muita metadatadokumenttia koskevia huomioita. Tarkempi kuvaus metadatadokumentin kuvailun sisällöstä löytyy liitteen 1 luettelosta, kohdasta 1.0, docDscr. (Data Documentation Initiative 2005b)

Kuviossa 3 on annettu esimerkki siitä, kuinka dokumentaatiota voidaan kuvailla DDI:n avulla. Kuviosta voidaan huomata, että siinä on käytetty vain osaa kaikista mahdollisista dokumentaation kuvailuun varatuista kohdista. Kuviossa on kuvailtu dokumentaation otsikko (*titl* ja *subTitl*), dokumentaation syöttäjä (*AuthEnty*), tuottaja ja tuotantopäivämäärä (*producer* ja *prodDate*), sekä dokumentaation liittyvä kontaktihenkilö (*contact*).



KUVIO 2. Metadatadokumentin kuvailussa käytettävä rakenne.

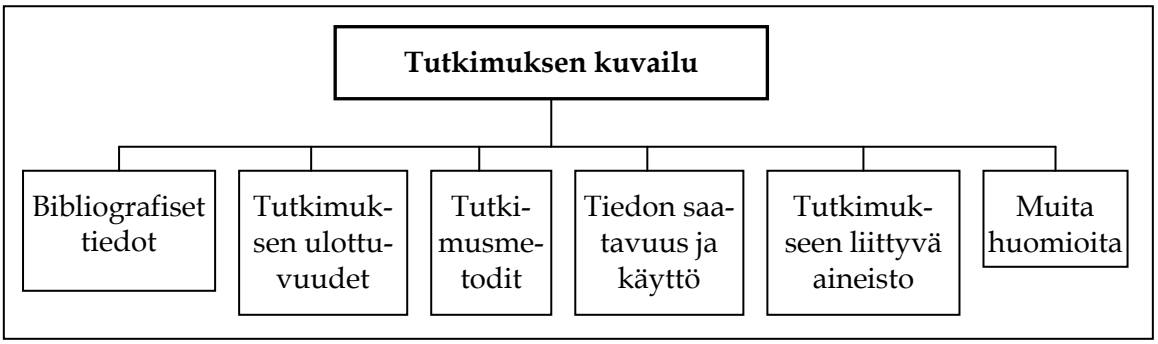
```
<codeBook version="2.0">
  <docDscr>
    <citation>
      <titlStmt>
        <titl>CS81sch1</titl>
        <subTitl>Lapsen kielen kehitys ja geneettinen dysleksiariski (LKK) -tutkimus</subTitl>
      </titlStmt>
      <rspStmt>
        <AuthEnty affiliation="LKK-tutkimus">Sinkkonen, Juha</AuthEnty>
      </rspStmt>
      <prodStmt>
        <producer affiliation="LKK-tutkimus" abbr="LKK">
          Jyväskylän yliopisto, Psykocenter, LKK-tutkimus
        </producer>
        <copyright>Copyright(c) Psykocenter, 2005</copyright>
        <prodDate date="2005-03-17"></prodDate>
      </prodStmt>
      <distStmt>
        <contact affiliation="LKK-tutkimus" email="jasinkko@cc.jyu.fi">
          Sinkkonen, Juha
        </contact>
      </distStmt>
    </citation>
  </docDscr>
```

KUVIO 3. Esimerkki DDI-spesifikaation mukaisesta dokumentaation kuvailusta.

3.4.2 Tutkimuksen kuvailu

DDI-spesifikaation mukaisen dokumentaation toinen osa-alue on tarkoitettu varsinaiseen tutkimukseen tai tutkimuskertaan liittyvän metadatan tallennukseen. Tutkimuksen kuvailu -kohdassa voidaan kuvailla myös useammista tutkimuksista tai tutkimuskerroista koostuvia tutkimussarjoja, jotka on kerätty pidemmän ajan kuluessa (Data Documentation Initiative 2005b). Kohta sisältää tiedot muun muassa tutkimusdatan keränneistä henkilöistä, tutkimuksen avainsanoista sekä käytetyistä tutkimusmetodeista (Data Documentation Initiative 2005b). Aikaisemmassa metadatan roolia ja merkitystä käsittelevässä luvussa (2.2.1) todettiin, että tutkimustietoon liittyvä metadata parantaa tutkimuksen käytettävyyttä. Marshall (1998, 162) korostaa, että juuri tutkijoiden itsensä kirjaama metadata (verrattuna koneellisesti luotuun) on ensisijaisen tärkeää tutkimukseen liittyvää taustatietoa tallentaessa ja välittäessä muille tutkijoille. Tästä syystä juuri tutkimukseen sekä sen suorittamiseen liittyvällä metadatalla on tärkeä rooli tutkimuksen käytettävyyden parantamisessa.

Kuviosta 4 voidaan huomata, että tutkimuksen kuvailu rakentuu kuudesta pääkohdasta. Ensimmäinen pääkohta on tutkimuksen bibliografinen kuvailu, joka sisältää samat alakohdat kuin metadatadokumentin kuvailun yhteydessä käytettävä bibliografinen kuvailu. Toisessa pääkohdassa kuvaillaan tutkimuksen sisällölliset, ajalliset ja maantieteelliset ulottuvuudet. Kolmannessa pääkohdassa kuvataan tiedon keräyksessä käytetyt metodit ja neljännessä kohdassa puolestaan keskitytään tutkimustiedon saatavuuden sekä käyttöön liittyvien ehtojen kuvailuun. Tutkimuksen kuvailun viidennessä pääkohdassa voidaan esitellä kuvailtavaan tutkimukseen läheisesti liittyvät materiaalit, muut tutkimukset, julkaisut sekä viittaukset. Kuudes ja viimeinen pääkohta on varattu lisätietojen tallennukselle. Jokainen pääkohta jakautuu lisäksi pienempiin kokonaisuuksiin, joiden tarkempi kuvaus löytyy liitteen 1 luettelosta, kohdasta 2.0, stdyDscr. (Data Documentation Initiative 2005b)



KUVIO 4. Tutkimuksen kuvailussa käytettävä rakenne.

```
<stdyDscr>
  <citation>
    <titlStmt>
      <titl>
        CS81sch1 - Group testing 1, November, Writing, Allu, Reading comprehension,
        Teacher evaluations
      </titl>
      <subTitl>Lapsen kielen kehitys ja geneettinen dysleksiariski (LKK) -tutkimus</subTitl>
    </titlStmt>
  </citation>
  <stdyInfo>
    <subject>
      <topcClas>Spelling</topcClas>
      <topcClas>Reading</topcClas>
    </subject>
    <sumDscr>
      <collDate event="start" date="2001-11-19"></collDate>
      <collDate event="end" date="2003-12-04"></collDate>
      <dataKind ID="DKSpel">Test</dataKind>
      <dataKind ID="DKRead">Test</dataKind>
    </sumDscr>
  </stdyInfo>
  <method>
    <dataColl>
      <collMode ID="Spel">
        1) Nonwords 1 (9 items: hoipa, oipus, ankula),
        2) Nonwords 2 (9 items: lirestemmä, märskynäivä, syönikeitväs)
      </collMode>
      <collMode ID="Read">
        1) Word recognition: ALLU, TL2 (2 min: matching word to picture),
        2) Sentence recognition: ALLU, TL4 (2 min: matching sentence to picture)
      </collMode>
    </dataColl>
  </method>
</stdyDscr>
```

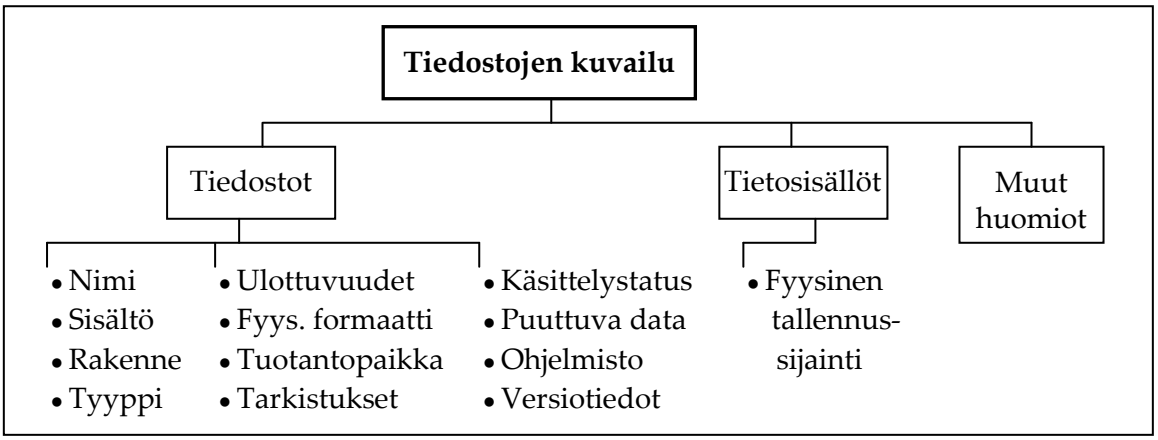
KUVIO 5. Esimerkki DDI-spesifikaation mukaisesta tutkimuksen kuvailusta.

Kuviossa 5 on annettu esimerkki DDI:n mukaisesta, XML-muotoisesta tutkimustietojen kuvailusta. Kuviosta voidaan huomata kuinka kukin metadatan osa tallennettaan sille tarkoitettujen tunnisteiden väliin. Esimerkissä tutkimustiedoista on kuvailtu tutkimuskerran otsikko *(titl, subTitl)*, aihepiiri *(topClas)*, tiedon keräyspäivämäärät *(collDate)*, tallennetut tietotyypit *(dataKind)* sekä tutkimusmetodina käytetty tiedonkeräysmittari *(collMode)*.

3.4.3 Tiedostojen kuvailu

Kolmas DDI-spesifikaation osa-alue muodostuu tiedostotason metadatasta, ja sen avulla kuvaillaan siis tutkimukseen tai tutkimuskertaan liittyvät tiedostot. Tällaisia tiedostoja ovat muun muassa tutkimusdataa sisältävät koontitiedostot (esimerkiksi SPSS-tiedostot) sekä tutkimuskertoja varten tehdyt testauslomakkeet, jotka sisältävät esimerkiksi haastattelussa esitettävät kysymykset ja ohjeet haastattelun läpiviennille.

Kuviossa 6 on esitelty DDI-spesifikaation mukaisessa tiedostojen kuvailussa käytettävä rakenne. Kuviosta ilmenee, että jokaisen tutkimukseen tai tutkimuskertaan liittyvän tiedoston kuvailun yhteydessä on mahdollista tallentaa metatietoa hyvin monipuolisesti. Dokumentaatiossa voidaan kuvailla muun muassa tiedostojen nimet ja tuotantopaikat, antaa kuvaus niiden tietosisällöistä ja rakenteista, sekä selvittää tiedostojen tuottamiseen käytetyt ohjelmistot ja tiedostoista mahdollisesti puuttuvat tiedot. Varsinaisten tutkimustiedostojen lisäksi tiedostojen kuvailu -kohdan avulla voidaan tallentaa tietoa myös niiden sisältämistä yksittäisistä tietosisällöistä sekä niiden sijainneista. Näin ollen tutkimukseen tai tutkimuskertaan liittyviä tiedostojen kuvailu voidaan viedä erittäin tarkalle tasolle pilkkomalla tiedosto pienempiin tietokokonaisuuksiin. Tarkempi kuvaus tiedostojen kuvailun sisällöstä löytyy liitteen 1 luettelosta, kohdasta 3.0, fileDscr.



KUVIO 6. Tiedostojen kuvailussa käytettävä rakenne.

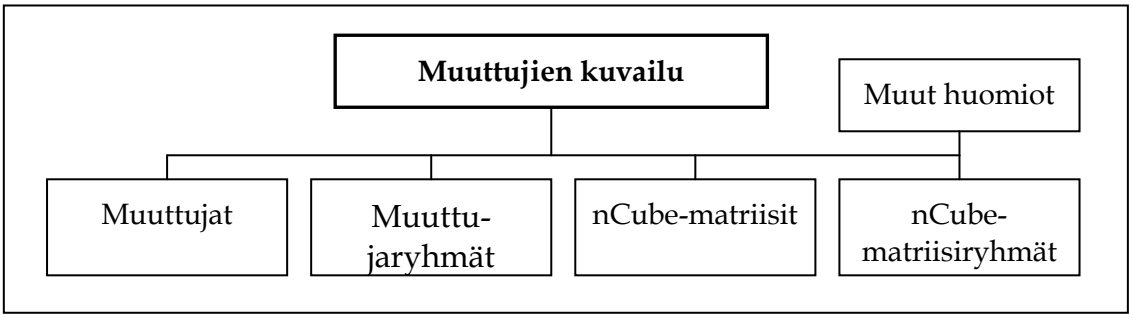
Kuviossa 7 on annettu esimerkki DDI-spesifikaation mukaisesta tiedostojen kuvailusta. Esimerkissä on kuvailtu kahden eri tiedoston tiedot, ja kummastakin tiedostosta on dokumentoitu tiedoston nimi (*fileName*), tietosisältö (*fileCont*), tyyppi (*fileType*), tuotantopaikka (*filePlac*) sekä tuottamiseen käytetty ohjelmisto (*software*). (Data Documentation Initiative 2005b)

```
<fileDscr ID="File1" URI="CS75ind3 - Individual 3, May, Phonology, Naming, Writing, Math,
  Common unit,Attributions.sav">
  <fileTxt>
    <fileName ID="CS75ind3">CS75ind3 - Individual 3, May, Phonology, Naming, Writing, Math,
      Common unit,Attributions.sav
    </fileName>
    <fileCont>Kerätty tutkimusdata</fileCont>
    <fileType>SPSS</fileType>
    <filePlac>Jyväskylän yliopisto, Psykocenter, LKK-tutkimus</filePlac>
    <software>SPSS (Statistical Package for the Social Sciences)</software>
  </fileTxt>
</fileDscr>
<fileDscr ID="File2" URI="May1lk.2004\May1lk.2004 yksilötutkimuksen testauslomake 75
  May  30042004.doc">
  <fileTxt>
    <fileName ID="May1lk.2004">May1lk.2004 yksilötutkimuksen testauslomake 75 May
      30042004.doc
    </fileName>
    <fileCont>Testauslomake</fileCont>
    <fileType>Word</fileType>
    <filePlac>Jyväskylän yliopisto, Psykocenter, LKK-tutkimus</filePlac>
    <software>Microsoft Word</software>
  </fileTxt>
</fileDscr>
```

KUVIO 7. Esimerkki DDI-spesifikaation mukaisesta tiedostojen kuvailusta.

3.4.4 Muuttujien kuvailu

DDI-dokumentaation neljäs osa-alue on tarkoitettu tutkimukseen tai tutkimuskertaan liittyvien muuttujien kuvailuun. Tässä kohdassa tarkastelu viedään siis tiedostotasoa tarkemmalle tasolle, ja siinä keskitytään kuvailemaan tutkimukseen liittyviä yksittäisiä muuttujia. Jokaiseen tutkimukseen tai tutkimuskertaan saattaa sisältyä satoja tai jopa tuhansia muuttujia, joista kukin on tarkoitettu jonkin pienen tietokokonaisuuden mittaamiseen ja tallentamiseen. Tällaisia tietokokonaisuuksia voi olla esimerkiksi haastatellun henkilön sukupuoli tai yksittäisen kysymyksen kautta saatu vastaus. Muuttujien sisältämät tiedot tallennetaan usein edellisessä kohdassa kuvailtuihin tiedostoihin, kuten SPSS-tiedostoihin. Muuttujien kuvailun yhteydessä jokaisesta muuttujasta voidaan tallentaa hyvin monipuolisesti tietoa, kuten esimerkiksi muuttujan tarkoitus, muuttujaan sisältyvä kysymys, siinä sallitut arvot sekä tietoja muuttujakategorioista ja muuttujien teknisestä formaatista. Tämän lisäksi joltain ominaisuuksiltaan tai aihepiiriltään toisiaan lähellä olevia muuttujia voidaan ryhmitellä muuttujaryhmiksi, joiden ominaisuuksia voidaan myös kuvailla. Kuviossa 8 on esitetty yksinkertainen graafinen kuvaus DDI-spesifikaation mukaisessa muuttujien kuvailussa käytettävästä rakenteesta.



KUVIO 8. Muuttujien kuvailussa käytettävä rakenne.

Kuviosta voidaan huomata, että muuttujien kuvailu jakaantuu viiteen eri pääalueeseen, joita ovat muuttujien ja muuttajaryhmien kuvailu, sekä nCube-

matriisien sekä niistä muodostuvien matriisiryhmien kuvailu. Viidentenä pääkohtana on muuttujien kuvailuun liittyvät muut huomiot.

Muuttujien ja muuttujaryhmien kuvailun lisäksi tässä kohdassa voidaan tutkimusaineistoa kuvailla siis myös nCube -matriisien sekä niistä muodostuvien matriisiryhmien avulla. (Data Documentation Initiative 2005b) Thomasin & Ryssevikin (2003) mukaan nCube -matriisien käyttö tiedon kuvailuun on hyödyllistä silloin kun tutkimusten tai tutkimuskertojen yhteydessä kerätty data on taulukkomuotoista koostedataa (aggregate data). Koostedata saadaan muokkaamalla kerättyä mikrodataa siten, että tietyt spesifit vaatimukset täyttyvät. Muokkaaminen voi koskea esimerkiksi tutkittujen tapausten määrää tai analysoinnissa käytettäviä muuttujia, ja sen avulla voidaan muun muassa luoda olemassa olevasta tutkimusryhmästä tietyt kriteerit täyttävä alitutkimusryhmä lähempää tarkastelua varten. Koostedata tallennetaan usein n-ulotteisiin tauluihin, joissa jokainen käytettävä muuttuja muodostaa oman, hierarkkisen, ulottuvuuden. Koostedataa sisältäviä tauluja käytetään ennen kaikkea tilastotieteellisen datan tallennukseen. DDI-spesifikaation versiosta 1.3 lähtien siinä on ollut mahdollista luoda tarkkoja kuvauksia moniulotteisista koostedataa sisältävistä tauluista nCube -matriisien avulla. (Thomas & Ryssevik 2003) Tarkempi kuvaus muuttujien kuvailun sisällöstä löytyy liitteen 1 luettelosta, kohdasta 4.0, dataDscr.

Kuviossa 9 on annettu esimerkki DDI-spesifikaation mukaisesta muuttujien kuvailusta. Esimerkissä muuttujatiedoista on kuvailtu muuttujan nimi *(name-attribuutti)*, muuttujan selite *(labl)*, muuttujaan liittyvä kysymys (qstnlit), testaajalle tarkoitetut kysymysohjeet *(ivuInstr)*, tiettyjen arvojen selitteet ja esiintymiskerrat *(catValu, labl ja freq-attribuutti)* sekä muuttujan tyyppi *(varFormat)*.

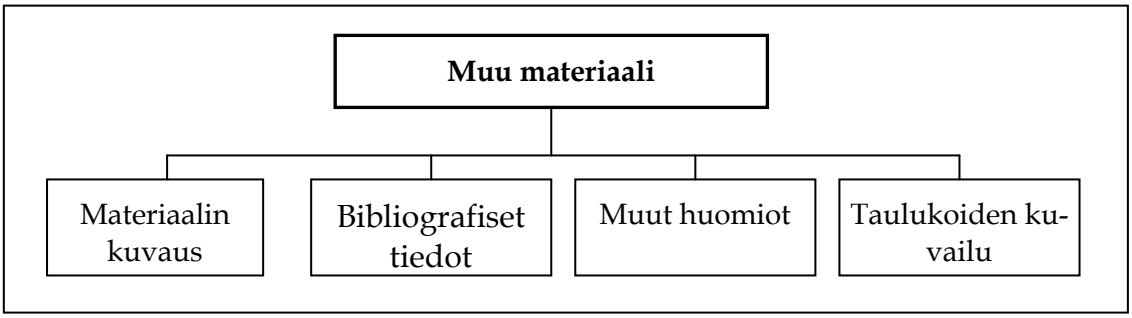
```
<var name="MAT75T11">
  <labl>math 1, numbers, 1st grade May, task 1</labl>
  <qstn>
    <qstnLit>
      Kuinka pitkälle osaat luetella lukuja. Aloita 1, 2, 3 ... Sanon, milloin voit lopettaa.
      Aloita nyt!
    </qstnLit>
    <ivuInstr>Pysäytä 31:een.</ivuInstr>
    <ivuInstr>
      Katkaise tehtävän suoritus, mikäli lapsi tekee yli 2 virhettä tai pysähtyy yli 10
      sekunninksi kesken suorituksen.
    </ivuInstr>
  </qstn>
  <catgry>
    <catValu>0</catValu>
    <labl>doesn't finish or over 2 errors</labl>
    <catStat type="freq">2</catStat>
  </catgry>
  <catgry>
    <catValu>1</catValu>
    <labl>1 or 2 errors</labl>
    <catStat type="freq">23</catStat>
  </catgry>
  <catgry>
    <catValu>2</catValu>
    <labl>correct</labl>
    <catStat type="freq">157</catStat>
  </catgry>
  <varFormat type="numeric" schema="other"/>
</var>
```

KUVIO 9. Esimerkki DDI-spesifikaation mukaisesta muuttujien kuvailusta.

3.4.5 Muu tutkimukseen liittyvä materiaali

DDI-spesifikaation viidennen osa-alueen yhteydessä on mahdollista kuvailla sellaisia tutkimukseen tai tutkimuskertaan liittyviä tietoa, joita ei ole kuvailtu dokumentaation aikaisempien osa-alueiden yhteydessä. Tutkimukseen liittyvän lisämateriaalin kuvailuun sisältyy materiaalista annetut kuvaukset, materiaalin bibliografiset tiedot sekä siihen liittyvät muut huomiot. Lisäksi DDI-spesifikaation viidennen osa-alueen avulla on mahdollista kuvailla taulukko-muodossa olevaa lisämateriaalia. Kuviossa 10 on esitetty graafinen kuvaus muuhun tutkimukseen liittyvän materiaalin kuvailussa käytetystä rakenteesta.

Tarkempi kuvaus kohdan sisällöstä löytyy liitteen 1 luettelosta, kohdasta 5.0, otherMat. (Data Documentation Initiative 2005b)



KUVIO 10. Rakenne muun tutkimukseen liittyvän materiaalin kuvailulle.

DDI-spesifikaation viisi osa-aluetta sisältävät yhteensä yli 300 erilaista kohtaa, johon metadataa voidaan tallentaa (Data Documentation Initiative 2005c). Kaikkia spesifikaation mahdollistamia kohtia ei suinkaan ole pakollista dokumentoida, eikä tämä yleensä ole edes suositeltavaa. Itse asiassa spesifikaation mukaisessa dokumentaatiossa ainoastaan dokumentoitavan tutkimuksen otsikon merkkaaminen on pakollista (Data Documentation Initiative 2005b). Huomioitavaa on myös se, että melkein kaikki DDI:n elementeistä ovat toistettavissa, mikä mahdollistaa sen, että niitä voidaan käyttää useaan kertaan. Tutkimukselle voidaan esimerkiksi mainita useampia rahoittajia, joiden erilaisia rooleja voidaan edelleen kuvata rooli-attribuutin avulla. (Data Documentation Initiative 2005b)

Elementtien ja attribuuttien laaja kirjo johtuu siitä, että DDI-spesifikaation laatijat ovat halunneet luoda mahdollisuudet hyvin erilaisten ja monimuotoisten tutkimuksien ja tutkimusaineistojen dokumentointiin. DDI-spesifikaation laatijat suosittelevatkin, että spesifikaation mukaisen dokumentaation toteutusta suunnittelevat organisaatiot tutustuvat aluksi DDI:n tarjoamiin dokumentointimahdollisuuksin. Tämän jälkeen organisaatioiden tulee tehdä päätös siitä, mitkä DDI-spesifikaation eri osa-alueiden sisältämistä kohdista tulee dokumentoida, jotta kaikki oleellinen metadata tulisi tallennettua kyseisen organisaation

tutkimuksia ja tutkimusaineistoa silmällä pitäen. Mikäli organisaation tutkimukset on jo dokumentoitu ja nykyinen metadata halutaan muuttaa DDI-spesifikaation mukaiseen muotoon, tulee organisaation tällöin harkita sitä, mihin DDI:n kohtiin olemassa olevan dokumentaation eri kohdat tallennetaan. Molemmissa tapauksissa juuri kyseiselle organisaatiolle tärkeiden dokumentointikohtien löytäminen on tärkeää, jotta dokumentaation toteuttaminen voisi jatkua muilta osin. (Data Documentation Initiative 2005c)

3.5 DDI-spesifikaation haasteet

Ryssevik (2001) esittelee artikkelissaan DDI-spesifikaatioon liittyviä rajoituksia ja ongelmia. Seuraavassa luettelossa esitetään aikaisemmin tiedossa olleita sekä vuonna 2001 tehdyssä DDI-hanketta koskevassa ulkoisessa arvioinnissa esiin tulleita haasteta. Yleisesti ottaen vuoden 2001 arvion tehneet arvioijat suhtautuivat DDI-allianssin työhön ja työn tuloksiin varsin positiivisesti, ja pitivät tärkeänä ennen kaikkea sitä, että DDI-spesifikaatio oli otettu nopeasti ja mieluisasti vastaan kohdeorganisaatioissa (Ryssevik 2001). Spesifikaation teknisestä laadusta sekä sen oikeista suuntaviivoista huolimatta arviossa haluttiin kuitenkin nostaa esille kaksi spesifikaatioon liittyvää ongelma-aluetta:

1) **DDI-spesifikaation painottuneisuus/rajoittuneisuus kyselytutkimusten dokumentointiin.** Spesifikaatio on alun perin suunniteltu kuvaamaan yhteiskuntatieteellisessä tutkimuksessa yleisimmin syntyvää tietoobjektia, joka on yksittäisessä kyselytutkimuksessa syntyvä tutkimustiedosto. Vaikka spesifikaatiosta tehdäänkin viittauksia myös muunlaiseen tutkimusdataan, on kaikki DDI:n tärkeimmät konseptit ja suurin osa sen logiikasta on johdettu juuri kyselytutkimusten pohjalta (Ryssevik 2001). Vaikka spesifikaation uusimmissa versioissa tätä ongelmaa on pyritty helpottamaan, ei sen käyttäminen kuitenkaan suju mutkattomasti monimutkaisempien tutkimustyyppien, kuten pitkittäistutkimusten yhtey-

dessä (Ryssevik 2001; Data Documentation Initiative 2005a). Jotta DDI-spesifikaation avulla voitaisiin saavuttaa varteenotettava tuki monimutkaisempien tutkimustyyppien dokumentoinnissa, täytyisi DDI:n alkuperäiseen arkkitehtuuriin tehdä luultavasti mittavia muutoksia. Kunnollista tukea ei voida siis saavuttaa alun perin kyselytutkimuksien dokumentointiin suunniteltua rakennetta venyttämällä. (Ryssevik 2001)

2) **Modulaarisuuden puute.** DDI-spesifikaatio pohjautuu ajatukseen, että se on digitaalinen vastine aikaisemmin paperimuodossa olleelle koodikirjalle (codebook) tai tutkimustiedon tietosanakirjalle (data dictionary), johon tallennettiin tutkimusta koskevaa metadataa. Vaikka DDI-spesifikaation koodikirja on jaettu viiteen eri osa-alueeseen, ei sitä kuitenkaan ole rakennettu modulaarisen arkkitehtuurin pohjalta. Modulaarisuus mahdollistaisi dokumentaatiopalasten valinnan sieltä täältä sekä niiden vapaan yhdistelyn dokumentaatiota suorittavan tahon haluamalla tavalla. Modulaarisuuden puute johtuu myös pitkälti DDI:n mukaisten XML-dokumenttien määrittelyyn valitusta DTD-määrittelykielestä. DTD ei salli dokumentaation pilkkomista pieniin ja uudelleen käytettäviin palasiin, vaan pakottaa dokumentaation noudattamaan yhtä järkälemäistä rakennetta. DTD-määrittelykielestä johtuvia modulaarisuusongelmia helpottaakseen DDI-allianssi on julkaissut spesifikaation mukaisia XML-dokumentteja koskevan määrittelyn myös luvussa 1.3 mainitulla XML Schema -määrittelykielellä. (Ryssevik 2001)

Seuraavaksi mainitut haasteet eivät tulleet esiin vuonna 2001 tehdyn arvioinnin yhteydessä vaan ovat olleet DDI-allianssin tiedossa jo ennen arvioinnin suorittamista.

3) **Kokoava lähestymistapa (bottom-up approach).** DDI-spesifikaatio on kehitetty kuvaamaan konkreettisia tutkimustiedostoja tai muita tilastollisen tutkimuksen tuotoksia ja voidaankin sanoa, että DDI:n päätehtävänä

on tarjota tutkimustiedostojen käyttäjille mahdollisimman paljon informaatiota niiden tuotantoprosessista. Tiedostojen ja niitä kuvaavan DDI-dokumentin välillä on siis yksi yhteen -suhde, eikä spesifikaatiossa näin ollen ole metodeja, joiden avulla voisi kuvailla yleisempiä tilastollisia käsitteistöjä, jotka soveltuisivat useampien tutkimusten kuvailuun. Kuitenkin esimerkiksi muuttujakuvausten kohdalla tulee usein vastaan tilanne, jossa eri tutkimusten ja saman tutkimuksen eri tutkimuskertojen yhteydessä käytetään samoja muuttujia, jolloin myös muuttujakuvaukset olisi tärkeää päästä tekemään yleisemmällä tasolla. DDI-spesifikaatiosta puuttuu keinot tämän tekemiseen ja siinä muuttujakuvaukset täytyykin tehdä joka tutkimuksen ja tutkimuskerran yhteydessä erikseen vaikka käytettävä muuttuja olisikin sama. (Ryssevik 2001)

4) **Laajennettavuuden puute.** Modulaarisuuden puuttumisen lisäksi toinen DTD-määrittelykielestä johtuva ongelma on laajennettavuuden puute. DTD:n rajoittuvuuksien takia DDI-spesifikaation mukaiseen kuvailuun ei voi lisätä omia laajennuksia niin että ne toimisivat yhdessä varsinaisen ydinspesifikaation kanssa. Toisin sanoen jotta dokumentaatio olisi spesifikaation mukaista, täytyy sen rakenne hyväksyä sellaisenaan ilman lisäyksiä. Tätä voidaan sinänsä pitää melko suurena heikkoutena sillä lähestulkoon jokaisessa tutkimusprojektissa on havaittavissa tiettyjä erityispiirteitä, jotka asettavat sen dokumentoinnille spesifejä tarpeita, jotka ovat ominaisia vain kyseisen tutkimuksen dokumentoinnissa. Tästä syystä mahdollisuus omien laajennusten tekemiseen tulisi ehdottomasti olla olemassa. (Ryssevik 2001)

5) **Suorituskykyongelmat.** Ryssevikin (2001) listaamien ongelmakohtien lisäksi Leighton (2002, 6) mainitsee artikkelissaan erään DDI-spesifikaation käyttöön liittyvän ongelman. Hänen mukaan DDI-dokumentaation käytön yhteydessä esiintyy usein suorituskykyyn liittyviä ongelmia. Tämä johtuu osaltaan spesifikaation massiivisesta raken-

> teesta, mikä usein johtaa siihen, että dokumentaatiota sisältävät XML-tiedostot paisuvat erittäin suurikokoisiksi. Tällä puolestaan voi olla se seuraus, että tiedostojen käsittely esimerkiksi hakujen yhteydessä on hidasta ja vie paljon koneen käyttömuistia. Toinen suorituskykyongelmia luova tekijä on DDI-spesifikaation mukaisten XML-dokumenttien sisältämä sisäisten viittausten runsas lukumäärä. Kustakin muuttujakuvauksesta on esimerkiksi viittaus siihen tiedostoon, johon kyseisen muuttujan tiedot on tallennettu sekä muuttujaryhmään, johon kyseinen muuttuja mahdollisesti kuuluu. Vastaavanlaisia viittauksia on spesifikaatiossa erittäin paljon, mikä vaatii DDI-dokumenttia käsittelevältä koneelta paljon prosessointitehoa. (Leighton 2002, 6)

DDI-spesifikaatiota kehittävä DDI-allianssi tekee aktiivisesti töitä edellä mainittujen ongelmien ratkaisemiseksi. Ryssevikin (2001) mukaan tällä hetkellä näyttää kuitenkin siltä, että ongelmien ratkaisemiseksi DDI-spesifikaatioon pitäisi tehdä melko suuria rakenteeseen ja toimintaperiaatteeseen liittyviä muutoksia. Eräs tällaisista muutoksia on hänen mukaan syntaksin ja semantiikan erottaminen toisistaan selkeämmin. Toisin sanoen DDI-spesifikaatiota kuvaava semanttinen malli tulisi tulevaisuudessa erottaa sen syntaksisesta esitysmuodosta (DTD-määrittely). Tulevaisuudessa DDI-spesifikaatiolla olisi siis yksi semanttinen malli, jolla voisi olla useampia syntaksisia esitysmuotoja. Muiden metadatastandardien tavoin DDI:n mukaista dokumentaatiota ei siis sidottaisi käyttämään jotain tiettyä syntaktista esitysmuotoa, vaan se voitaisiin valita tapauskohtaisesti olemassa olevien esitysmuotojen keskuudesta. Varteenotettavia syntaksisia esitysmuotoja ovat muun muassa RDF (Resource Description Framework) schema ja XML schema. (Ryssevik 2001)

3.6 Käytännön kokemuksia DDI-spesifikaatiosta - Tapaukset Counting California ja Nesstar

Tässä kohdassa käsitellään DDI-spesifikaation mukaisen dokumentaation käytöstä saatuja kokemuksia tutustumalla Counting California -projektiin, joka toteuttaa kokoamansa aineiston dokumentoinnin DDI-spesifikaatiota hyödyntäen. Lisäksi tutustutaan Nesstar-projektiin, joka kehittää spesifikaation mukaisen dokumentaation tuottamiseen tarkoitettuja työkaluja. Molemmat projektit toimivat myös yhteistyössä spesifikaatiota kehittävän DDI-allianssin kanssa.

3.6.1 Kalifornian osavaltion Counting California-projekti

Counting California -projekti on vuonna 2000 aloitettu hanke, jonka tavoitteena on parantaa Kalifornian osavaltion asukkaiden pääsyä julkisen hallinnon tuottamaan yhteiskuntatieteelliseen ja taloudellis-hallinnolliseen dataan. Projektin rahoittajana toimii Kalifornian digitaalinen kirjasto ja sen tarkoituksena on luoda järjestelmä, jonka käyttäjät voivat yksinkertaisen käyttöliittymän kautta selailla koko ajan kasvussa olevaa valtiollisten, osavaltiotasoisten ja paikallisten toimistojen keräämää julkiseen käyttöön tarkoitettua dataa. Järjestelmän avulla dataa voidaan myös koota yhteen ja yhdistellä useista eri lähteistä esimerkiksi jonkin otsikon tai maantieteellisen alueen perusteella. Lisäksi järjestelmä pyrkii vastaamaan ongelmaan, joka muodostuu jatkuvasti vaihtuvien ja päivittyvien teknologioiden ja tietoformaattien myötä. Counting California -järjestelmän avulla niin historiallista kuin ajankohtaistakin dataa voidaan tarkastella yhden järjestelmän avulla. (Counting California Project 2005)

Counting California -projekti tekee tiivistä yhteistyötä DDI-hankkeen kanssa osallistumalla DDI-standardin kehitykseen raportoimalla dokumentaation käytöstä sekä vastaan tulleista ongelmista ja niiden pohjalta heränneistä kehitystoiveista. Vastapainoksi Counting California -projekti saa tukea DDI-hankkeessa mukana olevilta organisaatioilta ja niissä työskenteleviltä asiantuntijoilta. DDI-

spesifikaation käyttökokemuksista kertovassa artikkelissaan Cruse, Einowski ja Stratford (2002, 11) toteavatkin, että tämän kaltainen spesifikaation käyttöä koskeva tuki on ensisijaisen tärkeää ja varsinkin projektin alkuvaiheissa tukea tarvittiin paljon. Asiantuntijatuen merkitystä kasvattaa erityisesti myös se seikka, ettei DDI-spesifikaation käytöstä ole saatavilla kovin monia konkreettisia esimerkkejä. (Cruse, Einowski & Stratford 2002, 11)

DDI-allianssin antamien suositusten mukaisesti Counting California-projekti aloitti DDI-spesifikaation käytön tutustumalla huolellisesti spesifikaation osa-alueisiin sekä niiden sisältämiin elementteihin ja attribuutteihin. DDI-spesifikaation 178 elementistä projekti valitsi aluksi käyttöönsä vain sellaisia elementtejä, joilla oli suora vaikutus kehitteillä olevan järjestelmän toimivuuteen. Projektin edetessä projektiryhmä kuitenkin huomasi, että tarvetta myös muiden DDI-spesifikaation sisältämien elementtien käytölle oli olemassa, ja lisäelementtejä otettiinkin käyttöön. Tällä hetkellä Counting California-projektin käytössä on 35 DDI-spesifikaation elementtityyppiä. Projekti käyttää dokumentaation syöttämiseen ja muokkaamiseen sekä valmiiden spesifikaation mukaisten XML-dokumenttien validointiin XMetal-nimistä XML-editoria.

Crusen, Einowskin ja Stratfordin (2002, 11–12) mukaan DDI-spesifikaation avulla saavutetut hyödyt ovat olleet huomattavia. Esimerkiksi kaiken metadatan tallentaminen spesifikaatiossa määritellyssä muodossa on vähentänyt metadatan käyttämiseen liittyviä ristiriitaisuusongelmia. Tämä johtuu siitä, että DDI-spesifikaation mukaisessa dokumentaatiossa jokaiseen tutkimukseen liittyvät tiedot, kuten otsikot ja viittaukset toisiin tutkimuksiin tulee tallennettua samassa muodossa, jolloin ristiriitoja ei ilmene esimerkiksi metadataan perustuvan haun yhteydessä. DDI-spesifikaation muotoinen metadata mahdollistaa myös monien tiedon dokumentointiin ja etsimiseen liittyvien toiminnallisuuksien toteuttamisen. Näitä ovat muun muassa synonyymisanaston (thesaurus) ja lähdeluettelon muodostaminen metadatan pohjalta sekä haettuun materiaaliin yhteydessä olevien tutkimusten ilmoittaminen. Toisaalta Cruse, Einowski & Strat-

ford (2002, 12) toteavat DDI-spesifikaation mukaisen dokumentaation syöttämisen ja ylläpitämisen olevan melko työlästä. Tosin he myös mainitsevat, että tämä johtuu pitkälti siitä, ettei prosessia ole automatisoitu tarpeeksi hyvin. Counting California-projektia hallinnoivat henkilöt pitävät projektiaan DDI-spesifikaation kannalta katsottuna jonkinlaisena pioneerihankkeena, jonka tehtävänä on oman järjestelmän lisäksi kehittää myös DDI-spesifikaatiota kohti yleisesti hyväksyttyä standardia.

3.6.2 Nesstar-projekti

Nesstar (Networked Social Science Tools and Resources) -projekti on Norjan, Iso-Britannian ja Tanskan tietoarkistojen yhteinen projekti, jonka tarkoituksena on kehittää työkaluja, jotka mahdollistavat DDI-spesifikaation mukaisen dokumentaation tehokkaan hyödyntämisen (Norwegian Social Science Data Services 1999). Nesstarin tarjoamat työkalut muodostuvat kolmesta eri ohjelmistokokonaisuudesta:

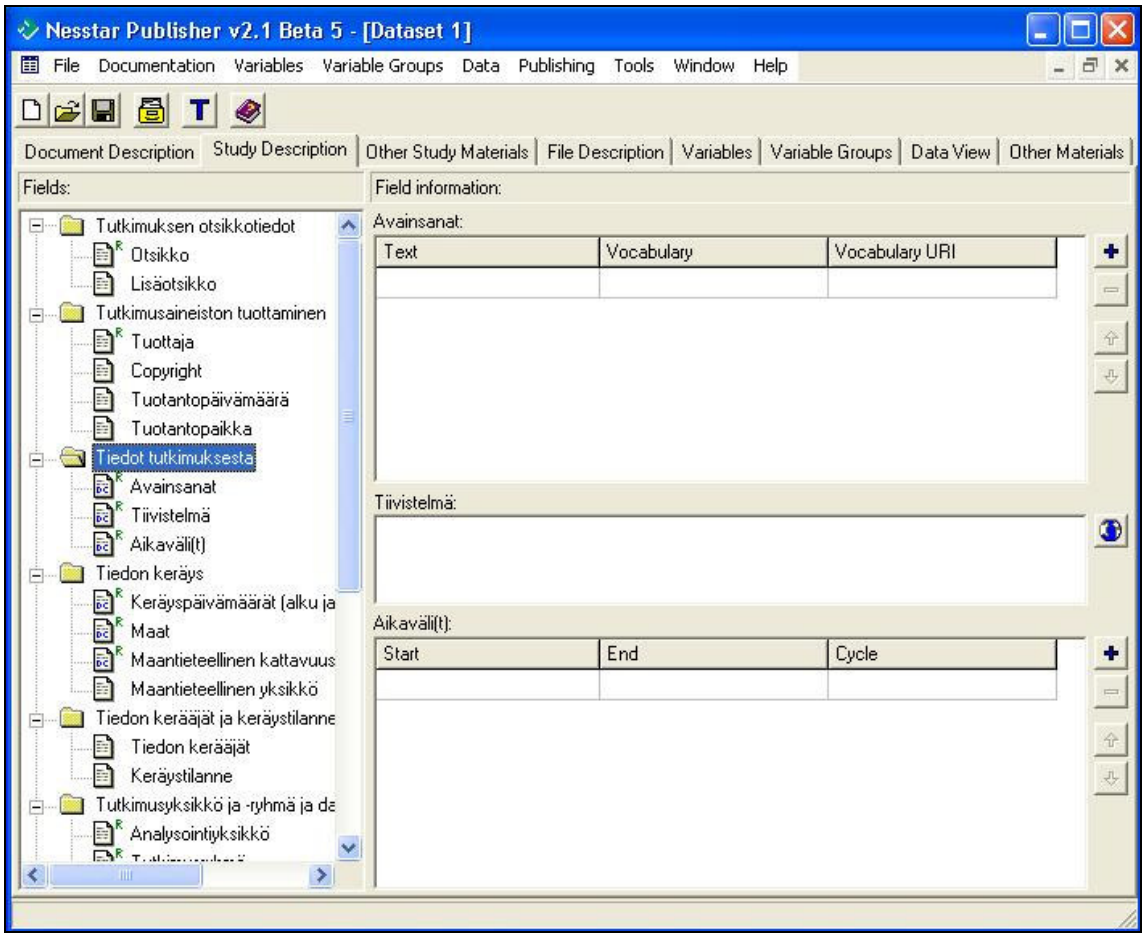
1) *Nesstar Server*. Palvelinohjelmisto, joka toimii varsinaisen palvelinohjelmiston lisäosana. Mahdollistaa yhteiskuntatieteellisen tutkimusaineiston julkaisun verkossa, sisältäen muun muassa haku-, selailu- ja analysointitoiminnot.

2) *Nesstar Publisher*. Kehittynyt tiedon hallintaohjelmisto, joka tarjoaa monipuoliset mahdollisuudet DDI-spesifikaation mukaisen dokumentaation luomiselle ja muokkaamiselle. Publisherin avulla dokumentaation syöttäminen voidaan hoitaa ilman XML-osaamista, koska siinä metadatan syöttäminen tapahtuu lomakepohjaisesti. Dokumentointimallien (template) avulla ohjelmistoa käyttävät organisaatiot voivat itse päättää, mitkä DDI-spesifikaation osa-alueiden sisältämistä elementeistä ja attribuuteista lomakkeissa esiintyy. Lomakkeeseen täytetyt kohdat tallentuvat automaattisesti DDI-spesifikaation mukaiseen XML-dokumenttiin.

Lisäksi Publisher mahdollistaa metadatan tuomisen (import) olevassa olevista DDI-dokumenteista sekä varsinaista tutkimusdataa sisältävistä koontitiedostoista, joista tuetaan kaikkia yleisimpiä tiedostoformaatteja. Esimerkiksi muuttujatiedot on mahdollista tuoda suoraan SPSS-tiedostosta. Muuttujien nimien ja kuvausten lisäksi Publisher osaa automaattisesti tallentaa esimerkiksi kunkin muuttujan kohdalla esiintyvien arvojen vaihteluvälin. Publisher tukee myös DDI-dokumentaation muuntamista (export) yleisempiin tiedostoformaatteihin.

Kuviossa 11 on annettu esimerkkinäkymä Nesstar Publisherin käytöstä. Kuviosta voidaan huomata Publisherin metadatan syötön lomakepohjaisuus: vasemman puoleisesta sarakkeesta voidaan valita kenttä, johon tietoa halutaan syöttää, kun taas oikean puoleinen sarake näyttä valitun kentän syöttölomakkeen. Sarakkeiden yläpuolella olevien välilehtien avulla voidaan lisäksi valita, mitä DDI-spesifikaation osa-aluetta halutaan käsitellä.

3) *Nesstar WebView*. Verkossa olevan yhteiskuntatieteellisen tutkimusaineiston hakemiseen ja selailuun tarkoitettu ohjelmisto, joka asennettaan osaksi internetselainta. WebViewin tarkoituksena on yksinkertaisesti tehostaa etsimistä, selailua ja analysointia.



KUVIO 11. Esimerkki Nesstar Publisherin lomakepohjaisesta metadatan syöttönäkymästä.

Nesstar tarjoaa siis käytännön työkaluja DDI-spesifikaation mukaisen dokumentaation tuottamiseen ja hallintaan. Norjan (Norwegian Social Science Data Services 1999) tietoarkiston julkaisemassa artikkelista ilmenee, että Nesstar-ohjelmistojen kehitystyön pohjana on ollut voimakas ajatus ”yhteiskuntatieteellisestä ihannejärjestelmästä” (Social Science Dream Machine), jonka vaatimuksina ovat olleet muun muassa seuraavat asiat:

- Kaikki olemassa oleva empiirinen tutkimusdata on saatavilla verkosta selailua ja analysointia varten.
- Edellä mainitun tutkimusdatan etsimiseen tarkoitettu integroitu järjestelmä, joka kohdistaa haut samalla kertaa useiden maiden arkistoihin.
- Tutkimusdataan liittyvää metadataa on tarjolla runsaasti.

- Mahdollisuus ladata tutkimusdata omalle koneelle useissa eri tiedostoformaateissa.
- Hakuagentit (knowbots), jotka etsivät aktiivisesti tiettyä tutkijaa kiinnostavaan aihepiiriin liittyvää tietoa.
- Hyperlinkki jokaiseen tietyn tutkimusdatan pohjalta tuotettuun julkaisuun. (Norwegian Social Science Data Services 1999)

Nesstar-projektin kehittelemät ohjelmistot sekä sen harjoittama yhteistyö monien Euroopan tietoarkistojen kanssa viittaa siihen, että useat yllä mainituista vaatimuksista on jo saavutettu tai ainakin ollaan päästy lähellä niiden saavuttamista. Joka tapauksessa Nesstar-projekti ja sen aikaansaannokset tarjoavat hyvän esimerkin siitä, kuinka DDI-spesifikaation mukaista metadataa voidaan hyödyntää käytännössä.

4 DDI-SPESIFIKAATION SOVELTAMINEN AINEISTON DOKUMENTOINTIIN LKK-TUTKIMUKSESSA

Tässä luvussa käsitellään DDI-spesifikaation soveltamista Jyväskylän yliopistossa toimivan Psykocenter-tutkimusorganisaation käyttöön. Soveltaminen on suoritettu Lapsen Kielen Kehitys ja Geneettinen Dysleksiariski (LKK) -tutkimuksen aineistolle siten, että osa tästä aineistosta dokumentoitiin DDI-spesifikaatiota hyödyntäen. Luvussa esitetään myös DDI-pohjainen metadataskeema, joka kehitettiin LKK-tutkimuksen dokumentointitarpeita silmällä pitäen. Tämän luvun tehtävänä on vastata johdannossa esitettyyn tutkimusongelmaan siitä, kuinka pitkittäistutkimukseen liittyvää metadataa voidaan syöttää ja hakea DDI-spesifikaatiota hyödyntäen. Luvun alussa tarkastellaan konstruktiivisen tutkimuksen kulkua ja sen tiedonkeruutapoja. Seuraavaksi esitellään tutkimuksen kohdeympäristö, ja tämän yhteydessä paneudutaan myös niihin tarpeisiin, joita tutkimusaineiston dokumentointiin kohdistuu tässä ympäristössä. Tämän jälkeen esitellään metadatan syöttämistä ja hakemista varten kehitetyt sovellusratkaisut, ja luvun lopuksi näitä ratkaisuja arvioidaan suhteessa kohdeympäristön dokumentointitarpeisiin.

4.1 Tutkimuksen kulku ja tiedonkeruutavat

Pohjatyönä tutkielmassa tehdylle käytännön DDI-toteutukselle voidaan pitää ennen sitä tehtyä, aihepiiriin liittyvää kirjallisuuskatsausta. Siinä selvitettiin olemassa olevan kirjallisuuden kautta tutkimustietoon ja sen hallintaan yleisesti liittyviä teorioita. Kirjallisuuskatsauksessa tutustuttiin myös siihen, miten metadataa tulisi tutkimusaineiston yhteydessä hyödyntää ja minkälainen rooli sillä tässä ympäristössä on. Kirjallisuuskatsauksen kautta tutkielman tekijä sai runsaasti tarvittavaa taustatietoa, joka auttoi käytännön työn läpiviennissä. Teoriaosuus tarjoaa myös tutkielman lukijoille johdatuksen tutkielman aihepiiriin.

DDI-spesifikaation soveltaminen LKK-tutkimuksen aineiston dokumentointiin aloitettiin tutustumalla kohdeorganisaation toimintaan sekä LKK-tutkimuksen tutkimusaineistoon. Hyvänä johdatuksena toimi tutkielman tekijän saama tutkimusraportti, jossa selvitettiin yksityiskohtaisesti muun muassa LKK-tutkimuksen taustalla vaikuttavia asioita, tutkimuksen tavoitteita sekä sen eri vaiheita ja kohdealueita. Tutustumista tehtiin myös paljon henkilökohtaisten keskustelujen sekä sähköpostiviestien kautta. Kohdeympäristöön tutustumisen jälkeen ja osittain sen kanssa päällekkäin tutkielman tekijä suoritti myös DDI-spesifikaation sisällön yksityiskohtaisen analysoinnin. Spesifikaatiota ylläpitävältä taholta sekä ulkopuolisista lähteistä saatujen tietojen avulla tarkasteltiin muun muassa DDI:n kehityshistoriaa ja spesifikaation yhteyttä XML-kieleen. Tarkastelussa kiinnitettiin erittäin paljon huomiota spesifikaation osa-alueisiin, koska haluttiin selvittää tarkasti sitä, minkälaisten tietosisältöjen dokumentointi DDI:n avulla on mahdollista ja löytyykö kaikki LKK-tutkimuksen aineiston dokumentoinnissa tarvittavat kohdat näiden joukosta. DDI-spesifikaatiota analysoidessa kiinnitettiin melko runsaasti huomiota myös sen käytön yhteydessä aikaisemmin havaittuihin haasteisiin.

Kun LKK-tutkimukseen ja DDI-spesifikaatioon liittyvää tuntemusta oli hankittu tarpeeksi, aloitettiin LKK-tutkimukseen ja sen aineistoon liittyvän metadatan syöttäminen DDI-muodossa. Metadatan muodostaminen aloitettiin varovasti valitsemalla dokumentoinnin kohteeksi aluksi vain yksi tutkimuskerta, jonka tietoja kuvailtiin. Tällöin dokumentaatiota syötettiin kaikkiin niihin DDI:n kohtiin, jotka vaikuttivat tutkimuskertaa silmällä pitäen mielekkäiltä. Työn edetessä dokumentoitavien kohtien määrää kuitenkin hiljalleen pienennettiin ja jäljelle jääneiden sisältöä tarkennettiin. Metadatarakenteen selkeytyessä lisättiin dokumentoitavien tutkimuskertojen määrää, jolloin kokeiluun saatiin lisää syvyyttä.

Seuraavaksi selvitettiin sitä, kuinka syötettyä dokumentaatiota voitaisiin hyödyntää parhaiten. Melko nopeasti selvisi, että metatiedon tehokas hyödyntämi-

nen vaatisi monipuolista käyttöliittymää, jonka avulla olemassa olevaa DDI-dokumentaatiota voitaisiin hakea ja selailla. Käyttöliittymän pohjaksi otettiin olemassa oleva, tekstinkäsittely ohjelmalla tehty metadatataulukko. Tätä taulukkopohjaista tekstidokumenttia päätettiin kehittää pidemmälle muuttamalla se Internet-selaimessa tarkasteltavaan muotoon ja lisäämällä taulukkoon tiedot tutkimuskertoihin liittyvistä tärkeimmistä muuttujista sekä linkit SPSS-tiedostoihin, testauslomakkeisiin, metodikuvauksiin sekä tutkimuskertojen pohjalta mahdollisesti julkaistuihin artikkeleihin. DDI-dokumentaation taulukkopohjaiseen käyttöliittymään lisättiin myös mahdollisuus valita mitkä edellä mainituista metatietokentistä halutaan näytettäväksi taulukossa ja mihin tutkimusaihepiireihin liittyvää tietoa tarkestellaan.

Kun olemassa olevan metadataan selailuun tarkoitettu käyttöliittymä oli saatu valmiiksi, tuli kehittää uuden dokumentaation syöttöön tarkoitettu käyttöliittymä. Tämän kehittelyssä päällimmäinen tavoite oli saada XML-muotoisen metadatan syöttäminen mahdolliseksi myös henkilöille, jotka eivät hallitse kyseistä kieltä. Tästä syytä päädyttiin metadatan syöttöliittymän kohdalla lomakepohjaiseen ratkaisuun, jossa lomakkeen eri kenttiin syötetyt tiedot tallentuvat automaattisesti valmiiksi määriteltyyn DDI-spesifikaation mukaisen rakenteen omaavaan XML-dokumenttiin.

Siinä missä tutkielman aikaisemmissa vaiheissa tietoa kerättiin ennen kaikkea olemassa olevaan tieteelliseen kirjallisuuteen tutustumalla, niin dokumentointityön aikana tiedonkeruu hoidettiin pääasiallisesti toteutetun metadatasovelluksen välityksellä. Sovelluksen kautta saatiin tietoa muun muassa niistä tarpeista, joita LKK-tutkimuksen ja sen aineiston dokumentointiin kohdistuu, ja lisäksi sen avulla voitiin myös selvittää DDI-spesifikaation tarjoamia käytännön dokumentointimahdollisuuksia. Metadatasovellusta kehitettiin yhteistyössä tutkielman tekijän sekä LKK-tutkimuksen yhteyshenkilön kanssa. Kussakin sovelluksen kehitysvaiheessa sen toimintaa ja kehitystarpeita arvioitiin viikoittaisten palaverien sekä sähköpostiyhteydenpidon avulla. Metadatasovellukselle asetet-

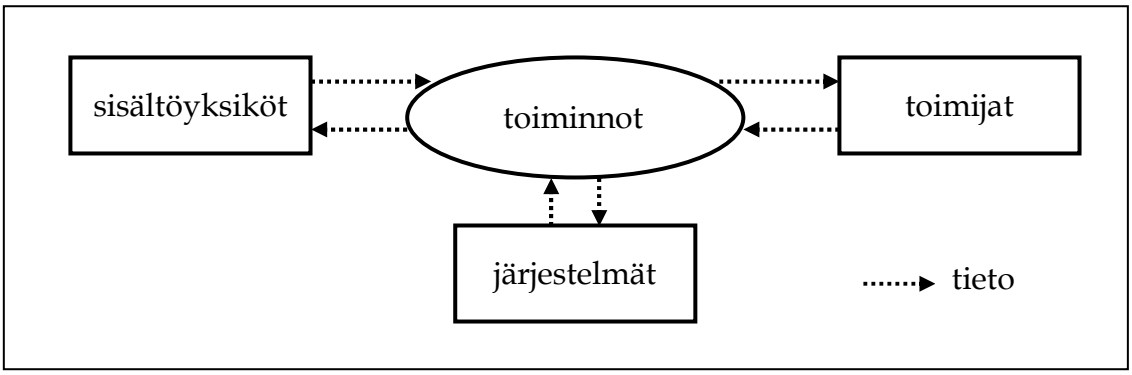
tujen tavoitteiden ja vaatimusten pohjalta voitiin suoraan päätellä DDI-spesifikaatioon kohdistuvat vaatimukset: mikäli metadatasovellukseen haluttiin lisätä jokin ominaisuus, piti siihen löytyä valmiudet DDI:stä. Metadatasovelluksen kehittäminen suoritettiin iteratiivisesti siten, että kunkin kehitysvaiheen tuloksia arvioimalla asetettiin tavoitteet seuraavalle kehitysvaiheelle. Sovellukseen ja sitä kautta koko tutkimukseen liittyvää tietoa kerättiin pääsääntöisesti näissä arviointipisteissä, mutta tietoa kertyi myös tasaisesti koko metadatakokeilun ajan.

4.2 Kohdeympäristön esittely

Tässä kohdassa tarkastellaan kohdeympäristön toimintaa ja resursseja sisällönhallinnan näkökulmasta, sekä kuvataan LKK-tutkimuksen tarpeita tutkimusaineiston dokumentointiin. Kohdeympäristön esittelyn lisäksi kohdan tarkoituksena on tarjota lukijalle esimerkki käytännön tutkimusympäristöstä, johon tutkielmassa esitettyä teoriaa voidaan soveltaa.

4.2.1 Viitekehys kohdeympäristön toiminnan ja sisällönhallinnan tarkastelulle

Kohdeympäristön tarkastelun viitekehyksenä käytetään Salmisen (2005b, 5-6) esittelemää, organisaation sisällönhallinnan tarkasteluun soveltuvaa viitekehystä. Tämän tutkielman yhteydessä voidaan puhua sisällönhallinnasta ja sen kehittämisestä, koska kohdeympäristön tiedonhallintaa tarkastellaan tallennetun tiedon käytettävyyden näkökulmasta. Viitekehyksen mukaan, voidaan organisaation sisällönhallintaympäristöä tarkasteltaessa erotella kahden tyyppisiä peruskomponentteja, toimintoja ja resursseja. Kuviossa 12 on esitelty sisällönhallinnan peruskomponentit sekä tiedonkulku niiden välillä. (Salminen 2005b, 5-6)



KUVIO 12. Sisällönhallinnan osa-alueet. (Salminen 2005b, 6)

Kuviossa oleva soikio esittää kohdeympäristön toimintaprosessien toimintoja ja suorakaiteet puolestaan toimintoihin liittyviä resursseja. Katkoviivat komponenttien välillä kuvaavat *tietoa*, joka syntyy tai jota käytetään toiminnoissa. *Toiminnoilla* tarkoitetaan erilaisia tehtäväsarjoja, joiden suorittamisesta vastaa yksi tai useampi toimija. *Toimijat* ovat puolestaan toimintoja suorittavia organisaatioita tai henkilöitä. *Järjestelmillä* tarkoitetaan niitä laitteistoja, ohjelmistoja, standardeja ja sopimuksia, joita käytetään toiminnoissa. *Sisältöyksiköt* ovat tallennettuja tietokokonaisuuksia, joita syntyy toiminnoissa ja joita käytetään toiminnoissa. Sisältöyksiköt voivat olla esimerkiksi organisaation verkkolevyille tallennettuja dokumentteja tai analogisessa muodossa olevia ääni- tai videonauhoja. (Salminen 2005b, 4-10)

Viitekehyksen esittelyn yhteydessä Salminen (2005b, 6-7) toteaa myös, että tallennettujen sisältöyksiköiden löydettävyyden ja käytettävyyden kannalta oleellista on, että niihin on liitetty tavalla tai toisella metadataa. Tutkimustietoon liittyvää metadataa sekä sen käyttöä käsiteltiin tarkemmin kohdassa 2.2.

4.2.2 Sisällönhallinta LKK-tutkimuksessa

Tämän tutkielman kohdeorganisaationa toimii Jyväskylän yliopistossa toimiva Psykocenter -niminen tutkimusorganisaatio. Psykocenter on Suomen Akatemian huippututkimusyksikköpolitiikkaan kuuluvan tukitoimintojen rahoituksen

piirissä oleva sateenvarjo-organisaatio erikseen valituille itsenäisille tutkimusryhmille. Keskuksen toiminta-ajatuksena on vahvistaa ymmärrystä ihmisen kasvuun, kehitykseen, toimintaan ja vanhenemiseen vaikuttavista tekijöistä. Psykocenterin piiriin kuuluu tässä vaiheessa 26 tutkimusryhmää. Tässä tutkielmassa keskitytään Lapsen Kielen Kehitys ja Geneettinen Dysleksiariski (LKK) -nimisen tutkimuksen ja sen aineiston dokumentoinnin tarkasteluun ja kehittämiseen. LKK-tutkimus on vuonna 1992 aloitettu poikkitieteellinen pitkittäistutkimushanke, joka pyrkii selvittämään kehityksellisen ja geneettisesti välittyvän lukemisvaikeuden, dysleksian prekursoreita. (Psykocenter 1998; Psykocenter 2003)

Seuraavissa kappaleissa hyödynnetään kohdassa 4.2.1 esitettyä viitekehystä tarkastelemalla LKK-tutkimuksen sisällönhallinnan toimintoja, toimijoita, sisältöyksiköitä ja järjestelmiä. Kuviossa 12 nuolilla havainnollistetut tietovirrat koostuvat siitä tiedosta, jota käytetään tai jota syntyy samassa kuviossa esitetyissä toiminnoissa. Käytettävästä tiedosta pääosan muodostaa LKK-tutkimuksessa kaikista tutkimushenkilöistä kerätty tutkimustieto, jota hyödynnetään tutkimukseen sekä sen aineistoon liittyvän metadatan kirjaamisessa. Muuta LKK-tutkimuksen sisällönhallinnassa hyödynnettävää tietoa ovat muun muassa varsinaisen tutkimustiedon pohjalta tuotetut artikkelit. Toiminnoissa syntyvä tieto koostuu puolestaan LKK-tutkimusta koskevasta metadatasta. Kohdeorganisaation tarkastelu perustuu LKK-tutkimuksen väliraportissa (Psykocenter 1998) esitettyihin sekä sähköpostin ja henkilökohtaisten tapaamisten kautta kerättyihin tietoihin.

Toiminnot ja toimijat

LKK-tutkimuksessa tutkittavien lasten joukko muodostuu noin 100 lapsen suuruisesta riskiryhmästä sekä samansuuruisesta seurantaryhmästä. Tutkimuksen päätavoitteena on siis selvittää dysleksian varhaiset ennusmerkit vertaamalla riskiryhmään kuuluvien lasten kehitystä kontrolliryhmän lasten kehitykseen

heti syntymästä lähtien. LKK-tutkimuksen alkuperäisen suunnitelman mukaan tutkimus oli tarkoitus lopettaa keväällä 2005. Tällöin myös nuorimmat vuosina 1993–1996 syntyneistä lapsista olisivat käyneet peruskoulun kolmannen luokan, ja näin ollen kolmannen luokan keväällä pidetyt testaukset olisi saatu suoritettua kaikkien lasten osalta. Alkuperäisestä suunnitelmasta poiketen on kuitenkin noussut selvä tarve jatkaa tutkimusta, koska on tehty niin iso perustyö lasten varhaiskehityksen kartoittamisessa. Tulevaisuudessa tutkimusta jatketaan joko vaikeasti lukiongelmallisiksi osoittautuneiden lasten erityisseurannalla tai genetiikkatutkimuksiin pohjautuva riskilasten luokittelulla, jonka tavoitteena on löytää lisää dysleksian kandidaattigeenejä ja tutkia lasten erilaisten genotyyppien (perimä) vaikutusta dysleksian erilaisiin ilmiasuihin. (Psykocenter 1998; Eklund 2005)

Kuviossa 12 esitetyssä viitekehyksessä olevilla toiminnoilla tarkoitettiin kohdeympäristön sisällönhallintaan liittyviä toimintoja. LKK-tutkimuksen kohdalla nämä toiminnot muodostuvat tutkimukseen liittyvien lukuisten tutkimuskertojen kuvailusta. Jokaisen tutkimuskerran osalta tallennetaan metatietoa muun muassa sen bibliografisista ulottuvuuksista, käsiteltävistä tutkimusaihepiireistä ja -metodeista sekä tutkimuskerran aikana tuotetuista tiedostoista. Lisäksi LKK-tutkimuksen sisällönhallinnan toimintoihin lukeutuu tutkimuskertoihin liittyvän materiaalin (esimerkiksi testauslomakkeet) sekä niiden kautta syntyvän tutkimusaineiston kuvailu metadatan avulla. Metadatan syöttämisestä vastaa tällä hetkellä yksi LKK-tutkimuksen työntekijä, joka tallentaa keskitetysti kaiken eri tutkimuskertoihin liittyvän metadatan. Tulevaisuudessa metadatan tallennuksesta on tarkoitus tehdä kiinteä osa varsinaisen tutkimuksen läpiviemistä siten, että tutkimukseen liittyvä metadata tallennettaan tutkimuksen suorittamisen yhteydessä tutkimusta tekevien tutkijoiden toimesta. Osa tästä tallennuksesta voitaisiin myös automatisoida, jolloin metadatan syöttämiseen tarvittavaa työmäärää voitaisiin pienentää.

Tallennettua metadataa käyttävät lukuisat LKK-tutkimuksen piirissä työskentelevät henkilöt, enimmäkseen tutkijat sekä tutkimusavustajat. Metadatan avulla voidaan selvittää muun muassa sitä, millaisilla tutkimusmenetelmillä tietyt tutkimustulokset on saatu aikaisiksi ja esimerkiksi sitä, minkälaisia lisätietoja liittyy käytettyihin muuttujiin. Metatiedoista on erityisesti hyötyä sellaisille henkilöille, joille LKK-tutkimus ja sen eri vaiheet ovat ennestään vieraita.

Sisältöyksiköt

LKK-tutkimuksen sisällönhallintatoiminnoissa muodostetaan ja niissä hyödynnetään suurta joukkoa erilaisia sisältöyksiköitä. Toimintojen kautta muodostettavilla sisältöyksiköillä tarkoitetaan luonnollisesti metadatan syöttämisen kautta syntyviä DDI-muotoisia metadatadokumentteja. Metadatan luomisessa hyödynnettävillä sisältöyksiköillä tarkoitetaan puolestaan kaikkia niitä digitaalisessa tai analogisessa muodossa olevia dokumentteja, jotka muodostavat LKK-tutkimuksen hallussa olevan tutkimusaineiston. Taulukossa 1 on esitetty LKK-tutkimuksen sisällönhallinnassa hyödynnettävät sisältöyksiköt sekä kuvaus kullekin sisältöyksikölle. Tutkimusta koskeva metadata muodostuu sekä näistä sisältöyksiköistä kerätyistä tiedoista että näiden sisältöyksiköiden tiedostotasoisesta kuvailusta.

TAULUKKO 1. Metadatan luomisessa hyödynnettävät sisältöyksiköt.

Sisältöyksikkö	**Kuvaus**
Tutkimusdata	Testeistä ja haastatteluista kerätty informaatio on koottu SPSS-tiedostoihin. Kaikki arvioitavaan henkilöön ja testaustilanteeseen liittyvät tiedot on tallennettu omiin muuttujiinsa.
Testauslomakkeet	Testauslomakkeet sisältävät ne tiedot, jotka tutkimusta suorittava henkilö tarvitsee kunkin tutkimuskerran läpiviemiseen. Lomakkeet sisältävät muun muassa yksityiskohtaiset ohjeet jokaisen kysymyksen ja testin suorittamiseen. Lisäksi lapsen antamat vastaukset ja tutkijan huomiot kirjataan testauslomakkeisiin, joista ne myöhemmin kirjataan tietokoneelle. Usein testauslomake sisältää myös testeissä käytettävät ärsykkeet.

Ärsykkeet	Joidenkin tutkimuskertojen yhteydessä käytettävät ärsykkeet on kirjattu erillisiin dokumentteihin. Näitä dokumentteja käytetään testauslomakkeiden rinnalla tutkimuskertojen läpiviemiseen.
Metodikuvaukset	Tutkimuskertoihin liittyvät metodikuvaukset on myös tallennettu erillisiin dokumentteihin. Metodikuvauksissa kuvaillaan käytetty metodi sekä annetaan tiedot metodin kehittäjästä sekä tutkimuksista, jossa sitä on käytetty aiemmin.
Ääni- ja videonauhat	Joihinkin tutkimuskertoihin sisältyy myös osioita, joissa lapsen antamat vastaukset tai lapsen reaktiot äänitetään tai videoidaan. Ääni- ja videonauhat on tallennettu suurimmilta osin analogisessa muodossa.
Tietolähteet	Dokumentteja, joissa kuvaillaan tutkimuksissa käytettävät tietolähteet. Tietolähteet voivat olla esimerkiksi tiettyyn tutkimusaihepiiriin liittyviä, aikaisemmin tehtyjä tutkimuksia tai niitä käsitteleviä tutkimusartikkeleja.
Artikkelit	Tutkimuksen tai siihen sisältyvien tutkimuskertojen pohjalta julkaistut tutkimusartikkelit.
Raportit	Tutkimusta kuvaavia raportteja, jossa annetaan selvitys muun muassa siitä missä vaiheessa tutkimus on kunkin raportin valmistushetkellä sekä millaisia jatkotutkimussuunnitelmia on vireillä. Raporttien avulla kerätään myös varsinaista tutkimusaineistoa. Tällöin lapsen kielen kehitystä arvioidaan testaajien, vanhempien tai opettajien tekemien raporttien avulla.

Sisältöyksiköiden tarkastelua voidaan pitää erittäin tärkeänä, koska juuri niiden käytettävyyttä metadatan avulla halutaan parantaa. Salminen (2005b) toteaa: ”Jotta sisältöyksiköt olisivat löydettävissä ja käytettävissä täytyy niihin liittää tavalla tai toisella metatietoa”. Kaikki edellä mainittuja sisältöyksiköitä käsittelevät ohjelmistot tallentavat metadataa, jonka avulla kukin ohjelmisto pystyy käyttämään tietovarantojaan. Ohjelmistokohtainen metadata on kuitenkin usein tallennettu sellaisessa muodossa, että se hyödyntäminen ei ole mahdollista ilman kyseistä ohjelmaa. (Salminen 2005b, 6) Koska LKK-tutkimus pyrkii löytämään pitkäaikaisen ja ohjelmistoriippumattoman metadataratkaisun, tulee metadataa tarkastella toimijoiden, toimintojen ja ohjelmistojen välisen yhteistyön näkökulmasta, ei yksittäisen ohjelmiston näkökulmasta. Metadata tulee myös

tallentaa sellaisessa muodossa, että sitä voidaan hyödyntää erilaisilla ohjelmistoilla (Salminen 2005b, 6).

Järjestelmät

Järjestelmillä tarkoitetaan niitä laitteistoja, ohjelmistoja, standardeja ja sopimuksia, joita käytetään LKK-tutkimuksen toiminnoissa. Järjestelmiä käytetään ennen kaikkea edellä mainittujen sisältöyksiköiden tuottamiseen ja analysointiin, mutta niitä voidaan hyödyntää tutkimustoiminnan tukena myös muilla tavoilla. Taulukossa 2 on esitelty LKK-tutkimuksessa käytettävät järjestelmät sekä kuvaus kullekin järjestelmälle.

TAULUKKO 2. LKK-tutkimuksen järjestelmät.

Järjestelmä	**Kuvaus**
SPSS (Statistical Package for the Social Sciences)	SPSS -ohjelmistoa käytetään tutkimuskerroissa saatujen tietojen tallentamiseen ja analysointiin. Ohjelmiston avulla voidaan tilastopohjaisesta tiedosta tehdä erilaisia tilastollisia mallintamisia, yhteenvetoja, graafeja ja taulukoita.
Microsoft Word	Tekstinkäsittelyohjelmalla laaditaan muun muassa testauslomakkeet, ärsykedokumentit, raportit sekä kaikki muut kirjalliset dokumentit.
Microsoft Access	Access-tietotokantoja käytetään tutkimuksessa mukana olevien perheiden yhteistietojen ja lomakkeiden lähettämisaikataulujen tallentamiseen. Lisäksi tietokantoihin on tallennettu tiedot tutkittavien lasten sairauksista.
Cognitive Workshop	Cognitive Workshop -ohjelmistoa käytetään testaustilanteissa ärsykkeiden esittämiseen lapselle sekä saatujen vastausten tallentamiseen. Lisäksi ohjelmiston avulla voidaan mitata reaktio- ja vastausaikoja automaattisesti.
Sähköposti	Sähköpostia käytetään monipuolisesti tutkimustoiminnan tukivälineenä. Sähköpostin avulla hoidetaan myös yhteydenpitoa ulkopuolisiin tahoihin.

4.2.3 LKK-tutkimuksen tarpeet tutkimusaineiston dokumentointiin

Tutkielman tekemisen alkuvaiheessa selvisi, että kiinnostus DDI-spesifikaatiota kohtaan oli LKK-tutkimuksen parissa herännyt Yhteiskuntatieteellisestä tietoarkistosta (FSD) tulleen aloitteen pohjalta. FSD toimii yhteiskuntatieteellisen tutkimuksen ja opetuksen valtakunnallisena palveluyksikkönä, joka arkistoi ja välittää elektronisia tutkimusaineistoja tutkimus- ja opetuskäyttöön. Tietoarkisto vastaanottaa ensisijaisesti aineistoja, jotka soveltuvat yhteiskuntatieteelliseen tutkimukseen, ja tästä syystä myös LKK-tutkimuksessa syntynyt tutkimusaineisto sopisi hyvin FSD:n arkistoitavaksi. FSD käyttää arkistoimiensa aineistojen kuvailukielenä DDI-spesifikaatiota. Jotta LKK-tutkimuksessa syntynyttä tutkimusaineistoa voitaisiin luovuttaa FSD:hen laajempaan tutkimus- ja opetuskäyttöön, täytyisi aineisto siis ensin dokumentoida DDI-spesifikaatiota käyttäen. FSD:n kautta LKK-tutkimuksen aineisto olisi helpommin myös muiden tutkijoiden saatavilla, minkä avulla tutkimustiedon jakamisen (kts. 2.1.3) kautta saavutettavia myönteisiä vaikutuksia, kuten esimerkiksi tiiviimpää yhteistyötä muiden tutkijoiden kanssa, voitaisiin tavoitella.

LKK-tutkimuksen sisällä tutkimusaineiston kuvailu nähdään yleisesti ottaen erittäin tärkeänä asiana, johon ei kuitenkaan ole kiinnitetty tarpeeksi huomiota. Aineistoa on vuosien mittaa kertynyt erittäin suuri määrä, eikä sitä ole kuvailtu millään standardoidulla tavalla. Dokumentaation vaillinaisuus tulee eritoten esiin silloin, kun aineistoa tulisi hyödyntää jonkun tutkimukselle vieraan henkilön toimesta. Muun muassa LKK-tutkimukseen tulevien uusien tutkimushenkilöiden sekä ulkopuolisten tutkijoiden on osittain erittäin vaikea päästä sisään tutkimusaineistossa käsiteltyihin asioihin metatiedon puuttumisesta johtuen. Osittain tutkimusaineistosta tekee tosin vaikeasti lähestyttävää myös sen laajuus ja LKK-tutkimuksen pitkä elinkaari.

LKK-tutkimuksen sisäisiä ja ulkoisia metadatatarpeita arvioitaessa nähtiin, että tämän tutkielman kohdalla on tärkeämpää pyrkiä vastaamaan tutkimuksen si-

sältä tuleviin tarpeisiin. Tästä syystä päätettiinkin, että tutkielman tekemisen yhteydessä lähdetään ennen kaikkea kehittämään ratkaisua, josta on hyötyä LKK-tutkimuksessa mukana oleville henkilöille. Koska dokumentaatio toteutettiin DDI-spesifikaatiota hyödyntäen, säilytettiin kuitenkin myös mahdollisuus sille, että aineistoa voitaisiin jatkossa mahdollisesti luovuttaa FSD:n arkistoitavaksi.

LKK-tutkimuksen dokumentointitavoitteiden selvittämisen pohjana toimii hyvin ennen DDI-kuvailua tehty dokumentaatio, jossa kaikki tutkimuskertoja koskeva metadata on tallennettu yhteen tekstinkäsittelyohjelmalla tehtyyn dokumenttiin. Kuviossa 13 on annettu esimerkki tämän dokumentin sisällöstä, ja siitä voidaan huomata että ennen tämän tutkielman suorittamista metadatan avulla on kuvailtu tutkittavan lapsen ikä, tutkimusmittarit ja tallennetun tiedon tyyppi.

3.5 (Lab 2 visits)	NEPSY - ***A Developmental Neuropsychological Assessment*** Korkman, M., Kirk, U., & Kemp, S. (1998) ***Attention/Executive***: Visual attention. ***Language***: Naming of body parts; Phonological processing A; Comprehension of instructions; Repetition of nonsense words. ***Sensorimotor***: Visuomotor precision; Finger discrimination. ***Visuospatial)***: Design copying; Block construction; Picture recognition. ***Memory***: Immediate memory for faces; Narrative memory; Sentence repetition.	Test
3.5 (Lab 2 visits)	Orthographic skills 1) Identification of letters 2) Naming letters 3) Logographic identification of words 4) Visual matching 5) Identification and writing of one=s own name	Test
3.5 (Lab 2 visits)	Other language measures: 1) Boston Naming Test 2) Peabody Picture Vocabulary Test (adaptively) 3) Morphology Test 4) Digit Span 5) Rapid naming (RAN): Objects	Test

KUVIO 13. Esimerkki aikaisemman metadatadokumentin sisällöstä.

DDI-spesifikaation avulla haluttiin olemassa olevaa dokumentaatiota viedä eteenpäin lisäämällä kuvailtavien asioiden määrää sekä tarkentamalla jo kuvail-

tuja seikkoja. Lisäksi massiivisen tekstidokumentin sisältö haluttiin jakaa pienempiin osiin tallentamalla kutakin tutkimuskertaa koskeva dokumentaatio erilliseen DDI-dokumenttiin. Tätä kautta metadatan hallintaa ja käsittelyä pysyttiin yksinkertaistamaan, ja lisäksi näin toimien pysyttiin kunkin tutkimuskerran tiedot tallentamaan huomattavasti monipuolisemmin. Aikaisempaan dokumentaatioon verrattuna DDI-dokumentaation haluttiin lisätä muun muassa kunkin tutkimuskerran käsittelemät aihepiirit sekä tutkimuskertoihin liittyvät tiedostot ja tärkeimmät muuttujat. Tarkennusta dokumentaatioon haluttiin puolestaan esimerkiksi lisäämällä lisätietoja useisiin dokumentaation eri kohtiin. Toteutettavien lisäysten ja tarkennusten avulla haluttiin dokumentaatiosta tehdä informatiivisempaa siten, että sen avulla LKK-tutkimuksen pitkälle aikavälille sijoittuvat tutkimusvaiheet voisi hahmottaa helpommin yhtenä kokonaisuutena ja, että sen koko tutkimusaineiston käytettävyyttä voitaisiin parantaa. Pidemmällä tähtäimellä tärkeä tavoite LKK-tutkimuksen sisällä on kuvailla koko tutkimus niin tarkasti, että sen keräämää aineistoa voidaan hyödyntää tehokkaasti myös tulevaisuudessa, tutkimuksen päätyttyä, ja että tutkimuksen eri vaiheet olisivat toistettavissa metadatan avulla.

4.3 Kehitetyt sovellusratkaisut

Tässä kohdassa esitellään LKK-tutkimuksen dokumentointitarpeiden pohjalta kehitetyt sovellusratkaisut. Kohdan käsittely aloitetaan perustelemalla valitut toteutustekniikat sekä esittelemällä LKK-tutkimuksen tarpeisiin kehiteltyä DDI-pohjaista metadataskeemaa. Tämän jälkeen käsitellään toteutettua metadataratkaisua esittelemällä aluksi sen yleistä toimintaperiaatetta. Kohdan tärkeimmän sisällön muodostaa metadatan syöttämiseen sekä sen hakemiseen ja selailuun kehitettyjen sovellusten esittely, joka tehdään kohdissa 4.3.2 ja 4.3.3.

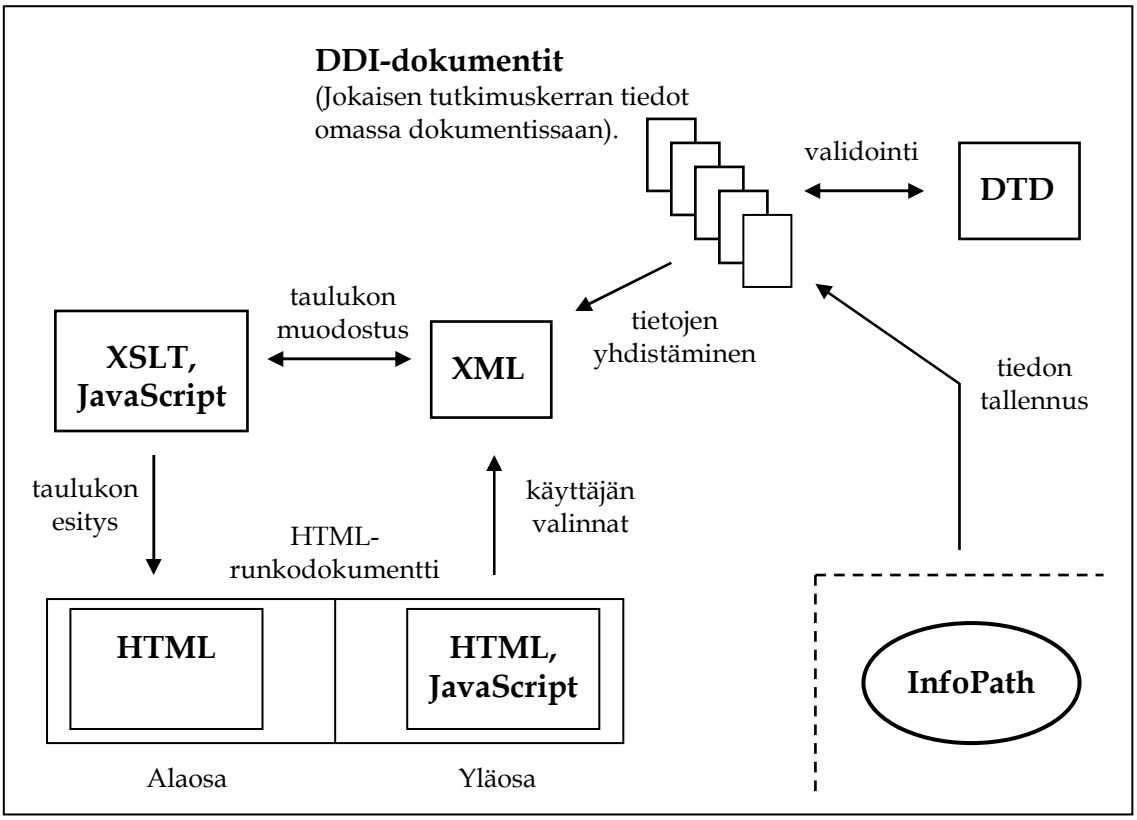
4.3.1 Valitut toteutustekniikat ja LKK-tutkimusta varten kehitetty DDI-skeema

Syöttöön tarkoitetun lomakepohjaisen käyttöliittymän toteutukseen näytti olevan tarjolla kaksi eri ohjelmistoratkaisua: Nesstar Publisher ja Microsoft Office InfoPath. Kokeilun aikaisessa vaiheessa Nesstar Publisheria pidettiin oikeana vaihtoehtona; olihan se kehitetty nimenomaan DDI-dokumentaation syöttämistä ja hallinnointia varten. Melko pian kuitenkin huomattiin, että Publisherissa tutkimuskertoihin liittyvien tiedostojen kuvailu ei vastannut LKK-tutkimuksen vaatimuksia. Jostain syystä Nesstar Publisherin avulla ei ollut mahdollista kuvailla kutakin tutkimusta tai tutkimuskertaa koskien kuin yhden tiedoston tiedot. LKK-tutkimuksessa jokaiseen tutkimuskertaan liittyy kuitenkin useita tiedostoja, joiden kaikkien kuvailu on tärkeää. Tästä Publisherin puutteesta johtuen päätettiin koettaa lomakepohjaisen syöttöliittymän toteuttamista Microsoftin Office-ohjelmistopaketin mukana tulevan InfoPath-ohjelman avulla. Sen etuina Publisheriin verrattuna on monipuolisemmat mahdollisuudet lomakkeen suunnitteluun, sisältäen mahdollisuudet useampien tutkimustiedostojen kuvailuun. InfoPath ei myöskään aiheuta LKK-tutkimukselle ohjelmistoinvestointeja, koska se tulee Office-paketin mukana. Lisäksi InfoPath on ulkoasultaan tuttu Office-paketin ohjelmia käyttäneille (esimerkiksi Word tai Excel). InfoPathin heikkoutena Nesstar Publisheriin verrattuna voidaan puolestaan pitää sitä, ettei siinä ole mahdollista muodostaa tutkimuskertaan liittyviä muuttujatietoja automaattisti tutkimusaineistosta, kuten Publisherissa.

InfoPath-lomakkeen avulla syötetyn DDI-muotoisen metadatan haku- ja selailukäyttöliittymän osalta päädyttiin itse kehiteltyyn ratkaisuun, koska dokumentoitavien asioiden tiimoilta liikuttiin erittäin spesifillä alueella, eikä valmiita sovelluksia näin ollen ollut olemassa. Sovelluksesta haluttiin myös luoda rakenteeltaan mahdollisimman kevyt siten, että se sisältäisi vain halutut ominaisuudet. Metadatasovelluksen kehittämiseen oli tarpeen valita ohjelmointikieli, jonka avulla olisi mahdollista sellaisten toimintojen toteutus, joiden avulla selailu-

liittymän käyttäjä pystyisi rajamaan näytettävän metatiedon määrää. Valitun ohjelmistokielen tuli olla suhteellisen helppokäyttöinen, yleisesti käytössä oleva sekä yhteensopiva yleisempien Internet-selainten sekä XSLT-muunnoskielen kanssa. Lisävaatimuksena oli myös, että valitun ohjelmointikielen piti olla palvelinriippumaton, eli sen tuli toimia kaikilla tietokoneilla ilman erillisten palvelinohjelmistojen tukea ja lisäksi kielen ohjelmoinnin tuli olla vaivatonta samoilla työkaluilla kuin XSLT-muunnoskielen kirjoittaminen. Edellä mainittujen vaatimusten nojalla päädyttiin JavaScript-ohjelmointikielen käyttöön. Kyseinen kieli on tarkoitettu nimenomaan parantamaan Internet-selaimissa tarkasteltavien dokumenttien toiminnallisuutta. Lisäksi JavaScript on erittäin laajalti käytössä maailmanlaajuisesti, minkä ansiosta Internetistä löytyy runsaasti laajoja ja monipuolisia sivustoja, joilta saa tukea ongelmatilanteissa. XSLT-muunnoskielen käyttäminen DDI-dokumenteissa olevan metadatan poimimiseen ja esittämiseen oli tutkielman tekijän mielestä erittäin luonnollinen ja perusteltu vaihtoehto, koska kyseinen on suunniteltu juuri kyseiseen käyttötarkoitukseen. Lisäksi XSLT kuuluu XML-kielin kanssa samaan, W3C:n hallinnoimaan kieliperheeseen. (W3C 1999)

Liitteessä 2 on esitelty LKK-tutkimuksen dokumentointitarpeiden pohjalta kehitelty DDI-pohjainen metadataskeema. Skeema on kehitelty yhteistyössä LKK-tutkimuksen yhteyshenkilön kanssa tarkastelemalla LKK-tutkimuksen asettamia vaatimuksia ja DDI-spesifikaation tarjoamia mahdollisuuksia. Liitteestä käy ilmi kaikki LKK-tutkimuksen ja sen aineiston dokumentoinnissa käytetyt metadatakentät sekä niiden tietosisältö.



KUVIO 14. Metadatasovelluksen toiminta.

Kuviossa 14 on esitelty kaavakuva toteutusta metadataratkaisusta. Kuviossa esitetyt suorakulmiot kuvastavat niitä sovellusdokumentteja, joiden avulla haku- ja selailukäyttöliittymän toiminta on toteutettu. Nuolet puolestaan kuvastavat dokumenttien välisiä suhteita. Kuvion oikeassa alakulmassa on katkoviivalla muusta kuviosta eroteltu soikio, joka kuvaa InfoPath-sovelluksella toteutettua metadatan syöttöliittymää. Kuhunkin tutkimuskertaan liittyvät metatiedot tallennetaan omaan DDI-dokumenttiinsa hyödyntäen InfoPathin syöttölomaketta. Syöttämisen jälkeen kaikki eri tutkimuskertoja kuvailemat DDI-dokumentit validoidaan DTD-dokumentilla, joka sisältää DDI-spesifikaation mukaisen dokumenttityypin määrittelyn. Validoinnin avulla voidaan varmistaa, että InfoPath-lomakkeen avulla tallennetut metadatadokumentit ovat spesifikaation mukaisia. DDI-dokumenttien tuottaminen InfoPath-lomakkeen avulla esitellään tarkemmin seuraavaksi. Muut kuviossa 14 esiintyvät dokumentit liittyvät metadatan selailuun, joka esitetään hieman edempänä, kohdassa 4.3.4.

4.3.2 Metadatan syöttäminen

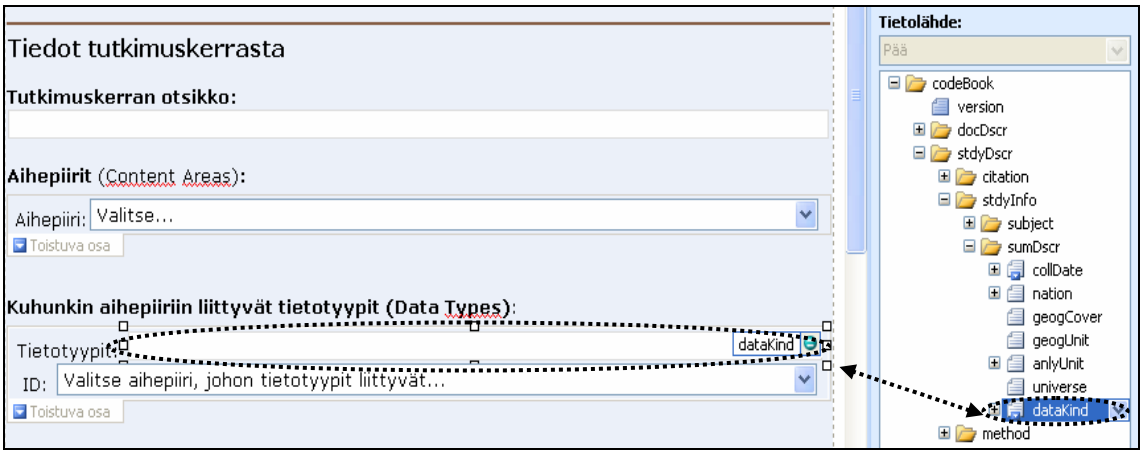
Metadatan syöttämistä varten suunniteltiin ja toteutettiin käyttöliittymä Microsoft Office 2003 -ohjelmistopakettiin sisältyvällä InfoPath-sovelluksella. InfoPath on kehitetty nimenomaan XML-pohjaisen tiedon lomakepohjaiseen syöttämiseen, joten se sopii hyvin myös DDI-muotoisen metadatan syöttöliittymäksi.

Liitteessä 3 on esitelty LKK-tutkimuksen henkilökunnan käyttöön kehitelty DDI-sovelluksen käyttöohje. Siinä on kuvattu vaihe vaiheelta se, kuinka uusi, tiettyä tutkimusvaihetta kuvaileva DDI-dokumentti luodaan, ja kuinka sitä pystytään muokkaamaan jälkeenpäin. Ohjeessa on selitetty myös metadatan hakemiseen ja selailuun kehitetyn sovelluksen käyttö sekä DDI-dokumenttien luomiseen tarkoitetun InfoPath-lomakepohjan muokkaaminen.

InfoPath-sovelluksessa on kaksi erillistä toimintatilaa, joita ovat lomakkeiden suunnittelutila ja lomakkeiden täyttötila. Suunnittelutilassa voidaan määritellä minkälaisia kenttiä metadatalomakkeeseen halutaan sisällyttää ja mihin XML-dokumentin elementteihin tai attribuutteihin nämä kentät halutaan yhdistää. Lisäksi suunnittelutilassa voidaan suunnitella lomakkeen ulkoasu ja lisätä mahdollisia tekstiosuuksia (otsikot, ohjetekstit jne.). Lomakkeen suunnittelu aloitetaan siten, että muodostettavan lomakkeen pohjaksi otetaan olemassa oleva XML-dokumentti tai XML Schema -määrittelykielillä toteutettu rakennemäärityys. InfoPath analysoi pohjadokumenttina käytettävässä XML-dokumentissa tai annetussa rakennemäärityksessä olevan, elementtien ja attribuuttien muodostaman rakenteen, ja käyttää tätä rakennetta InfoPath-lomakkeen pohjana toimivan tietolähteen luomiseen. Tietolähteen avulla InfoPath-lomakkeen suunnittelu on mahdollista, ja oikean rakenteen omaava XML-dokumentti voidaan muodostaa lomakkeen pohjalta. Mikäli lomakkeen tietolähde muodostetaan XML-dokumentin avulla, voidaan pohjadokumentissa olevien elementtien ja attribuuttien sisältöä käyttää InfoPath-lomakkeen kenttien oletusarvoina. LKK-

tutkimuksen ja sen aineiston dokumentoinnissa tällaisiksi oletusarvoiksi on tallennettu sellaisia tietoja, jotka toistuvat samanlaisina kaikissa tutkimuskerroissa. Näitä tietoja ovat esimerkiksi copyright-tiedot sekä yhteyshenkilöiden nimet ja sähköpostiosoitteet. Oletusarvojen avulla voidaan metatietojen syöttämiseen tarvittavaa työmäärää hieman pienentää.

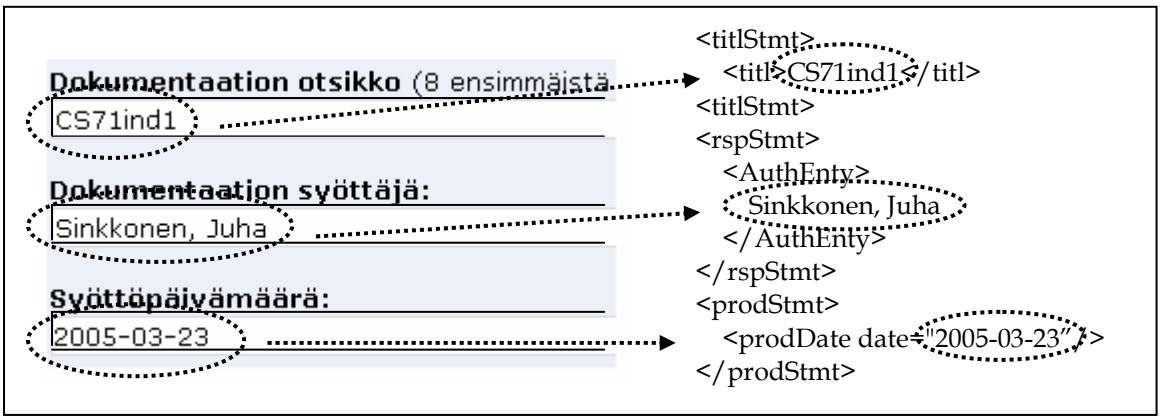
Kun InfoPath-lomakkeen tietolähde on muodostettu, lomakkeen kenttien suunnittelu voidaan aloittaa. InfoPathin avulla lomakkeeseen voidaan liittää monia erilaisia tiedon syöttökenttiä, kuten luetteloruutuja, päivämäärävalitsimia, valintaruutuja sekä avoimia tekstin syöttökenttiä. Lisäksi voidaan luoda osioita, jotka ovat toistuvia tai valinnaisia. Toistuvien osioiden avulla lomakkeen täyttäjä voi syöttää tietoja sellaisten XML-elementtien sisään, jotka toistuvat muodostettavassa DDI-dokumentissa useaan kertaan. Esimerkiksi tutkimuskertaan liittyvien tietojen tallennuksen yhteydessä kuhunkin tutkimuskertaan saattaa sisältyä useita tutkimusaihepiirejä. Tällöin tiedot tallennetaan toistuvien <subject> ja </subject> tunnisteparien sisään. Jotta lomakkeeseen voitaisiin luoda toistuva osio, täytyy elementin toistettavuuden olla sallittua lomakkeen pohjana toimineessa rakennemäärityksessä ja sitä kautta lomakkeen tietolähteessä. Valinnaisten osioiden avulla lomakkeen täyttäjä voi puolestaan päättää sisällyttääkö tietyn kentän lomakkeeseen vai ei. Myös valinnaisten osioiden kohdalla tulee elementtien valinnaisuuden olla sallittua lomakkeen tietolähteessä. Jokainen kenttä tai osio tulee liittää johonkin tietolähteessä olevaan elementtiin tai attribuuttiin (kts. kuvio 15). Liittämisen kautta InfoPath osaa tallentaa kuhunkin kenttään syötetyt tiedot tulosdokumentin oikeaan kohtaan.



KUVIO 15. Kentän liittäminen tietolähteeseen InfoPath-sovelluksessa.

Kun kaikki halutut kentät on sisällytetty lomakkeeseen ja ne on liitetty InfoPathin tietolähteeseen, voidaan lomakepohja tallentaa ja saattaa käyttäjien saataville, jotta XML-pohjaisen tiedon syöttäminen sen avulla voitaisiin aloittaa.

InfoPath-sovelluksen lomakkeiden täyttötilan käyttämien on erittäin yksinkertaista. Suunnittelutilassa lomakkeeseen sijoitetut kentät täytetään asiaankuuluvalla tiedolla minkä jälkeen lomakkeen tiedot tallennetaan. Tällöin lomakkeeseen syötetyt tiedot tallentuvat tietolähteessä määritellyn rakenteen omaavaan XML-dokumenttiin (kts. kuvio 16). Lomakkeen avulla muodostetut XML-dokumentit voidaan myös avata InfoPathiin uudelleenkäsiteltäväksi, mikäli niihin tulee tehdä muutoksia tai lisäyksiä. Näin ollen myös metatietojen ylläpito on tehty lomakepohjaisuuden kautta yksinkertaiseksi.



KUVIO 16. InfoPath-lomakkeen yhteys muodostettavaan XML-dokumenttin.

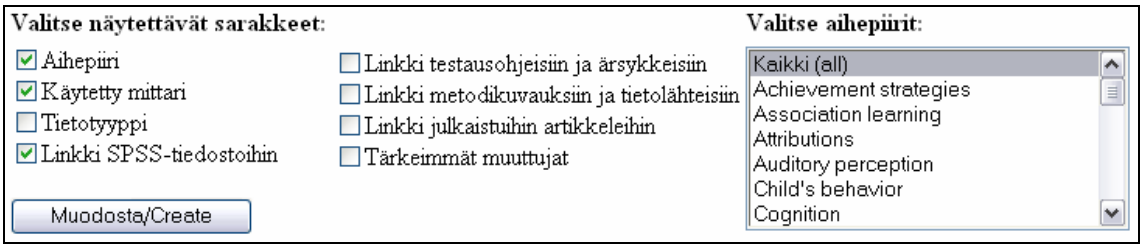
Liitteessä 4 on esitelty tarkemmin ne kentät jotka on sisällytetty LKK-tutkimuksen metatietojen syöttämiseen tarkoitettuun InfoPath-lomakkeeseen.

4.3.3 Metadatan haku ja selailu

Kuviosta 14 nähdään, että metadatan selailukäyttöliittymän keskeisimpänä osana toimii XML-dokumentti, joka sisältää listan kaikista olemassa olevista, eri tutkimuskertojen tietoja sisältävistä DDI-dokumenteista. Kun uusi DDI-dokumentti on luotu InfoPath-lomakkeen avulla, lisätään XML-dokumentissa olevaan listaan uuden DDI-dokumentin nimi ja tallennussijainti. Listan avulla pidetään siis huoli siitä, että kaikkien DDI-dokumenttien tiedot näkyvät muodostettavassa metadatataulukossa. Esimerkki XML-dokumentin sisällöstä on annettu seuraavalla sivulla olevassa kuviossa 17.

DDI-muotoisen metadatan selailu aloitetaan siten, että käyttäjä avaa Internet-selaimeen HTML-runkodokumentin, joka on esitetty kuvion 14 alaosassa. Tämä dokumentti toimii selailun runkona ja se on jaettu kehyksiä käyttämällä kahteen eri osaan, joista kumpikin saa sisältönsä erillisestä dokumentista. Yläosa muodostetaan JavaScriptejä ja HTML-koodia sisältävästä lomakkeesta, jonka avulla käyttäjä tekee valinnat koskien sitä, mitä metatiedon osia hän haluaa tarkastella. Yläosa pysyy kokoajan muuttumattomana, minkä ansiosta käyttäjä voi halutessaan muuttaa lomakkeen valintoja ja näin ollen rajoittaa tai laajentaa tarkasteltavan metatiedon määrää. Kuviossa 18 on annettu esimerkkikuva HTML-runkodokumentin yläosasta, ja siitä voidaan huomata, että käyttäjä pystyy valitsemaan lomakkeesta muodostettavaan metadatataulukkoon tulevat sarakkeet sekä näytettävät tutkimusaihepiirit.

KUVIO 17. XML-muotoinen lista tutkimuskertoja koskevista DDI-dokumenteista.



KUVIO 18. HTML-runkodokumentin yläosa.

HTML-runkodokumentin alaosa puolestaan muodostetaan käyttäjän tekemien valintojen mukaisesti. Alaosa on siis tarkoitettu DDI-dokumenteissa kuvailtujen tutkimuskertojen metatietojen esittämiseen. Metatiedot näytetään käyttäjälle dynaamisesti muodostetun HTML-dokumentin avulla, joka sisältää taulukkomuotoisen esityksen metatiedoista. Taulukko muodostetaan hyödyntämällä kuviossa 14 esiintyviä XML- ja XSLT-dokumentteja. Yläosan lomakkeeseen teh-

dyt käyttäjän valinnat välitetään XML-dokumentille, joka sisältää listan kaikista DDI-dokumenteista. Tämän XML-dokumentin ulkoasumääritykset on tehty erillisessä XSLT-dokumentissa, ja käyttäjän valinnat välitetään edelleen tälle dokumentille. XSLT-dokumentin avulla poimitaan käyttäjän valintojen mukaisesti XML-dokumentista ja sitä kautta DDI-dokumenteista ne metatietokentät, joiden tulee näkyä taulukossa. XSLT-dokumentissa on käytetty myös JavaScriptejä, joiden avulla käyttäjän valintamahdollisuuksia on voitu parantaa monella tapaa. Alla on esitetty kaksi esimerkkikuvaa XSLT-dokumentista. Ylemmässä kuvassa (kuvio 19) muodostetaan taulukon otsikkorivi JavaScriptin avulla ja alemmassa kuvassa (kuvio 20) puolestaan muodostetaan tutkimusmittarit ja tietotyypit sisältävät taulukon solut yhdistämällä JavaScript-ohjelmointia XSLT-määrityksiin.

```
<script language="JavaScript">
  if (show_measure == 1) {
    document.write('<td width="30%"><b>Content/Measure</b></td>');
  }
  if (show_data_type == 1) {
    document.write('<td width="12%"><b>Data Types</b></td>');
  }
  if (show_spss == 1) {
    document.write('<td width="17%"><b>SPSS Data Files</b></td>');
  }
  if (show_word == 1) {
    document.write('<td width="12%"><b>Test Instructions and Stimuli</b></td>');
  }
  if (show_method == 1) {
    document.write('<td width="6%"><b>Method Descriptions and Sources</b></td>');
  }
  if (show_publ == 1) {
    document.write('<td width="6%"><b>Published articles</b></td>');
  }
  if (show_variable == 1) {
    document.write('<td width="6%"><b>Important variables</b></td>');
  }
</script>
```

KUVIO 19. Metadatataulukon otsikkorivin muodostaminen.

```
if (show_measure == 1) {
    document.write('<td width="30%">');
        document.write('<xsl:value-of select="method/dataColl/collMode"/>');
        document.write('<xsl:for-each select="method/notes[string(.)]">');
            document.write(' |<a><xsl:attribute name="href"><xsl:value-of select="."/></xsl:attribute>NOTE</a>| ');
        document.write('</xsl:for-each>');
    document.write('</td>');
}
if (show_data_type == 1) {
    document.write('<td width="12%">');
        document.write('<xsl:for-each select="stdyInfo/sumDscr/dataKind"><xsl:value-of select="."/><br/></xsl:for-each>');
    document.write('</td>');
}
```

KUVIO 20. Tutkimusmittarin ja tietotyypin sisältävien solujen muodostaminen.

XSLT-dokumentin avulla muodostetaan siis dynaamisesti HTML-runkodokumentin alaosaan HTML-tulosdokumentti, joka sisältää metadatataulukon. Taulukko muodostetaan siten, että kullekin tutkimuskerroissa esiintyvälle aihepiirille tehdään taulukkoon oma rivinsä. Yhden tutkimuskerran metatiedot saattavat siis tuottaa taulukkoon useamman kuin yhden rivin, mikäli kyseisen tutkimuskerran testeissä on tutkittu useampaa aihepiiriä. Lisäksi on huomioitava, että käyttäjän valitsemien sarakkeiden lisäksi taulukkoon tulostetaan aina tiedot tutkittavan lapsen iästä, ja että taulukon rivit lajitellaan lapsen iän mukaan nousevaan järjestykseen. Kuviossa 21 nähdään metadatasovelluksen avulla muodostettu taulukko. HTML-runkodokumentin yläosassa olevasta valintalomakkeesta (kts. Kuvio 18) on valittu näytettäväksi ainoastaan fonologiset taidot (Phonological skills) -tutkimusaihepiiriin liittyvä tutkimus. Näytettäviksi sarakkeiksi on valittu tutkimusaihepiiri (content areas), käytetty tutkimusmittari (content/measure) sekä tallennetut tietotyypit (data types). Lapsen ikä (age in years) näytetään taulukossa automaattisesti.

Age in Years	Content Areas	Content/Measure	Data Types
7.3	Phonological skills	Phonological awareness: 1) Initial phoneme production & deletion (asuu, suu), 2) Non-word repetition (18 items: sipa, kutta, tiippa, tati, kuuki ...)	Test, Cognitive Workshop
7.5	Phonological skills	Phonological awareness: 1) Phoneme blending (Poskiparta ym.: 10 items: p-u-u), 2) Syllable deletion (Poskiparta ym.: 10 items: paluu --> luu), 3) Initial phoneme production & deletion (Poskiparta ym.: 10 items: asuu --> suu), 4) Phonological Common Unit Task: Syllable & phoneme level (lehmä – lehti: noita - kukka)	Test, Cognitive Workshop
9.3	Phonological skills	Common Unit Task: Phoneme level (lauhkua-terike upittaa- homppe)	Cognitive Workshop

KUVIO 21. Esimerkki metadatasovelluksen avulla muodostetusta metadatataulukosta.

4.4 Toteutuksen evaluointi

Tässä kohdassa evaluoidaan toteutetun metadatasovelluksen toimintaa suhteessa LKK-tutkimuksen dokumentointitarpeisiin. Kohdassa 4.4.1 arvioidaan toteutuksen soveltuvuutta LKK-tutkimuksen aineiston dokumentointiin ja kohdassa 4.4.2 esitellään puolestaan DDI:n soveltamiseen liittyviä haasteita.

4.4.1 Ratkaisun soveltuvuus kohdeorganisaatioon

Toteutetun metadatasovelluksen avulla voitiin mahdollistaa metatietojen lomakepohjainen syöttö sekä niiden taulukkopohjainen haku ja selailu. DDI-spesifikaation kautta pystyttiin LKK-tutkimuksen dokumentointia myös kehittämään tavalla, joka tehostaa aineiston käyttöä. Lisäksi kehitettyyn DDI-pohjaiseen metadataskeemaan pystyttiin lisäämään kaikki LKK-tutkimuksen kannalta oleelliset dokumentointikohdat. Näiltä osin metadatasovelluksen kehittämisessä onnistuttiin, ja sen toimintaan oltiin kohdeorganisaatiossa tyytyväisiä.

Myös kohdassa 4.1.3 esitettyjä, LKK-tutkimuksen tutkimusaineiston dokumentointiin kohdistuvia tarpeita tarkastelemalla voidaan toteutettua sovellusta pitää onnistuneena. Sen avulla voidaan hahmottaa LKK-tutkimuksen pitkän elinkaareen aikana tehtyjä tutkimusvaiheita siten, että laajaa tutkimusta ja sen aikana kerättyä aineistoa voidaan käsitellä paremmin. Ennen kaikkea tutkimukselle ennestään vieraat henkilöt, kuten tutkimuksen parissa työskentelevät uudet tutkijat, hyötyvät suuresti syötetystä DDI-dokumentaatiosta ja toteutetuista metadatasovelluksista. Lisäksi LKK-tutkimuksen ja sen aineiston kuvailua vietiin merkittävin askelin eteenpäin siirtymällä tekstinkäsittelyohjelmalla tuotetusta listauksesta, hajautettuun ja rakenteiseen dokumentaatioon. Hajauttamisen myötä pystyttiin kunkin tutkimuskerran metatiedot tallentamaan huomattavasti aikaisempaa yksityiskohtaisemmin, joka vastasi dokumentointitarpeissa esiin tulleeseen vaatimukseen kuvailtavien kohtien lisäämisestä ja olemassa olevien tarkentamisesta. Rakenteisuuden kautta voitiin puolestaan muun muassa parantaa metatietoihin kohdistuvia hakutoimintoja sekä parantaa tietojen siirrettävyyttä eri järjestelmiin.

LKK-tutkimuksessa tehdyn dokumentointityön kautta voitiin myös selvittää spesifikaation tarjoamia soveltamismahdollisuuksia sekä erilaisia metadatan syöttö- ja selailuvaihtoehtoja. Metadatasovellus tuottikin tutkielman kannalta arvokasta tietoa, jota ilman ei olisi voitu arvioida DDI-spesifikaation soveltamista käytäntöön. Lisäksi sovelluksen kehityksen kautta pystyttiin LKK-tutkimuksen piirissä hahmottamaan konkreettisesti niitä hyötyjä, joita aineiston dokumentoinnin avulla voidaan saavuttaa ja lisäksi sovellus mahdollisti sen, että tavoitteet tuleville metadatahankkeille voitiin asettaa realistiselle tasolle. Toteutettua metadatasovellusta voidaankin pitää varsinaisen sovelluksen prototyyppinä, joka toimii pohjana tulevalle kehitystyölle. Sen avulla hahmoteltiin niitä vaatimuksia, joihin sen pohjalta mahdollisesti kehiteltävän pitkäkestoisemman metadatasovelluksen tulee pystyä vastaamaan.

Tutkielman yhteydessä toteutetun metadatasovelluksen käyttöä jatketaan luultavasti siten, että sen avulla kuvaillaan kaikkien eri tutkimuskertojen tiedot. Näin sovelluksen syöttöliittymän testaamista ja kehitystä voidaan jatkaa, ja kuvaillun aineiston määrän kasvaessa myös sovelluksen selailuliittymän toiminta tulee kunnolla arvioiduksi. Vaikka tulevaisuudessa käyttöön otettaisiin pidemmälle kehitelty metadatasovellus, ei metadatan XML-pohjaista syöttämistä prototyyppisovelluksen avulla voida pitää turhana työnä. Olemassa oleva metadata voidaan helposti muuntaa sellaiseen muotoon, että sitä voidaan jatkohyödyntää myös tulevien sovellusten käytössä. Tässä suhteessa XML-kielen laitteisto- ja ohjelmistoriippuvuutta voidaan pitää erittäin positiivisena tekijänä.

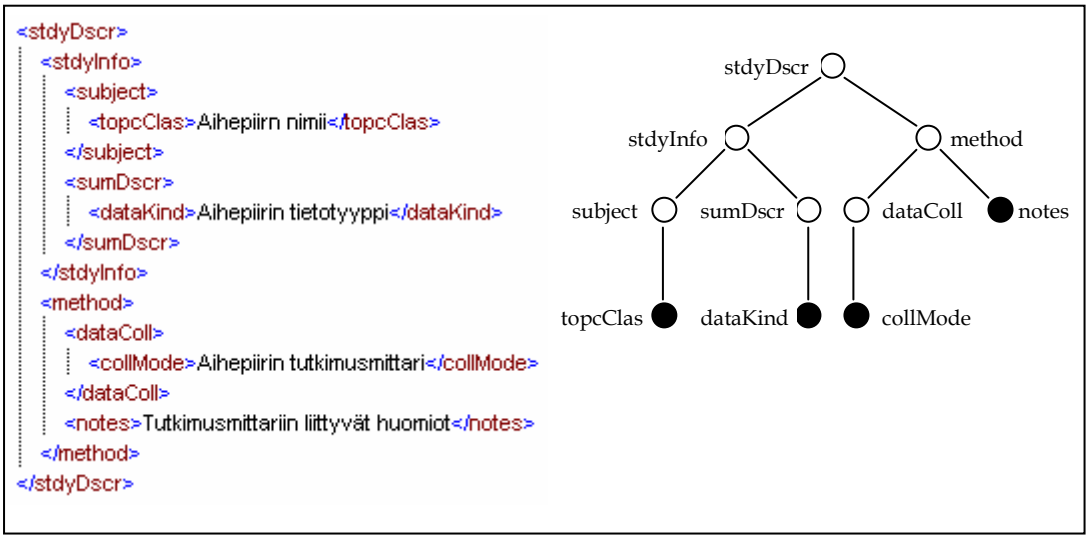
Kehitetyn metadatasovelluksen jatkokehitykseen vaikuttaa paljolti myös se, mihin suuntaan tutkimusaineistoa koskevan dokumentaation kehittämistä halutaan LKK-tutkimuksen parissa viedä. Jatketaanko mahdollisesti sillä linjalla, mihin tämän kokeilun yhteydessä lähdettiin, eli kehitetäänkö oma järjestelmä metatiedon syöttämistä ja selailua varten, jolloin lähtökohtana olisi ennen kaikkea metadatan hyödyntäminen LKK-tutkimuksen sisällä. Toinen mahdollinen kehityssuunta on se, että dokumentaation painopistettä viedään siihen suuntaan, että tutkimusaineistoa voidaan tallentaa Yhteiskuntatieteellisen tietoarkiston (FSD) tietokantoihin. Myös tutkimuksen rahalliset ja ajalliset resurssit määrittelevät pitkälti sen, kuinka paljon dokumentaation kehittämiseen voidaan tulevina vuosina satsata.

4.4.2 DDI:n soveltamisen haasteet

DDI-spesifikaatiota käsittelevässä luvussa (kts. 3.5) mainittiin viisi ongelmaa, jotka usein tulevat vastaan DDI-pohjaisia ratkaisuja kehittäessä. Seuraavaksi käsitellään näiden ongelmien ilmenemistä LKK:lle kehitetyn ratkaisun yhteydessä.

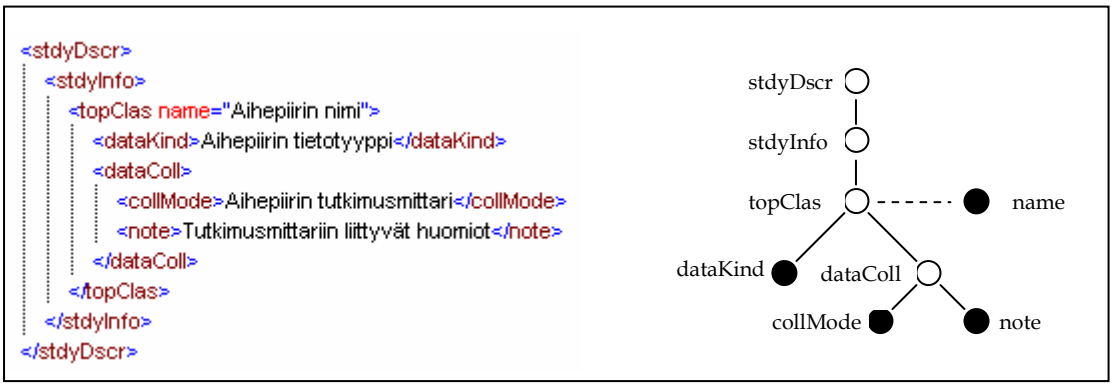
1) **Painottuneisuus/rajoittuneisuus kyselytutkimuksien dokumentointiin.** Jo tutkimuksen aikaisessa vaiheessa voitiin havaita, että DDI:n rajoittuneisuus kyselytutkimuksiin tulisi vastaan myös käsillä olevan kuvailun yhteydessä. DDI-spesifikaation rakennetta tarkasteltaessa voitiin huomata, että se on alun pitäen selvästi suunniteltu ennen kaikkea yksittäisten kyselytutkimusten dokumentointia varten. Tämä ilmeni muun muassa tutkimuksen kuvailuun tarkoitetusta osiosta, jossa ei ollut metatietokenttiä, jotka olisivat mahdollistaneet useita tutkimuskertoja sisältävään pitkittäistutkimukseen liittyvien tietojen tallennuksen. DDI-spesifikaation painottuneisuus kyselytutkimuksien dokumentointiin aiheutti LKK-kuvailun yhteydessä ongelmia lähinnä metadatarakenteen sopimattomuutena. Kyselytutkimuksille suunniteltua rakennetta jouduttiin sovittamaan ja paikoin venyttämään sopivaksi LKK-tutkimuksen dokumentointitarpeisiin.

2) **Modulaarisuuden puute**. Myös modulaarisuuden puute aiheutti vaikeuksia metadatasovelluksen kehittämisen aikana, etenkin sen alkuvaiheessa. DDI-spesifikaatio sisältää suuren määrän erilaisia kenttiä, joihin metatietoa voidaan tallentaa. LKK-tutkimuksen dokumentointitarpeita tarkastellessa kuitenkin huomattiin, että suurin osa näistä kentistä oli tutkimuksen kannalta tarpeettomia ja tarkoitettu täysin erilaisten tutkimusaineistojen kuvailuun. Nämä LKK-tutkimuksen kannalta tarpeettomat metatietokentät vaikeuttivat olennaisten kenttien löytämistä ja hyödyntämistä. Lisäksi LKK-tutkimuksen tutkimusaineiston kuvailun kannalta oleelliset kentät sijaitsivat hajallaan DDI-spesifikaation eri osa-alueissa, mikä aiheutti sen, että muodostettujen DDI-dokumenttien XML-puurakenne on monimutkainen ja epälooginen. Esimerkiksi yhden tutkimuskerran jokaisen tutkimusaihepiirin tiedot on tallennettu kuviossa 22 esitetyn puurakenteen mukaisesti.



KUVIO 22. Tutkimusaihepiirin kuvailuun käytetty XML-puurakenne.

Mikäli DDI-spesifikaatio mahdollistaisi modulaarisen rakenteen käytön, voitaisiin hajallaan olevat metadatakentät koota yhteen ja muodostaa niistä LKK-tutkimuksen dokumentointitarpeet huomioon ottaen mielekäs XML-puurakenne. Tämä rakenne voisi olla esimerkiksi kuviossa 23 esitetyn rakenteen mukainen.



KUVIO 23. Esimerkki rakenteeltaan yksikertaisemmasta ja loogisemmasta aihepiirin kuvailuun käytetystä XML-puurakenteesta.

LKK-tutkimuksen dokumentointitarpeiden pohjalta modulaarisesti muodostettu metadatarakenne helpottaisi myös metadatan syöttämiseen ja selailuun tarkoitettujen työkalujen kehittämistä. Juuri esimerkkikuvissa (kuvio 22 ja kuvio 23) käsitellyn tutkimusaihepiirin kuvailu tuotti modulaarisen rakenteen puuttumisen johdosta ongelmia. Yksi tutkimuskerta voi sisältää useita aihepiirejä ja lisäksi kuhunkin aihepiiriin liittyy useita tietotyyppejä

sekä tutkimusmittareita. Kuviossa 22 on esitetty DDI-spesifikaation mukainen rakenne, jota noudattaen siis myös LKK-tutkimuksen aineisto on kuvailtu. Kuvion oikeassa reunassa käytetty puurakenne on lisäksi havainnollistettu graafisesti siten, että tiedon tallennukseen käytetyt elementit on merkattu mustilla palloilla. Kuten kuviosta voidaan huomata, tässä rakenteessa tutkimusaihepiirin nimi ja siihen liittyvät tiedot on tallennettu XML-dokumentissa hajalleen erinäisten juurielementtien sisään. Tämä tekee tutkimusaihepiireihin liittyvien metatietojen käsittelystä erittäin hankalaa niin tietoja syöttäessä kuin niitä selaillessa. Esimerkiksi LKK-tutkimukseen liittyviä metatietoja kuvaillessa täytyi yhtä tutkimuskertaa koskevien useiden aihepiirien tiedot tallentaa toistamalla <stdyDscr> -elementtiä, jotta aihepiirit ja niihin liittyvät tietotyypit ja tutkimusmittarit saatiin yhdistetty toisiinsa oikein. Tästä puolestaan oli seurauksena se, että muut <stdyDscr> -juurielementin sisällä olevat tiedot toistuivat turhaan jokaisen aihepiirin kohdalla.

Kuviossa 23 on esitetty esimerkki siitä, kuinka DDI-spesifikaation modulaarinen rakenne voisi helpottaa tutkimusaihepiireihin liittyvien tietojen käsittelyä. Tämän hypoteettisen rakenteen avulla useat aihepiirit voitaisiin kuvailla toistamalla kunkin aihepiirin kohdalla <topClas> -elementtiä. Tämän elementin lapsielementit sisältäisivät puolestaan tiedot kuhunkin aihepiiriin liittyvistä tietotyypeistä ja tutkimusmittareista, ja näin ollen tiedot olisi liitetty toisiinsa yksinkertaisesti ja loogisesti. Muun muassa metatiedon selailuliittymän toteutuksessa käytetyn XSLT-muunnoskielen polkumääritykset muuttuisivat huomattavasti yksikertaisimmiksi, mikäli DDI-spesifikaatio sallisi modulaarisen rakenteen käytön.

3) **Kokoava lähestymistapa**. DDI-spesifikaatiosta puuttuu keinot, joiden avulla voisi kuvailla yleisempiä käsitteistöjä, jotka soveltuisivat useampien tutkimusten tai tutkimuskertojen kuvailuun (Ryssevik 2001). LKK-tutkimuksen yhteydessä tällaisia käsitteistöjä ovat muun muassa tutkimusmetodien, tut-

kimustiedostojen sekä muuttujatietojen kuvailut, jotka toistuvat usein samanlaisina useissa eri tutkimuskerroissa. Erityisesti koko LKK-tutkimusta koskevat yleiset tiedot, kuten tutkimuksen otsikko, kontaktihenkilöt ja tutkimuksen otos täytyi toistaa jokaisen tutkimuskerran kuvailun yhteydessä, koska yleisemmällä tasolla liikkuvien kuvausten tekeminen ei ole mahdollista. Myös monet käytetyistä muuttujakuvauksista ovat sellaisia, että ne toistuvat samanlaisina useissa eri tutkimuskerroissa koko pitkittäistutkimuksen elinkaaren ajan. Tällaisia muuttujakuvauksia ovat muun muassa tutkittavan lapsen sukupuoli, testausajankohta sekä kuvaukset, jotka liittyvät eri tutkimusaihepiirien tutkimisessa käytettäviin vuosittain toistettaviin kysely-/testausmuuttujiin. Kuvausten tekeminen yleisemmällä tasolla vähentäisi toistettavan metatiedon määrää ja leikkaisi näin osaltaan tutkimusaineiston dokumentointiin tarvittavaa työmäärää. Kokoava lähestymistapa ei aiheuttanut LKK-tutkimuksen kuvailun yhteydessä varsinaisia ongelmia, mutta yleisellä tasolla tehtävien kuvausten puuttuminen lisäsi tiedon kuvailuun kulutettua aikaa, ja vei näin ollen resursseja dokumentoinnin muilta osa-alueilta.

4) **Laajennettavuuden puute.** Kuten DDI-spesifikaation ongelmia käsittelevässä luvussa todettiin, spesifikaation mukaiseen kuvailuun ei voi lisätä omia laajennuksia niin että ne toimisivat yhdessä varsinaisen ydinspesifikaation kanssa. Spesifikaation rakenne täytyy siis hyväksyä sellaisenaan ilman lisäyksiä. (Ryssevik 2001) Kuten lähestulkoon kaikki dokumentoitavat aineistot, myös LKK:n aineisto asettaa sitä kuvailevalle dokumentointikielelle spesifejä vaatimuksia, joiden täyttäminen vaatii mahdollisuutta omien laajennuksien ja muutoksien tekemiseen. Laajennettavuuden puutteen takia jouduttiin tehdyn LKK-tutkimuksen dokumentoinnin yhteydessä usein soveltamaan DDI-spesifikaatiosta löytyviä metadatakenttiä, jotta ne saatiin sovitettua LKK:n tarpeisiin. Koska spesifikaatioon lukeutuvien elementtien joukosta ei aina löytynyt sellaista elementtiä, joka olisi annetun käyttötarkoituskuvauk-

sen perusteella täysin sopinut tietyn LKK-tutkimuksen osa-alueen dokumentointiin, jouduttiin elementtejä paikoin käyttämään vastoin alkuperäistä käyttötarkoitusta. Toisin sanoen kun tiettyyn spesifiin tarpeeseen ei löytynyt DDI:n elementtien joukosta oikeaa metadataelementtiä, jouduttiin dokumentointitarve tyydyttämään ikään kuin ”väärää” elementtiä hyödyntäen. Tästä esimerkkinä voidaan pitää tutkittavan lapsen iän dokumentoimista. LKK-tutkimuksessa lapsen iän dokumentoiminen on erittäin tärkeää, koska kussakin tutkimuskerrassa keskitytään tarkastelemaan tietyn ikäisiä lapsia, ja lisäksi samat tutkimukset suoritetaan samoille lapsille eri ikävaiheissa. DDI-spesifikaatiosta ei kuitenkaan löydy kenttää tutkimuksessa tarkasteltujen henkilöiden iän kuvaamiseen. Koska iän dokumentointi oli kuitenkin välttämätöntä, päätettiin se kuvailla DDI-spesifikaatiosta löytyvän <anlyUnit> -elementin avulla. DDI:n mukaan tämän elementin avulla on kuitenkin tarkoitus kuvata tutkimuksen analysointiyksikköä (esimerkiksi perhe, lapsi tai organisaatio), mutta koska paremmin soveltuvaa elementtiä ei ollut tarjolla, käytettiin sitä LKK-tutkimuksen dokumentaatiossa tutkittavan lapsen iän kuvailuun. Painottuneisuus kyselytutkimuksien dokumentointiin toimii pitkälti tekijänä, joka aiheuttaa DDI-spesifikaation osittaista sopimattomuutta LKK-tutkimuksen aineiston dokumentointikieleksi. Osaltaan tätä sopimattomuutta voidaan helpottaa soveltamalla ja venyttämällä spesifikaatiosta löytyviä metadatakenttiä, mutta mahdollisuus laajennuksien tekemiseen tekisi spesifikaatiosta varmasti helpommin sovitettavan erilaisten aineistojen dokumentointikieleksi.

5) **Suorituskykyongelmat.** Kohdassa 3.5 mainittiin, että DDI:n suorituskykyongelmat johtuvat yleensä spesifikaation massiivisesta rakenteesta sekä runsaasta sisäisten viittausten lukumäärästä. LKK-tutkimukselle tehdyssä sovellusratkaisussa suorituskykyongelmia aiheutti kuitenkin spesifikaation modulaarisuuden puute, minkä seurauksena toteutuksen XSLT-polkumäärityksistä muodostui monimutkaisia ja paljon suorituskykyä vaa-

tivia. Toinen sovelluksen toimintaa hidastavista seikoista oli sen rakenne, jossa kuhunkin tutkimuskertaan liittyvä kuvailu on tallennettu erilliseen dokumenttiin. Useiden dokumenttien sisältämien tietojen hakeminen ja käsittely vaati paljon suoritustehoa. LKK-toteutuksessa esiintyneet suorituskykyongelmat olivat kuitenkin erittäin lieviä ja niitä esiintyi ainoastaan vanhemmilla tietokoneilla.

Vaikka edellä mainitut vaikeudet tulivatkin kehityksen edessä vastaan aiheuttaen erisuuruisia vastoinkäymisiä, ei yhdestäkään niistä kuitenkaan koitunut ylitsepääsemätöntä estettä kokeilun läpiviennille. Ongelmatilanteita ilmeni ennen kaikkea silloin, kun DDI-spesifikaation mukaista dokumentaatiota piti muokata ja uudelleen sovittaa aina uusien dokumentointitarpeiden ilmetessä. Ongelmien esiintyminen teki spesifikaation käytöstä paikoin joustamatonta sekä työlästä, ja jätti sen avulla työskentelystä jäykän kuvan.

Muut kehitystyön yhteydessä vastaan tulleet haasteet liittyivät lähinnä valittujen teknologioiden ja ohjelmistojen käytössä ilmenneisiin vaikeuksiin. Esimerkiksi XML- ja XSLT-kielten käyttö yhdessä ohjelmointikielten kanssa aiheutti hankaluuksia. Internet-selaimilla oli muun muassa ongelmia käsitellä sellaisia dokumentteja, jotka sisälsivät sekä XSLT-muunnoskielellä että JavaScript-ohjelmointikielellä toteutettuja osioita. Lisätyötä aiheutti myös InfoPath-ohjelmistossa esiintyneet puutteet ja epäjohdonmukaisuudet. Suurin puutteista oli se, ettei InfoPath tukenut lomakkeen suunnittelun taustana olevaan rakennemääritykseen tehtäviä muutoksia. Kun LKK-tutkimuksen dokumentointiin käytettävien XML-dokumenttien rakennetta muutettiin esimerkiksi lisäämällä sinne jokin uusi elementti tai attribuutti, ei tätä muutosta voinut tehdä InfoPathiin yksinkertaisesti rakennemääritystä muuttamalla. Tämän sijasta metatietojen syöttämiseen tarkoitettu InfoPath-lomake piti rakentaa alusta asti uudelleen, jotta halutut muutokset pystyttiin toteuttamaan. Ohjelman epäjohdonmukaisuuksilla tarkoitetaan lähinnä sen lyhyestä elinkaaresta luultavasti johtuvia

ongelmia, jolloin ohjelman toiminnassa ilmenee käyttökerrasta riippuen arvaamatonta virheellistä toimintaa.

5 JOHTOPÄÄTÖKSET

Tässä luvussa paneudutaan johdannossa esitettyyn tutkimusongelmaan siitä, kuinka hyvin DDI-spesifikaatio pystyy vastaamaan niihin vaatimuksiin joita pitkittäistutkimus ja sen aineisto sille asettaa. Vastaus pohjautuu LKK-tutkimukseen kehitetystä ratkaisusta saatuihin kokemuksiin. Tässä kohdassa ei keskitytä tarkastelemaan niitä hyötyjä mitä tutkimusaineistojen kuvailulla voidaan yleisemmin saavuttaa. Sen sijaan arvioidaan sitä, mitkä seikat tukevat DDI-spesifikaation valintaa organisaation dokumentointikieleksi, ja toisaalta kuinka siinä havaitut heikkoudet tekevät spesifikaation valinnasta ongelmallisen. Tavoitteena on ennen kaikkea arvioida DDI-spesifikaation soveltuvuutta metadatan hallintavälineeksi pitkittäistutkimusta tekevissä tutkimusorganisaatioissa.

DDI-spesifikaation valintaa organisaation dokumentointikieleksi tukee ennen kaikkea se seikka, että sen avulla voidaan saavuttaa standardoitu tapa tallentaa tutkimusaineistoa koskevaa metatietoa. Käyttämällä standardoitua kuvailukieltä voidaan varmistua siitä, että eri tutkimusorganisaatiot tallentavat metatietoa saman ohjeiston mukaisesti. Tätä kautta useiden eri organisaatioiden keräämää metatietoa voidaan tarkastella ja vertailla keskenään suhteellisen helposti. Yhtenäinen metadatastandardi mahdollistaa myös eri tutkimusryhmien keräämien aineistojen kokoamisen erilaisiin tietokantoihin, kuten esimerkiksi kirjastoihin tai arkistoihin. Standardoidun kuvailukielen kautta voidaan varmistua myös siitä, että tutkimusaineisto tulee dokumentoiduksi tarkasti siten, että kaikki aineistoa koskevat oleelliset metatiedot tulee tallennetuksi. Tämä johtuu siitä, että sekä yleisesti hyväksytyn standardin asemaa tavoittelevat spesifikaatiot, kuten DDI, tai sen jo saavuttaneet, käytössä olevat standardit on kehitetty eri alojen asiantuntijoiden toimesta pitkän kehitysprosessien kuluessa. Standardia käyttämällä kukin tutkimusorganisaatio voi hyödyntää standardin sisältämää tietämystä, ja näin ollen säästää omia resurssejaan varsinaiseen tutkimustyöhön.

Haittapuolena standardien käytössä on se, että ne eivät sovi kaikille organisaatioille. Tämä ongelma tuntuu vaivaavan myös DDI-spesifikaatiota, ja syynä tähän voidaan pitää ennen kaikkea sen joustamattomuutta. DDI-spesifikaatio määrittelee erittäin tarkasti sen mitä tutkimusaineistoa koskevia seikkoja sen avulla voidaan tallentaa ja minkälaista rakennetta tallennetun metatiedon tulee noudattaa. Kullekin tutkimusaihepiirille tarkoitettujen spesifien metadatakenttien lisääminen tai omien rakennemääritysten tekeminen ei siis ole mahdollista DDI-spesifikaatiota hyödyntävissä tutkimusorganisaatioissa. Tätä voidaan pitää suurena heikkoutena sellaisen spesifikaation kohdalla, josta halutaan kehittää kansainvälinen standardi kaikkien yhteiskuntatieteellisten tutkimusten ja niissä syntyvän tutkimusaineiston dokumentointia varten.

LKK-tutkimukselle kehitetyn ratkaisun yhteydessä DDI-spesifikaation joustamattomuus aiheutti monenlaisia ongelmia, joita käsiteltiin tarkemmin kohdassa 4.4.2. Suurin osa näistä ongelmista aiheutui siitä, että LKK-tutkimus on luonteeltaan useista eri tutkimuskerroista muodostuva pitkittäistutkimus. DDI-spesifikaatio on puolestaan alun perin kehitetty yksittäisten kyselytutkimusten dokumentointiin. Vastaavanlainen ristiriita yhdistettynä spesifikaation joustamattomuuteen aiheuttaa todennäköisesti ongelmia myös muissa DDI:tä hyödyntävissä pitkittäistutkimuksissa. Ongelmat voivat koskea muun muassa dokumentoitavien metadatakenttien valintaa tai monimutkaiseksi muodostuneen dokumentointirakenteen hallintaa. LKK-tutkimuksen kuvailun ja kehitettyjen metadatasovellusten perusteella voidaan sanoa, että DDI-spesifikaation hyödyntäminen yhteiskuntatieteellisten pitkittäistutkimusten yhteydessä on erittäin työlästä, koska spesifikaatio on suunniteltu yksittäisten kyselytutkimusten tarpeiden pohjalta. Sen tarjoamat dokumentointimahdollisuudet eivät täysin kohtaa niitä tarpeita, joita pitkittäistutkimus asettaa sitä kuvaavalle kuvailukielelle. Vaillinaisuutta esiintyy esimerkiksi pitkittäistutkimuksen yleistietojen kuvailumahdollisuuksissa. Lisäksi puute omien muokkausten tai laajennusten tekemiseen aiheuttaa sen, ettei spesifikaatiota voida muunnella paremmin sopi-

vaksi kunkin pitkittäistutkimuksen tarpeisiin. LKK-tutkimuksen tutkimusaineistoa ja tallennetun metadatan käyttötarkoitusta silmällä pitäen, voidaan perustellusti sanoa, ettei DDI-spesifikaation käytöstä saatu sellaista hyötyä, jota standardin asemaa tavoittelevan spesifikaation avulla pyrittiin saavuttamaan. Monessa tilanteessa spesifikaation sopimattomuus tuntui aiheuttavan jopa enemmän työtä verrattuna siihen, että käyttötarkoitusta varten olisi kehitetty oma metadatarakenne.

DDI-spesifikaation joustamattomuudesta johtuvien ongelmien lisäksi sen valitsemiskynnystä tutkimusorganisaation kuvailukieleksi nostaa olemassa olevien, kehittyneiden metadatan syöttö- ja selailuratkaisuiden puute. DDI-allianssi on kehittänyt yhteistyössä Norjan, Iso-Britannian ja Tanskan tietoarkistojen kanssa Nesstar-ohjelmistopaketin, joka on esitelty tämän tutkielman kohdassa 3.6.2. Tämän ohjelmistopaketin tarkoituksena on tarjota DDI-spesifikaatiota käyttävien organisaatioiden käyttöön dokumentaation syöttämiseen ja hallintaan tarkoitetut työkalut. Nesstar-ohjelmisto ei kuitenkaan onnistu täysin täyttämään sille asetettuja tavoitteita, ja ainakin LKK-tutkimuksen yhteydessä sen toimissa voitiin havaita selviä puutteita. Nesstar-ohjelmiston vaillinaisuudet tehdyn kokeilun yhteydessä johtuvat pitkälti samasta syystä, mistä suuri osa koko DDI-spesifikaation ongelmista johtuu, eli painottuneisuudesta/rajoittuneisuudesta yksittäisten kyselytutkimusten kuvailuun. Toimivien metadatan syöttö- ja hallintaratkaisuiden puute heikentää siis osaltaan mielenkiintoa DDI-spesifikaatiota kohtaan, koska omien työkalujen kehittäminen on usein erittäin työlästä ja kallista. Tällöin myös metadatan tallentamisen ja hyödyntämisen ansiosta saavutettavien näkyvien tulosten aikaansaaminen vie pitkän ajan. Tiettyyn tarkoitukseen räätälöityjen ohjelmistotyökalujen positiivisena puolena voidaan pitää juuri niiden täydellistä sopivuutta kyseiselle sovellusalueelle.

Eräs DDI-spesifikaation käyttöönotossa askarruttava asia on sen hitaasti etenevä kehitystyö. Vaikkakin spesifikaation julkaistaan pieniä versiopäivityksiä aika ajoin, ei suurempia muutoksia/parannuksia sisältäviä päivityksiä ole julkaistu

moneen vuoteen. Spesifikaation tulevaisuuden kannalta on huolestuttavaa huomata, ettei sen kehittämisessä tapahdu jatkuvaa ja näkyvää edistystä, joka osaltaan edesauttaisi huomattavasti DDI-allianssin pyrkimyksissä luoda DDI-kuvailukielestä yleisesti hyväksytyn standardin omaava kuvailukieli. On myös mielenkiintoista huomata, että vaikka spesifikaatiota koskevat ongelmat (kts. kohta 3.5) on selvitetty sisäisten katselmusten sekä ulkoisesti teetettyjen arviointien avulla, ei näiden ongelmien ratkaisemiseksi ole ryhdytty vaadittaviin toimenpiteisiin. Ilmeisesti DDI-allianssi ei halua tehdä spesifikaation suurempia rakenteellisia muutoksia, koska sen toimintaan ollaan tyytyväisiä sitä tällä hetkellä hyödyntävissä tutkimusorganisaatioissa. Tästä syystä johtuen spesifikaation kehitys tuntuu kuitenkin junnaavan paikoillaan, eikä uusia merkittäviä yhteistyöorganisaatioita ole lähivuosina ilmaantunut. DDI-spesifikaation kotisivuilla esitettyjä esimerkkiprojekteja tarkastellessa herää myös kysymys, että onko DDI tarkoitettu enemminkin erilaisiin tietokantoihin arkistoivan tiedon hallintaan, vai voiko sitä hyödyntää yhtä tehokkaasti myös LKK-tutkimuksen tyyppisissä, vielä aktiivisesti jatkuvissa tutkimushankkeissa.

Kaiken kaikkiaan DDI-spesifikaation soveltuvuudesta pitkittäistutkimusten sekä niissä syntyvän aineiston dokumentoitiin, voitaisiin sanoa seuraavaa. Spesifikaation voidaan katsoa soveltuvan myös pitkittäistutkimusten kuvailuun, mutta sitä ei voida missään nimessä pitää kovin helposti käyttöönotettavana tai toimintalogiikaltaan erityisen hyvänä vaihtoehtona kuvailukieleksi. Tehdyn johtopäätökseen syntyyn vaikutti paljolti se, että LKK-tutkimuksen DDI-pohjaisen dokumentaation kehittämisessä onnistuttiin lopulta hyvin ja saavutettuihin tuloksiin oltiin tyytyväisiä. Spesifikaation soveltamisessa pitkittäistutkimuksien kuvailuun esiintyy kuitenkin vielä paljon sellaisia ongelmia, joihin spesifikaation kehittäjien pitäisi puuttua.

6 YHTEENVETO

Tämä tutkielma käsitteli tieteellisissä tutkimuksissa syntyvän tutkimustiedon ominaisuuksia ja erityispiirteitä, sekä sitä kuinka tutkimusaineistoja voidaan hallita metadatan avulla. Lisäksi tutkielmassa käsiteltiin yhteiskuntatieteellisten tutkimusaineiston dokumentointia varten kehitetyn DDI-spesifikaation ominaisuuksia ja sen tarjoamia dokumentointimahdollisuuksia. Tutkielman pääpaino oli kuitenkin konstruktiivisen tutkimuksen suorittamisessa. Tutkielmassa toteutetun DDI-pohjaisen ratkaisun avulla haluttiin tutkia sitä, kuinka DDI-muotoista dokumentaatiota voidaan hyödyntää kohdeympäristön sisällönhallinnassa mahdollisimman tehokkaasti. DDI-spesifikaation soveltaminen suoritettiin kehittämällä LKK-tutkimuksen käyttöön sen dokumentointitarpeita vastaava DDI-pohjainen metadataskeema, ja lisäksi metadatan syöttämistä sekä sen hakemista ja selaamista varten kehitettiin omat käyttöliittymät.

Tutkielman toisessa luvussa havaittiin, että tutkimustietoa syntyy pääasiallisesti tieteellisen tutkimuksen kautta, mutta sitä syntyy myös tutkimustoiminnan sivutuotteena. Lisäksi tutkimustiedon syntyyn vaikuttaa se minkälaisesta tutkimuksesta on kysymys sekä se millainen tallennustapa tutkimusta tekevällä tutkijalla on. Toinen näkökulma tutkimustiedon syntyyn oli Birnholtzin ja Bietzin (2003, 341–342) esittämä ajatus siitä, että tutkimustietoa voidaan kerätä joko pitkän ajanjakson tasaisena pysyvistä tietovirroista tai yksittäisistä, kertaalleen esiintyvistä tietotapahtumista. Tutkimustiedon katsottiin puolestaan koostuvan ilmiöitä tutkimalla saatavasta raakadatasta, eli mikrodatasta, siitä analysoitavasta makrodatasta sekä näitä kuvailevasta metadatasta. Tallennetun tutkimustiedon erityispiirteitä olivat sen pitkäikäisyys, monimuotoisuus sekä maksuttomuus. Tieteellisen toiminnan kannalta tärkeintä on se, että tutkimustieto tallennetaan ja kuvataan mahdollisimman tarkasti, jotta siitä olisi hyötyä muille tutkijayhteisön tutkijoille sekä tulevaisuuden tutkijoille. Tieteellisen toiminnan perusedellytyksenä pidetään myös tutkimustiedon jakamista, koska usein oman tutkimuksen suorittaminen on mahdollista vain toisilta tutkijoilta saadun

tiedon avulla, ja myös omien tutkimustulosten tieteellinen validointi tapahtuu tarjoamalla ne muiden tutkijoiden käytettäviksi ja arvioitaviksi. Tutkimustiedon jakamisessa ilmenee kuitenkin usein muun muassa taloudellisista intresseistä tai tieteellisestä kilpailusta johtuvia ongelmia.

Kuten edellisessä kappaleessa mainittiin, metadatalla tarkoitettiin sellaista dataa, jonka avulla kuvaillaan tutkimuksissa syntyvää mikro- ja makrodataa. Metadatan avulla tutkimustiedon hallintaa ja käytettävyyttä voidaan oleellisesti parantaa. Metadata mahdollistaa kaikkien tutkimukseen liittyvien ja siihen vaikuttavien tekijöiden kuvailun, ja lisäksi metadatan avulla voidaan mahdollistaa tutkimustiedon jakaminen sekä sen hyödyntäminen myös tulevaisuudessa. Tutkimustietoon liittyvää metadataa tallennetaan yleensä viidellä eri tasolla, joita ovat tutkimustaso, tiedostotaso, muuttujataso, hallinnollinen metadata sekä meta-metadata. Tämän lisäksi metadata voidaan tallentaa osaksi kuvattavan dokumentin sisältöä tai vaihtoehtoisesti se voidaan tallentaa myös erilleen kuvattavasta dokumentista. Metadatan tallentamisen ja ylläpitämisen merkitys on korostunut entisestään sen myötä kun tutkijayhteisöt ovat levittäytyneet maantieteellisesti yhä laajemmalle alueelle, jolloin tarvittava tutkimustieto löydetään pääasiassa juuri metadatan avulla. Vaikka metadatan merkitys tunnustetaankin sen käyttöönottoa jarruttaa usein kuitenkin siitä aiheutuvat kustannukset, oikean metadatarakenteen löytäminen sekä metadatan luomisessa ja käytössä vastaan tulevat ongelmat.

Tutkielman kolmannessa luvussa siirryttiin tarkastelemaan metadataspesifikaatiota, joka kantaa nimeä Data Documentation Initiative (DDI). DDI-spesifikaatiosta pyritään kehittämään kansainvälinen XML (Extensible Markup Language) -pohjainen standardi yhteiskuntatieteellisten tutkimusten ja niissä syntyvän tutkimusaineiston dokumentointia varten. DDI:n avulla tuotetun dokumentaation tarkoituksena on mahdollistaa tutkimusaineiston tehokas ja kokonaisvaltainen hyödyntäminen, ja lisäksi sen avulla halutaan korvata teknisesti vanhentuneet metadatastandardit. DDI-spesifikaation on siis tarkoitus mah-

dollistaa edellisessä kappaleessa mainitut, tallennetun metadatan avulla tavoiteltavat hyödyt. DDI-spesifikaatio mukainen dokumentaatio on toteutettu rakenteisella XML-kielellä, jonka avulla jokainen yksittäinen metatiedon osa voidaan kuvailla oman elementin tai elementtiryhmän sekä attribuuttien avulla. DDI-spesifikaatio jakautuu viiteen osa-alueeseen, jotka sisältävät yhteensä noin 300 eri kohtaa, joihin metatietoa voidaan tallentaa. Spesifikaation osa-alueita ovat: 1) Metadatadokumentin kuvailu, 2) Tutkimuksen kuvailu, 3) Tiedostojen kuvailu, 4) Muuttujien kuvailu ja 5) Muu tutkimukseen liittyvä materiaali. Vaikka spesifikaation mukainen dokumentaatio onkin otettu onnistuneesti käyttöön useissa yhteiskuntatieteellistä tutkimusta tekevissä, sekä tietoa arkistoivissa organisaatioissa, liittyy sen käyttöön kuitenkin joitain melko suuria ongelmia, jotka esiteltiin kohdassa 3.5.

Tutkielman neljännessä luvussa esiteltiin DDI-spesifikaation soveltamista LKK-tutkimuksen aineiston dokumentointiin. Kohdeympäristön dokumentaatiota lähdettiin kehittämään tarkastelemalla sen sisällönhallintaa siinä esiintyvien toimintojen, toimijoiden, sisältöyksiköiden, järjestelmien sekä näiden välillä esiintyvien tietovirtojen näkökulmasta. Lisäksi selvitettiin ne tarpeet, joita tutkimusaineiston dokumentointiin kohdistuu kohdeorganisaatiossa. Näiden tarpeiden pohjalta kehitettiin LKK-tutkimuksen käyttöön oma DDI-pohjainen metadataskeema, jota hyödyntäen osa tutkimuksen aineistosta dokumentoitiin. Lisäksi dokumentointitarpeiden pohjalta kehiteltiin sovellusratkaisu, joka koostuu metadatan syöttöön tarkoitetusta lomakepohjaisesta käyttöliittymästä sekä syötetyn metatiedon hakemiseen tarkoitetusta taulukkopohjaisesta käyttöliittymästä.

Luvussa 5 arvioitiin puolestaan konstruktiivisen tutkimuksen kautta saatujen tietojen perusteella sitä, kuinka hyvin DDI-spesifikaation voidaan katsoa soveltuvan pitkittäistutkimusten dokumentointiin yleisemmällä tasolla. Arviointi suoritettiin puntaroimalla keskenään DDI:n soveltamisen aikana vastaan tullei-

ta, spesifikaation valinnan puolesta puhuvia seikkoja, sekä sen käytössä havaittuja heikkouksia.

Tutkielman tuloksena LKK-tutkimuksen dokumentointitarpeet saatiin selvitetyksi sekä sen sisällönhallinta analysoiduksi. Keskeisimpinä tuloksina voidaan kuitenkin pitää kehiteltyä LKK-metadataskeemaa sekä sitä hyödyntäen tuotettua DDI-ratkaisua. Tulosten kautta LKK-tutkimuksen dokumentointia pystyttiin kehittämään merkittävästä ja lisäksi toteutetut sovellukset mahdollistavat DDI-pohjaisen dokumentaation syöttämisen, ylläpidon sekä hyödyntämisen tutkimuksen piirissä myös jatkossa.

Tutkielman tuloksina voidaan pitää myös DDI-spesifikaation yleisestä soveltuvuudesta tehtyä arviointia sekä tutkimustietoon ja sen dokumentoitiin liittyvien seikkojen selvittämistä sekä niiden avaamista lukijalle. Lisäksi tutkielma antaa kattavan tietopaketin yhteiskuntatieteellisten tutkimusten dokumentointiin tarkoitetusta DDI-spesifikaatiosta, selvittämällä sen tarjoamat dokumentointimahdollisuudet.

Tutkielman tuloksista voidaan katsoa olevan hyötyä tietojärjestelmätieteen tutkimukselle, koska se käsittelee spesifin sovellusalueen (pitkittäistutkimus) dokumentointiin liittyviä kysymyksiä, ja näin ollen laajentaa sekä tarkentaa aineiston dokumentointiin liittyvää tietämystä. Tutkielma kokoaa myös yhteen useista lähteistä peräisin olevaa, DDI-spesifikaatiota koskevaa tietoa, minkä ansiosta siitä kiinnostuneet henkilöt voivat saada tarkkaa suomenkielistä tietoa muun muassa spesifikaation osa-alueista. Käytännön työn kannalta tuloksista on hyötyä ennen kaikkea LKK-tutkimuksella, jossa uusi metatietojärjestelmä otettiin käyttöön. Käytännön tutkimustyön kannalta tutkielman tuloksista on varmasti myös hyötyä lukuisille muille yhteiskuntatieteellistä tutkimusta tekeville tutkimusorganisaatiolle, jotka painivat LKK-tutkimuksen kanssa samankaltaisten dokumentointiongelmien kanssa.

Mielenkiintoinen jatkotutkimusaihe voisi olla vertailla DDI-spesifikaatiota muihin olemassa oleviin metadatastandardeihin tarkastellen sitä, voitaisiinko LKK-tutkimuksen aineiston dokumentointia helpottaa/tehostaa muiden standardien avulla. Lisäksi toteutettua DDI-ratkaisua voitaisiin viedä eteenpäin kuvailemalla koko LKK-tutkimuksen aineisto spesifikaation avulla, ja ottamalla se osaksi jokapäiväistä tutkimustyötä. Tämä toisi DDI:n käyttöön lisäsyvyyttä antaen paremman kuvan sen kaikista dokumentointimahdollisuuksista.

LÄHDELUETTELO

Beedham H. 2004. The role of metadata in a major European data archive. Teoksessa Proceedings of the Czech Statistical Office International Conference Prague, Czech Rebuplic, September 6-7 [viitattu 16.11.2004]. Saatavilla www-muodossa <http://www.czso.cz/sif/conference2004.nsf/i/the_role_of_metadata_in_a_major_european_data_archive>.

Birnholtz J. & Bietz M. 2003. Data at work: supporting sharing in science and engineering. Teoksessa Proceedings of the 2003 international ACM SIGGROUP conference on supporting group work Sanibel Island, Florida, November 9-12. New York: ACM Press, 339-348.

Bose R. 2002. A conceptual framework for composing and managing scientific data lineage. Teoksessa D. Bren Proceedings of the 14th international Conference on Scientific and Statistical Database Management Santa Barbara, California, July 24-26. Los Alamitos: IEEE Computer Society, 15-19.

Counting California Project. 2005. About Counting California. [online]. [Viitattu 8.3.2005]. Saatavilla www-muodossa <http://countingcalifornia.cdlib.org/about.html>.

Cruse P., Einowski I. & Stratford J. 2002. Applications in the real world: the counting California experience with the DDI. IASSIST Quarterly 25(1), 10-12.

Data Documentation Initiative. 2005a. About the specification [online]. [Viitattu 25.1.2005]. Saatavilla www-muodossa <http://www.icpsr.umich.edu/DDI/codebook/index.html>.

Data Documentation Initiative. 2005b. DDI Tag Library [online]. [Viitattu 25.1.2005]. Saatavilla www-muodossa <http://www.icpsr.umich.edu/DDI/users/dtd/index.html#a03>

Data Documentation Initiative. 2005c. Getting started with the DDI [online]. [Viitattu 4.3.2005]. Saatavilla www-muodossa <http://www.icpsr.umich.edu/DDI/users/intro-use.html>

Eklund K. 2005. Sähköpostiviesti koskien LKK-tutkimuksen tulevaisuutta. Psykocenter, Jyväskylä.

Jacobs J. & Humphrey C. 2004. Preserving research data. Communications of the ACM 47(9), 27-29.

Leighton V. 2002. Developing a new data archive in a time of maturing standards. IASSIST Quarterly 26(1), 5-9.

Marshall C. 1998. Making metadata: a study of metadata creation for a mixed physical-digital collection. Teoksessa Proceedings of the third ACM conference on Digital libraries Pittsburgh, Pennsylvania, June 23–26. New York: ACM Press, 162-171.

Norwegian Social Science Data Services. 1999. Providing global access to distributed data through metadata standardisation: the parallel stories of NESSTAR and the DDI. Working Paper No. 10, UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, September 22-24.

Psykocenter. 1998. Appendix II: CoreII. Humandevelopment and its risk factors: Congnitive development. Jyväskylä: University of Jyväskylä.

Psykocenter. 2003. Tutkimusryhmät [online]. [Viitattu 17.5.2005]. Saatavilla www-muodossa <http://www.jyu.fi/agora/psykocenter/docs/PSC-vihkoesite2003.pdf>

Ryssevik J. 2001. The Data Documentation Initiative (DDI) metadata specification [online]. [viitattu 9.12.2004]. Saatavilla www-muodossa <http://www.icpsr.umich.edu/DDI/papers/ryssevik.pdf>.

Ryssevik J. & Musgrave S. 2001. The social science dream machine: resource discovery, analysis and delivery on the web. Social Science Computer Review 19(2), 163-174.

Salminen A. 2003. Document analysis methods. Teoksessa C.L. Bernie (toim.) Encyclopedia of Library and Information Science, Second Edition, Revised and Expanded. New York: Marcel Dekker, 916-927.

Salminen A. 2005a. XML-kieli, Jakso 1: XML pähkinänkuoressa. Jyväskylä: Jyväskylän yliopisto, Tietojenkäsittelytieteiden laitos [viitattu 23.2.2005]. Saatavilla www-muodossa <http://www.cs.jyu.fi/~airi/opetus/xml/xml-kieli/xml-jakso1-100205.pdf>.

Salminen A. 2005b. Metatiedot organisaatioiden sisällönhallinnassa. Julkaisussa Lehtinen A., Salminen A. & Nurmeksela R. Metatiedot suomalaisen lainsäädäntöprosessin tiedonhallinnassa. RASKE2-projektin II väliraportti, Eduskunnan kanslian julkaisu 7/2005.

Thomas W. & Ryssevik J. 2003. Multidimensional table description [online]. [Viitattu 3.3.2005]. Saatavilla www-muodossa <http://ils.unc.edu/%7Eohjs/stats/tutorial/nCube.pdf>.

Thomson J., Adams D., Cowley P. & Walker K. 2003. Metadata's role in a scientific archive. IEEE Computer 36(12), 27-34.

Toivonen H., Salmenkivi M. & Verkamo I. 2004. Luentomateriaali: Tutkimustiedonhallinnan peruskurssi. Helsinki: Helsingin yliopisto, Tietojenkäsittelytieteen laitos [viitattu 15.11.2004]. Saatavilla www-

muodossa
<http://www.cs.helsinki.fi/u/htoivone/teaching/tutihaK04/>.

W3C. 1999. XSL Transformations (XSLT) Version 1.0, W3C Recommendation 16 November 1999 [online]. [viitattu 24.5.2005]. Saatavilla www-muodossa <http://www.w3.org/TR/1999/REC-xslt-19991116 >

W3C. 2004a. Extensible Markup Language (XML) 1.0 (Third Edition), W3C Recommendation 04 February 2004 [online]. [viitattu 9.12.2004]. Saatavilla www-muodossa <http://www.w3.org/TR/2004/REC-xml-20040204/>.

W3C. 2004b. XML Schema Part 0: Primer (Second Edition), W3C Recommendation, 28 October 2004 [online]. [viitattu 23.2.2005]. Saatavilla www-muodossa <http://www.w3.org/TR/2004/REC-xmlschema-0-20041028/>.

Yang R., Kafatos M. & Wang X.S. 2002. Managing scientific matadata using XML. IEEE Internet computing 6(4), 52-59.

Yau H. & Hawker S. 2004. SA_MetaMatch: relevant document discovery through document metadata and indexing. Teoksessa Proceedings of the ACM 42nd annual Southeast regional conference Huntsville, Alabama, April 2-3. New York: ACM Press, 385-390.

LIITE 1

Alla olevassa luettelossa on esitetty DDI-spesifikaation mukaisessa dokumentaatiossa käytettävät elementit ja attribuutit sekä niiden noudattama rakenne. Kuvaus on tehty DDI-hankkeen Internetsivuilta löytyvän sanastokirjaston pohjalta (Data Documentation Initiative 2005b). Luetteloa tarkastellessa on syytä huomioida, että:

- Attribuuteille annetaan selitys vain ensimmäisen esiintymiskerran yhteydessä
- Kunkin elementin yhteydessä on mainittu vain ne attribuutit, jotka esiintyvät ainoastaan kyseisen elementin kohdalla. Jokaisen DDI-spesifikaation mukaisen elementin yhteydessä voidaan lisäksi käyttää attribuutteja ID (elementin identifioiva numero- tai kirjainsarja), xml:lang (elementin sisällössä käytetty kieli) ja source (erottelemaan tutkimusaineiston tuottajalta ja arkistoijalta saatuja tietoja).

codeBook (0.0) - Juurielementti koko DDI-dokumentaatiolle.
.
.
.
docDscr (1.0) - Metadatadokumentin kuvailu.
|
|-------- **citation** (1.1) - Bibliografisen tiedon kuvaaminen. Samaa elementtiä lapsielementteineen ja attribuutteineen käytetään myös myöhemmissä dokumentaation vaiheissa. Attribuutit: MARCURI (linkitys MARC -luetteloihin).
| |
| |-------- **titlStmt** (1.1.1) - Otsikkotietojen kuvailu.
| | |-------- **titl** (1.1.1.1) - Otsikko
| | |-------- **subTitl** (1.1.1.2) - Lisäotsikko
| | |-------- **altTitl** (1.1.1.3) - Vaihtoehtoinen otsikko
| | |-------- **parTitl** (1.1.1.3) - Otsikko toisella kielellä
| | +-------- **IDNo** (1.1.1.4) - Yksilöivä ID-numero. Attribuutit: agency (taho) ja level (taso).
| |
| |-------- **rspStmt** (1.1.2) - Vastuutietojen kuvailu.
| | |-------- **authEnty** (1.1.2.1) - DDI-dokumentaation sisällön syöttämisestä vastuussa oleva henkilö, taho tai yksikkö. Attribuutit: affiliation (yhteys, esim. yritys tai yksikkö, johon dokumentaation syöttänyt henkilö kuuluu).
| | +-------- **othId** (1.1.2.2) - Muut DDI-dokumentaation tuottamiseen osallistuneet henkilöt tai tahot, joita ei ole mainittu edellisessä kohdassa. Attribuutit: affiliation, type (tyyppi) ja role (rooli).
| |
| |-------- **prodStmt** (1.1.3) - Tuotantotietojen kuvailu.
| | |-------- **producer** (1.1.3.1) - Tuottajalla tarkoitetaan henkilöä tai organisaatiota, joka on taloudellisessa ja hallinnollisessa vastuussa dokumentaation tuottamisesta. Attribuutit: affiliation, abbr (lyhenne) ja role.
| | |-------- **copyright** (1.1.3.2) - Copyright
| | |-------- **prodDate** (1.1.3.3) - Tuotantopäivämäärät. Atribuutit: date (päivämäärän merkkaus muodossa VVVV-KK-PP).
| | |-------- **prodPlac** (1.1.3.4) - Tuotantopaikat.
| | |-------- **software** (1.1.3.5) - Käytetyt ohjelmistot. Attribuutit: date ja version (versio).
| | |-------- **fundAg** (1.1.3.6) - Mahdolliset rahoittajat. Attribuutit: abbr ja role.
| | +-------- **grantNo** (1.1.3.7) - Mahdolliset apurahat. Attribuutit: agency ja role.
| |
| |-------- **distStmt** (1.1.4) - Jakelutietojen kuvailu.
| | |-------- **distrbtr** (1.1.4.1) - Jakelijalla tarkoitetaan tahoa, jonka tehtävänä on luoda ja jakaa

| | | kopioita dokumentaatiosta. Attribuutit: abbr, affiliation ja URI (verkko-osoite).
| | |-------- **contact** (1.1.4.2) - Yhteyshenkilö. Attribuutit: affiliation, URI ja email (sähköpostiosoite).
| | |-------- **depositr** (1.1.4.3) -Taho, joka luovutti dokumentaation sitä varastoivalle arkistolle. Attribuutit: abbr ja affiliation.
| | |-------- **depDate** (1.1.4.4) - Arkistointipäivämäärä. Attribuutit: date.
| | +-------- **distDate** (1.1.4.5) - Päivämäärä, jolloin dokumentaatio luovutettiin jakeluun. Attribuutit: date.
| |
| |-------- **serStmt** (1.1.5) - Sarjatietojen kuvailu. Attribuutit: URI.
| | |-------- **serName** (1.1.5.1) - Dokumentaatiosarjan nimi. Attribuutit: abbr.
| | +-------- **serInfo** (1.1.5.2) - Tiedot dokumentaatiosarjasta.
| |
| |-------- **verStmt** (1.1.6) - Versiotietojen kuvailu.
| | |-------- **version** (1.1.6.1) - Versionumero. Attribuutit: date ja type.
| | |-------- **verResp** (1.1.6.2) - Kyseisestä versiosta vastuussa oleva taho. Attribuutit: affiliation.
| | +-------- **notes** (1.1.6.3) - Versioon liittyviä muita huomioita. Attribuutit: type, subject (aihe), level ja resp (huomion laatija).
| |
| |-------- **biblCit** (1.1.7) - Bibliografinen esimerkkiviittaus, jonka avulla dokumentaatioon voidaan viitata. Attribuutit: format (viittauksen muoto/tyyppi).
| |-------- **holdings** (1.1.8) - Tiedot dokumentaation kuvailemasta fyysisestä tai sähköisestä tietokokoelmasta. Attribuutit: location (fyysinen tallennussijainti), callno (fyysisen kokoelman tunnusnumero), URI, media (sähköinen tallennusmedia).
| +-------- **notes** (1.1.9) - Bibliografista informaatiota koskevia muita huomioita. Attribuutit: type, subject, level ja resp.
|
|-------- **guide** (1.2) - Tietoja termeistä ja määrityksistä, joita dokumentaatiossa käytetään. Kohdan tarkoituksena toimia oppaana dokumentaation käytölle
|
|-------- **docStatus** (1.3) - Metadatadokumentin käsittelystatus
|
|-------- **docSrc** (1.4) - Mahdollisen lähdedokumentin kuvailulle. Sisältää elementit titlStmt, rspStmt, prodStmt, distStmt, serStmt, verStmt, biblCit, holdings ja notes sekä niiden lapsielementit ja attribuutit. (Tarkempi rakenne kts. kohta 1.1, citation). Attribuutit: MARCURI.
|
+-------- **notes** (1.5) - Metadatadokumentin kuvailua koskevia muita huomioita. Attribuutit: type, subject, level ja resp.

.
.
.

stdyDscr (2.0) - Tutkimuksen kuvailu.
|
|-------- **citation** (2.1) - Tutkimukseen liittyvän bibliografisen tiedon kuvailu. Elementit titlStmt, rspStmt, prodStmt, distStmt, serStmt, verStmt, biblCit, holdings ja notes sekä niiden lapsielementit ja attribuutit. (Tarkempi rakenne kts. kohta 1.1, citation). Attribuutit: MARCURI. Erotuksena kohtaan 1.1. tässä käsitellään otsikko-, vastuu-, tuotanto-, sarja- sekä versiotietojen osalta dokumentaatioon liittyvien tietojen sijasta tutkimukseen liittyviä tietoja. Kohtien 1.1 ja 2.1 sisältö on usein identtinen tapauksissa, jossa tutkimuksen suorittanut organisaatio on tuottanut myös siihen liittyvän dokumentaation.
|
|-------- **stdyInfo** (2.2) - Tutkimuksen ulottuvuuksien kuvailu.
| |
| |-------- **subject** (2.2.1) - Tutkimuksen aihepiiriin liittyvien tietojen kuvailu.

| | |-------- **keyword** (2.2.1.1) - Tutkimuksen avainsanat. Attribuutit: vocab (kontrolloitujen
| | | avainsanastojen luomista/käyttöä varten) ja vocabURI (verkko-osoite, josta
| | | kontrolloitu avainsanasto on saatavilla).
| | +-------- **topcClas** (2.2.1.2) - Aihepiirin luokittelu kontrolloitujen luokittelukirjastojen
| | avulla (Esim. Yhdysvaltain kongressin kirjaston luokittelukirjasto). Attribuutit:
| | vocab ja vocabURI.
| |
| |-------- **abstract** (2.2.2) - Tutkimuksen tiivistelmä. Tiivistelmässä voidaan kuvailla mm. tutki-
| | muksen tarkoitus, luonne, tiedon keräyksen laajuus, erityspiirteet, aihepiiri sekä tutki-
| | muskysymykset. Attribuutit: date.
| |
| |-------- **sumDscr** (2.2.3) - Tutkimuksen ajallisen ja maantieteellisen kattavuuden sekä analysoin-
| | tiyksikön kuvailu.
| | |-------- **timeprd** (2.2.3.1) - Tutkimusaineiston kattava aikaväli. Attribuutit: date, event
| | | (halutaanko ilmaista alkupvm, loppupvm vai yksittäinen pvm) ja cycle (tutki-
| | | mussykli).
| | |-------- **collDate** (2.2.3.2) - Tutkimusaineiston keräyspäivämäärät. Attribuutit: date,
| | | event ja cycle.
| | |-------- **nation** (2.2.3.3) - Keräysmaa. Attribuutit: abbr (maan lyhenne).
| | |-------- **geogCover** (2.2.3.4) - Maantieteellinen kattavuus (esim. Keski-Suomi).
| | |-------- **geogUnit** (2.2.3.5) - Maantieteellinen yksikkö (esim. Kunta).
| | |-------- **geoBndBox** (2.2.3.6) - Elementin avulla voidaan ilmaista tutkimuksen suorakul-
| | | mion muotoinen maantiet. kattavuus käyttämällä pituus- ja leveysasteita.
| | |-------- **boundPoly** (2.2.3.7) - Monikulmion muotoinen maantieteellinen kattavuus.
| | |-------- **anlyUnit** (2.2.3.8) - Analysointiyksikkö. Attribuutit: unit (kontrolloidun sanaston
| | | käyttöä varten).
| | |-------- **universe** (2.2.3.9) - Tutkimusryhmä. Attribuutit: level (tutkimusryhmien tasot) ja
| | | clusion (ryhmään kuuluvien ja siitä ulosluettujen henkilöiden erottelu).
| | +-------- **dataKind** (2.2.3.10) - Tutkimusdatan tyyppi (esim. Kyselytulos).
| |
| +-------- **notes** (2.2.4) - Muita tutkimuksen ulottuvuutta koskevia huomioita Attribuutit: type,
| subject, level ja resp.
|
|-------- **method** (2.3) - Tutkimusmetodien kuvailu.
| |
| |-------- **dataColl** (2.3.1) - Tiedon keräyksessä käytetyn metodin kuvailu.
| | |-------- **timeMeth** (2.3.1.1) - Käytetty aikametodi. Attribuutit: method (kontrolloidun
| | | metodisanaston käytölle).
| | |-------- **dataCollector** (2.3.1.2) - Taho, joka keräsi tutkimusdatan esim. pitämällä haastat-
| | | telun. Attribuutit: abbr ja affiliation.
| | |-------- **frequenc** (2.3.1.3) - Tutkimusdata keräyksen tiheys. Attribuutit: frec (kontrol-
| | | loidun sanaston käytölle).
| | |-------- **sampProc** (2.3.1.4) - Tutkimusotos ja sen valintaprosessi.
| | |-------- **deviat** (2.3.1.5) - Tietoa siitä, kuinka hyvin otos vastaa sitä ryhmää, jota sen on
| | | tarkoitus kuvata (perusjoukko).
| | |-------- **collMode** (2.3.1.6) - Tiedon keräyksessä käytetty metodi (esim. tietokoneavusteinen haastattelu).
| | |-------- **resInstru** (2.3.1.7) - Tutkimuksessa käytetty kyselymenetelmä. Rakenteinen (kaikille samat kysymykset ja vastausvaihtoehdot), puolirakenteinen (avoimet kysymykset) ja ei-rakenteinen (syväluotaava haastettelu). Attribuutit: type (kontrolloidun sanaston käytölle).
| | |-------- **sources** (2.3.1.8) - Tiedon keräysmenetelmään liittyvät tietolähteet.
| | | Sisältää viisi lapsielementtiä jokaisen lähteen tarkemmalle kuvailulle.
| | |-------- **collSitu** (2.3.1.9) - Tiedon keräystilanteen kuvailu.
| | |-------- **actMin** (2.3.1.10) - Toimenpiteet tiedon hukkaantumisen minimoimiseksi.
| | |-------- **ConOps** (2.3.1.11) - Kerätylle tiedolle suoritetut kontrollointitoimenpiteet. Attri-

| | | buutit: agency.
| | |-------- **weight** (2.3.1.12) - Käytetyt painomuuttujat.
| | +-------- **cleanOps** (2.3.1.13) - Kerätylle tiedolle suoritetut puhdistusoperaatiot. Attribuutit: agency.
| |
| |-------- **notes** (2.3.2) - Muita tutkimusmetodia koskevia huomioita Attribuutit: type, subject, level ja resp.
| |
| |-------- **anlyInfo** (2.3.3) - Kerätyn tiedon arviointi.
| | |-------- **respRate** (2.3.3.1) - Vastausprosentti.
| | |-------- **EstSmpErr** (2.3.3.2) - Arvio siitä kuinka hyvin otoksesta kerätyt tiedot kuvaavat perusjoukkoa.
| | +-------- **dataAppr** (2.3.3.3) - Muu kerätyn tiedon arviointi, kuten vastaamatta jättäneiden määrä ja saatujen vastausten vaihtelevuus.
| |
| +-------- **stdyClas** (2.3.4) - Tutkimuksen luokka ja mahdollinen statusnumero. Attribuutit: type (kontrolloidun sanaston käytölle).
|
|-------- **dataAccs** (2.4) - Tiedon saatavuuden ja käytön kuvailu.
| |
| |-------- **setAvail** (2.4.1) - Tutkimustiedon saatavuuden kuvailu. Attribuutit: media (tallennusmedia), callno (arkistointinumero), label (nimike) ja type (saatavuustyyppi).
| | |-------- **accsPlac** (2.4.1.1) - Tiedon tallennussijainti. Attribuutit: URI.
| | |-------- **origArch** (2.4.1.2) - Arkisto, johon tutkimustieto on alun perin tallennettu.
| | |-------- **avlStatus** (2.4.1.3) - Saatavuusstatus.
| | |-------- **collSize** (2.4.1.4) - Tutkimusdataa sisältävien tiedostojen lukumäärä sekä mainintа muista, lisätietoa sisältävistä tiedostoista.
| | |-------- **complete** (2.4.1.5) - Mainінta tiedosta, joka on kerätty, mutta jota ei syystä tai toisesta ole tallennettu tutkimukseen liittyviin tiedostoihin.
| | |-------- **fileQnty** (2.4.1.6) - Kaikkien tutkimukseen liittyvien tiedostojen lukumäärä.
| | +-------- **notes** (2.4.1.7) - Muita tutkimuksen saattavuutta koskevia huomioita. Attribuutit: type, subject, level ja resp.
| |
| |-------- **useStmt** (2.4.2) - Tutkimustiedon käytön kuvailu.
| | |-------- **confDec** (2.4.2.1) - Vaatiiko tiedon käyttö luottamuksellisuussopimuksen tekemistä. Attribuutit: required (koneellista lukua varten, yes/no), formNo (täytettävän lomakkeen numero) ja URI.
| | |-------- **specPerm** (2.4.2.2) - Vaatiiko tiedon käyttö erityisluvan saamista. Attribuutit: required, formNo ja URI.
| | |-------- **restrctn** (2.4.2.3) - Käyttöön liittyvät mahdolliset rajoitukset.
| | |-------- **contact** (2.4.2.4) - Kontaktihenkilö käyttöön liittyviä kysymyksiä varten. Attribuutit: affiliation, URI ja email.
| | |-------- **citReq** (2.4.2.5) - Tutkimustietoa koskevat viittaussäännöt.
| | |-------- **deposReq** (2.4.2.6) - Tiedon arkistointiin liittyvät säännöt.
| | |-------- **conditions** (2.4.2.7) - Muita tutkimustiedon käyttöön liittyviä ehtoja.
| | +-------- **disclaimer** (2.4.2.8) - Vastuuvapauslauseke.
| |
| +-------- **notes** (2.4.3) - Muita tutkimuksen käyttöä koskevia huomioita. Attribuutit: type, subject, level ja resp.
|
|-------- **othrStdyMat** (2.5) - Muu tutkimuksen kuvailuun liittyvä materiaali.
| |
| |-------- **relMat** (2.5.1) - Tutkimuksen kuvailuun liittyvä materiaali. Attribuutit: callNo, label, media ja type
| |-------- **relStdy** (2.5.2) - Tutkimuksen kuvailuun liittyvät tutkimukset.

| |-------- **relPubl** (2.5.3) - Tutkimuksen kuvailuun liittyvät julkaisut.
| |-------- **othRefs** (2.5.4) - Tutkimuksen kuvailuun liittyvät muut viittaukset.
|
+-------- **notes** (2.6) - Muita tutkimuksen kuvailua koskevia huomioita. Attribuutit: type, subject, level ja resp.

.
.
.

fileDscr (3.0) - Tiedostojen kuvailu. Toistetaan kunkin tiedoston kohdalla. Attribuutit: URI, sdatrefs (viittaus sumDscr (2.2.3) -kohdan sisältämien lapsielementtien ID-numeroihin), methrefs (viittaus method (2.3) -kohdan sisältämien lapsielementtien ID-numeroihin) , access (viittaus dataAccs (2.4) -kohdan sisältämien lapsielementtien ID-numeroihin) ja pubrefs (viittaus othrStdyMat (2.5) ja otherMat (5.0) -kohtien sisältämien lapsielementtien ID-numeroihin).

|
|-------- **fileTxt** (3.1) - Tiedostoihin liittyvien tietojen kuvailu.
| |
| |-------- **fileName** (3.1.1) - Tiedoston nimi.
| |-------- **fileCont** (3.1.2) - Lyhyt kuvat tiedoston sisällöstä. Kuvataan mm. tiedoston merkitys,
| | luonne ja alueet, jotka tiedoston sisältö kattaa.
| |
| |-------- **fileStrc** (3.1.3) - Tiedoston rakenne. Vaihtoehdot: hierarkkinen (hierarchical), suorakulmainen (rectangular), relaationalinen (relational) ja sisäkkäinen (nested).
| | |-------- **recGrp** (3.1.3.1) - Hierarkkisten, relaationalisten ja sisäkkäisten tiedostoluetteloiden kuvailu. Sisältää myös attribuutteja ja lapsielementtejä tarkempaa kuvailua varten.
| | +-------- **notes** (3.1.3.2) - Muita tiedoston rakennetta koskevia huomioita. Attribuutit: type, subject, level ja resp.
| |
| |-------- **dimensns** (3.1.4) - Tiedoston ulottuvuuksien kuvailu.
| | |-------- **caseQnty** (3.1.4.1) - Tutkittavien tapausten tai havaintojen lukumäärä.
| | |-------- **varQnty** (3.1.4.2) - Muuttujien lukumäärä.
| | |-------- **logRecL** (3.1.4.3) - Tiedoston looginen pituus (merkkien lukumäärä).
| | |-------- **recPrCas** (3.1.4.4) - Yhtä tapausta koskevien tietueiden lukumäärä.
| | +-------- **recNumTot** (3.1.4.5) - Tietueiden kokonaislukumäärä.
| |
| |-------- **fileType** (3.1.5) - Tiedoston tyyppi. Voi sisältää raakadataa (esim. ASCII) tai ohjelmistoriippuvaista dataa (esim. SPSS-tiedostot). Attribuutit: charset (sanaston määrittely, esim. UTF-8).
| |-------- **format** (3.1.6) - Tiedostossa olevan datan fyysinen formaatti.
| |-------- **filePlac** (3.1.7) - Tiedoston tuotantopaikka.
| |-------- **dataChck** (3.1.8) - Tiedostotasoisten tarkastusten ja operaatioiden kuvailu.
| |-------- **ProcStat** (3.1.9) - Prosessointistatus.
| |-------- **dataMsng** (3.1.10) - Tiedostosta puuttuva data.
| |-------- **software** (3.1.11) - Tiedoston tuottamiseen käytetyt ohjelmisto.
| | Attribuutit: date ja version
| +-------- **verStmt** (3.1.12) - Tiedoston versiotiedot. Sisältää samat lapsielementit kuin kohta 1.1.6.
|
|-------- **locMap** (3.2) - Tiedostojen fyysisen tallennussijainnin kuvailu.
| |
| +-------- **dataItem** (3.2.1) - Yksittäisen tietosisällön fyysinen tallennussijainti. Attribuutit: varRef (viittaus kyseistä tietoa keräävän muuttujan ID-numeroon), nCubeRef (viittaus kyseistä tietoa keräävän nCube -matriisin ID-numeroon).
| kuvaukseen).
| |-------- **cubeCoord** (3.2.1.1) - Mikäli tiedon tallennukseen käytetään nCube -matriiseja,
| | voidaan tämän elementin attribuuttien avulla ilmoittaa sijainti matriisissa.

| | Attribuutit: coordNo (koordinaation numero), coordVal (koordinaation arvo) ja
| | coordValRef (viittaus muuttujaan johon koordinaatin arvo on tallennettu).
+-------- **physLoc** (3.2.1.2) - Tietosisällön fyysinen sijainti tiedostossa. Kuvailu attribuuteilla: type, recRef (viittaus kohdan 3.1.3.1 ID-numeroihin), startPos (muuttujan tai tietueen aloitussijainti), endPos (lopetussijainti) ja width (koko).
|
+-------- **notes** (3.3) - Muita tiedoston kuvailua koskevia huomioita. Attribuutit: type, subject, level ja resp.

.
.
.

dataDscr (4.0) - Muuttujatietojen kuvailu.
|
|-------- **varGrp** (4.1) - Muuttujaryhmien kuvailu. Attribuutit: type, var (vittaus ryhmän sisältämien muuttujien ID-numeroihin, kohta 4.3), varGrp (viittaus toissijaisten muuttujaryhmien ID-numeroihin), sdatrefs, methrefs, access ja pubrefs.
| |
| |-------- **labl** (4.1.1) - Muuttujaryhmän lyhyt kuvaus. Attribuutit: level (taso), vendor (otsikot eri ohjelmistoille), country (otsikot eri maille), sdatrefs.
| |-------- **txt** (4.1.2) - Pidempi kuvaus muuttujaryhmälle. Attribuutit: level ja sdatref.
| |-------- **concept** (4.1.3) - Aihepiiri, jota muuttujaryhmä käsittelee.
| | Attribuutit: vocab ja vocabURI.
| |-------- **defntn** (4.1.4) - Perustelut muuttujien ryhmittelylle.
| |-------- **universe** (4.1.5) - Tutkimusryhmä tai sen osa, jota muuttujaryhmä avulla kuvaillaan.
| | Attribuutit: level ja clusion.
| +-------- **notes** (4.1.6) - Muita muuttujaryhmää koskevia huomioita. Attribuutit: type, subject, level ja resp.
|
|
|-------- **nCubeGrp** (4.2) - nCube -matriisiryhmien kuvailu. Attribuutit: type (ryhmittelyn tyyppi), nCube (viittaus ryhmän sisältämiin matriisien ID-numeroihin, kohta 4.4), nCubeGrp (viittaus toissijaisiin nCube -ryhmien ID-numeroihin), name (ryhmän nimi) sekä sdatrefs, methrefs, access ja pubrefs. Kohta sisältää myös samat lapsielementit kuin kohta 4.1.
|
|
|-------- **var** (4.3) - Muuttujien kuvailu. Toistetaan kunkin muuttujan kohdalla.
| Attribuutit: name (nimi), wgt (onko painomuut.), wgt-var (viittaus painomuut.), weight (viittaus kohtaan 2.3.1.12), qstn (viittaus kohtaan 4.3.8), files (viittaus kohtaan 3.0), vendor (muuttujan ohjelmistoformaatti), dcml (desimaalien lkm.), intrvl (jaksotustyyppi, erillinen/jatkuva), rectype (viitaus kohtaan 3.1.3.1), sdatrefs, methrefs, pubrefs, access, aggrMeth (yhteen kokoamisen metodi), measUnit (mittausyksikkö), scale (mittasuhde), origin (mitta-asteikon alkupiste), nature (muuttujan luonne), additivity (lisättävyyden tyyppi), temporal (sisältääkö aikariippuvaista tietoa), geog (sisältääkö maantieteellistä tietoa), geoVocab (tiedon tallennukseen käytetty malli), catQnty (kategorioiden lkm.).
| |
| |-------- **location** (4.3.1) - Sijainti tiedostossa. Kuvailu attribuuteilla: startPos, endPos, width, RecSegNo (Luettelosegmentin numero), fileid (viittaus kohtaan 3.0) ja locMap (viittaus kohtaan 3.2).
| |-------- **labl** (4.3.2) - Muuttujan lyhyt kuvaus. Attribuutit: kts. kohta 4.1.1.
| |-------- **imputation** (4.3.3) - Arvio syistä tiedon puuttumiselle.
| |-------- **security** (4.3.4) - Muuttujan luottamuksellisuus/salaisuus. Attribuutit: date.
| |-------- **embargo** (4.3.5) - Muuttujan tietojen sulkeminen ulkopuolisista.
| | Attribuutit: event (sulku alkaa/päättyy), date ja format (sulun muoto).
| |-------- **respUnit** (4.3.6) - Taho, jolta muuttujan tiedot saatiin.
| |-------- **anlysUnit** (4.3.7) - Taho, jota muuttujan tiedot kuvaavat.

| |
| |-------- **qsnt** (4.3.8) - Muuttujaan liittyvä kysymys. Attribuutit: qstn (ID-attribuutti, jonka avulla
| | sama kysymys voidaan liittää useampaan muuttujaan), var (viittaus kohtaan 4.3), seqNo
| | (kysymyksen seqvenssinumero) ja sdatrefs.
| | |-------- **preQtxt** (4.3.8.1) - Esikysymys.
| | |-------- **qstnLit** (4.3.8.2) - Varsinainen kysymys. Attribuutit: sdatrefs.
| | |-------- **postQTxt** (4.3.8.3) - Jälkikysymys.
| | |-------- **forward** (4.3.8.4) - Etenemisohjeet kysymyksen jälkeen. Attribuutit: qstn (viittaus
| | | kysymykseen, johon tulisi siirtyä).
| | |-------- **backward** (4.3.8.5) - Ohjeet taaksepäin siirtymiselle. Attribuutit: qstn.
| | +-------- **ivuInstr** (4.3.8.6) - Ohjeet kysymyksen esittäjälle.
| |
| |-------- **valrng** (4.3.9) - Hyväksyttävien arvojen vaihteluväli.
| | |-------- **range** (4.3.9.1) - Vaihteluväli. Attribuutit: UNITS (yksikkö), min (minimi, >=),
| | | max (maksimi, <=), minExclusive (minimi, >) ja maxExclusive (maksimi, <).
| | |-------- **item** (4.3.9.2) - Yksittäisten arvojen ilmoittaminen. Attribuutit: UNITS (yksikkö)
| | | ja VALUE (arvo).
| | |-------- **key** (4.3.9.3) - Arvoja vastaava sisältö. Voidaan ilmoittaa myös kategorioiden
| | | kuvaamisen yhteydessä.
| | +-------- **notes** (4.3.9.4) - Muita hyväksyttäviä arvoja koskevia huomioita.
| | Attribuutit: type, subject, level ja resp.
| |
| |-------- **invalrng** (4.3.10) - Ei-hyväksyttävien arvojen vaihteluväli. Samat lapsielementit ja attri-
| | buutit kuin kohdassa 4.3.9.
| |
| |-------- **undocCod** (4.3.11) - Tuntemattomien arvojen kuvailu.
| |-------- **universe** (4.3.12) - Tutkimusryhmä tai sen osa jota muuttuja kuvaa. Attribuutit:
| | kts. kohta 2.2.3.9.
| |-------- **TotlResp** (4.31.3) - Vastauksien määrä muuttujassa.
| |-------- **sumStat** (4.3.14) - Yhteenvetotilasto muuttujasta. Attribuutit: wgtd (onko tilasto paino-
| | tettu, wgt-var (viitaus painomuuttujaan), weight (viittaus kohtaan 2.3.1.12), type.
| |-------- **txt** (4.3.15) - Pidempi kuvaus muuttujalle. Attribuutit: level ja sdatref.
| |-------- **stdCatqry** (4.3.16) - Käytettävät standardikategoriat. Attribuutit: date ja URI.
| |
| |-------- **catgryGrp** (4.3.17) - Muuttujakategoriaryhmien kuvailu. Attribuutit: missing (sisältääkö
| | puuttuvaa dataa), missType (puuttuvan datan tyyppi), catgry (viitaus ryhmän sisältämi-
| | en kategorioiden ID-numeroihin, kohta 4.3.18), catGrp (viittaus toissijaisten katego-
| | riaryhmien ID-numeroihin), levelno (ryhmän tasonumero), levelnm (ryhmän tasonimi),
| | compl (onko ryhmä lopullinen vai ei) ja excls (onko ryhmä ainutkertainen vai ei).
| | |-------- **labl** (4.3.17.1) - Kategoriaryhmän lyhyt kuvaus. Attribuutit: kts. kohta
| | | 4.1.1.
| | |-------- **catStat** (4.3.17.2) - Kategoriaryhmän tilasto. Attribuutit: type, wgtd, wgt-var,
| | | weight, sdatrefs, methrefs ja URI.
| | +-------- **txt** (4.3.17.3) - Pidempi kuvaus kategoriryhmälle. Attribuutit: level ja sdatrefs.
| |
| |-------- **catgry** (4.3.18) - Muuttujakategorioiden kuvailu. Attribuutit: missing, missType,
| | | country, excls ja sdatrefs.
| | |-------- **catValu** (4.3.18.1) - Kategorian arvo.
| | |-------- **labl** (4.3.18.2) - Kategorian lyhyt kuvaus. Attribuutit: kts. kohta 4.1.1.
| | |-------- **txt** (4.3.18.3) - Pidempi kuvaus kategorialle. Attribuutit: level ja sdatrefs.
| | |-------- **catStat** (4.3.18.4) - Kategorian tilasto. Attribuutit: kts. kohta 4.3.17.2.
| | +-------- **mrow** (4.3.18.5) - Kategorian ominaisuuksien kuvaaminen yhden yhtenäisen
| | rivin avulla. Sisältää lapsielementin identifiointia varten.
| |
| |-------- **codInstr** (4.3.19) - Tietoa, jota tarvitaan jos muuttujan sisältämä tieto muutetaan toiseen
| | muotoon.

| |-------- **verStm** (4.3.20) - Muuttujan versiotiedot. Sisältää samat lapsielementit kuin kohta 1.1.6.
| |-------- **concept** (4.3.21) - Aihepiiri, jota muuttuja käsittelee. Attribuutit: vocab ja vocabURI.
| |
| |-------- **derivation** (4.3.22) - Muista muuttujista johdettujen muuttujien kuvailu.
| | Attribuutit: var (viittaus niiden muuttujien ID-numeroihin, joista kyseisen muuttuja
| | tiedot johdetaan).
| | |-------- **drvdesc** (4.3.22.1) - Kuvaus sille kuinka tiedot on johdettu.
| | +-------- **drvcmd** (4.3.22.2) - Komento, jolla tietojen johtaminen tehdään. Attribuutit:
| | syntax (komennon kieli).
| |
| |-------- **varFormat** (4.3.23) - Muuttujan tekninen formaatti. Attribuutit: type (formaatin tyyppi,
| | merkki/numero), formatname (formaatin nimi), schema (käytetty ohjelmisto), category
| | (minkä tyyppistä dataa formaatti sisältää) ja URI.
| |-------- **geoMap** (4.3.24) - Linkki karttaan, josta nähdään muuttujan käsittelemä maantiet. alue.
| | Attribuutit: URI, mapformat (kartan formaatti) ja levelno (kartan taso).
| +-------- **notes** (4.3.25) - Muita muuttujaa koskevia huomioita. Attribuutit: type, subject, level ja
| resp.
|
|-------- **nCube** (4.4) - nCube -matriisien kuvailu. Kuvailu sisältää lapsielementit location, labl, txt,
| universe, imputation, security, embargo, respUnit, verStmt, purpose, dmns, measure ja notes.
|
+-------- **notes** (4.5) - Muita muuttujatietojenkuvailua koskevia huomioita. Attribuutit: type, subject, level
ja resp.

.
.
.

otherMat (5.0) - Muu tutkimukseen liittyvä dokumentaatio. Toistetaan eri materiaalien kohdalla erikseen. Attribuutit: type (materiaalin tyyppi), level (materiaalin taso) ja URI (materiaalin osoite).

|
|-------- **labl** (5.1) - Materiaalin lyhyt kuvaus. Attribuutit: kts. kohta 4.1.1.
|
|-------- **txt** (5.2) - Pidempi kuvaus materiaalille. Attribuutit: level (taso) ja sdatrefs (viittaus
| kohtaan 2.2.3).
|
|-------- **notes** (5.3) - Muita materiaalia koskevia huomioita. Attribuutit: type, subject, level ja resp.
|
|-------- **table** (5.4) - Taulukkomuotoisen lisämateriaalin kuvailu. Sisältää lapsielementtejä sekä attri-
| buutteja tarkemman kuvailun suorittamiseksi.
|
+-------- **citation** (5.5) - Tutkimukseen liittyvän lisämateriaalin bibliografinen kuvailu Elementit titlStmt, rspStmt, prodStmt, distStmt, serStmt, verStmt, biblCit, holdings ja notes sekä niiden lapsielementit ja attribuutit. (Tarkempi rakenne kts. kohta 1.1, citation). Attribuutit: MARCURI.

LIITE 2

LKK-tutkimuksen dokumentointitarpeiden pohjalta kehitelty DDI-pohjainen metadata-taskeema.

codeBook.
.
.
docDscr - Metadatadokumentin kuvailu.
|
+-------- **citation** - Bibliografisen tiedon kuvaaminen.
|
|-------- **titlStmt** - Otsikkotietojen kuvailu.
| |-------- **titl** - Otsikko.
| +-------- **subTitl** - Lisäotsikko.
|
|-------- **rspStmt** - Vastuutietojen kuvailu.
| +-------- **authEnty** - DDI-dokumentaation sisällön syöttämisestä vastuussa oleva
| henkilö, taho tai yksikkö. Attribuutit: affiliation.
|
|-------- **prodStmt** - Tuotantotietojen kuvailu.
| |-------- **producer –** Tuottaja (henkilö tai organisaatio). Attribuutit: affiliation ja abbr.
| |-------- **copyright** - Copyright.
| +-------- **prodDate** - Tuotantopäivämäärät. Attribuutit: date.
|
+-------- **distStmt** - Jakelutietojen kuvailu.
+-------- **contact** - Yhteyshenkilö. Attribuutit: affiliation ja email.
.
.
stdyDscr - Tutkimuksen kuvailu.
|
|-------- **citation** - Bibliografisen tiedon kuvaaminen.
| |
| |-------- **titlStmt** - Otsikkotietojen kuvailu.
| | |-------- **titl** - Otsikko.
| | +-------- **subTitl** - Lisäotsikko.
| |
| |-------- **rspStmt** - Vastuutietojen kuvailu.
| | +-------- **authEnty** - Aineiston tuottamisesta vastuussa oleva taho. Attribuutit: affillation.
| |
| |-------- **prodStmt** - Tuotantotietojen kuvailu.
| | |-------- **producer –** Tuottaja (henkilö tai organisaatio). Attribuutit: affiliation ja abbr.
| | +-------- **copyright** - Copyright.
| |
| |-------- **distStmt** - Jakelutietojen kuvailu.
| | +-------- **contact** - Yhteyshenkilö. Attribuutit: affiliation ja email.
| |
| +-------- **serStmt -** Sarjatietojen kuvailu.
| +-------- serName -Sarjan nimi.
|
|-------- **stdyInfo** - Tutkimuksen ulottuvuuksien kuvailu.
| |

| |-------- **subject** - Tutkimuksen aihepiiriin liittyvien tietojen kuvailu.
| | |
| | +-------- **topcClas** - Aihepiirin luokittelu.
| |
| +-------- **sumDscr** - Tutkimuksen ajallisen ja maantieteellisen kattavuuden sekä analysointiyksikön kuvailu.
| |-------- **collDate** - Tutkimusaineiston keräyspäivämäärät. Attribuutit: event ja date.
| |-------- **nation** - Keräysmaa. Attribuutit: abbr.
| |-------- **geogCover** - Maantieteellinen kattavuus (esim. Keski-Suomi).
| |-------- **geogUnit** - Maantieteellinen yksikkö (esim. Kunta).
| |-------- **anlyUnit** - Analysointiyksikkö. Attribuutit: unit.
| |-------- **universe** - Tutkimusryhmä.
| +-------- **dataKind** - Tutkimusdatan tyyppi (esim. Kyselytulos).
|
|-------- **method** - Tutkimusmetodien kuvailu.
| |
| |-------- **dataColl** - Tiedon keräyksessä käytetyn metodin kuvailu.
| | +-------- **collMode** - Tiedon keräyksessä käytetty metodi.
| |
| +-------- **notes** - Muita tutkimusmetodia koskevia huomioita
|
+-------- **dataAccs** - Tiedon saatavuuden ja käytön kuvailu.
|
+-------- **setAvail** - Tutkimustiedon saatavuuden kuvailu.
+-------- **accsPlac** - Tiedon tallennussijainti.

.
.

fileDscr - Tiedostojen kuvailu. Toistetaan kunkin tiedoston kohdalla. Attribuutit: URI.
|
|-------- **fileTxt** - Tiedostoihin liittyvien tietojen kuvailu.
| |
| |-------- **fileName** - Tiedoston nimi. Attribuutit: ID.
| |-------- **fileCont** - Lyhyt kuvat tiedoston sisällöstä.
| |-------- **fileType** - Tiedoston tyyppi.
| +-------- **filePlac** - Tiedoston tuotantopaikka.
|
+-------- **notes** - Muita tiedoston kuvailua koskevia huomioita.

.
.

dataDscr - Tärkeimpien muuttujien kuvailu.
|
+-------- **var** - Muuttujien kuvailu. Toistetaan kunkin muuttujan kohdalla. Attribuutit: name.
|
|-------- **labl** - Muuttujan lyhyt kuvaus.
|-------- **universe** - Tutkimusryhmä tai sen osa jota muuttuja kuvaa.
+-------- **notes** - Muita muuttujaa koskevia huomioita. Attribuutit: type.

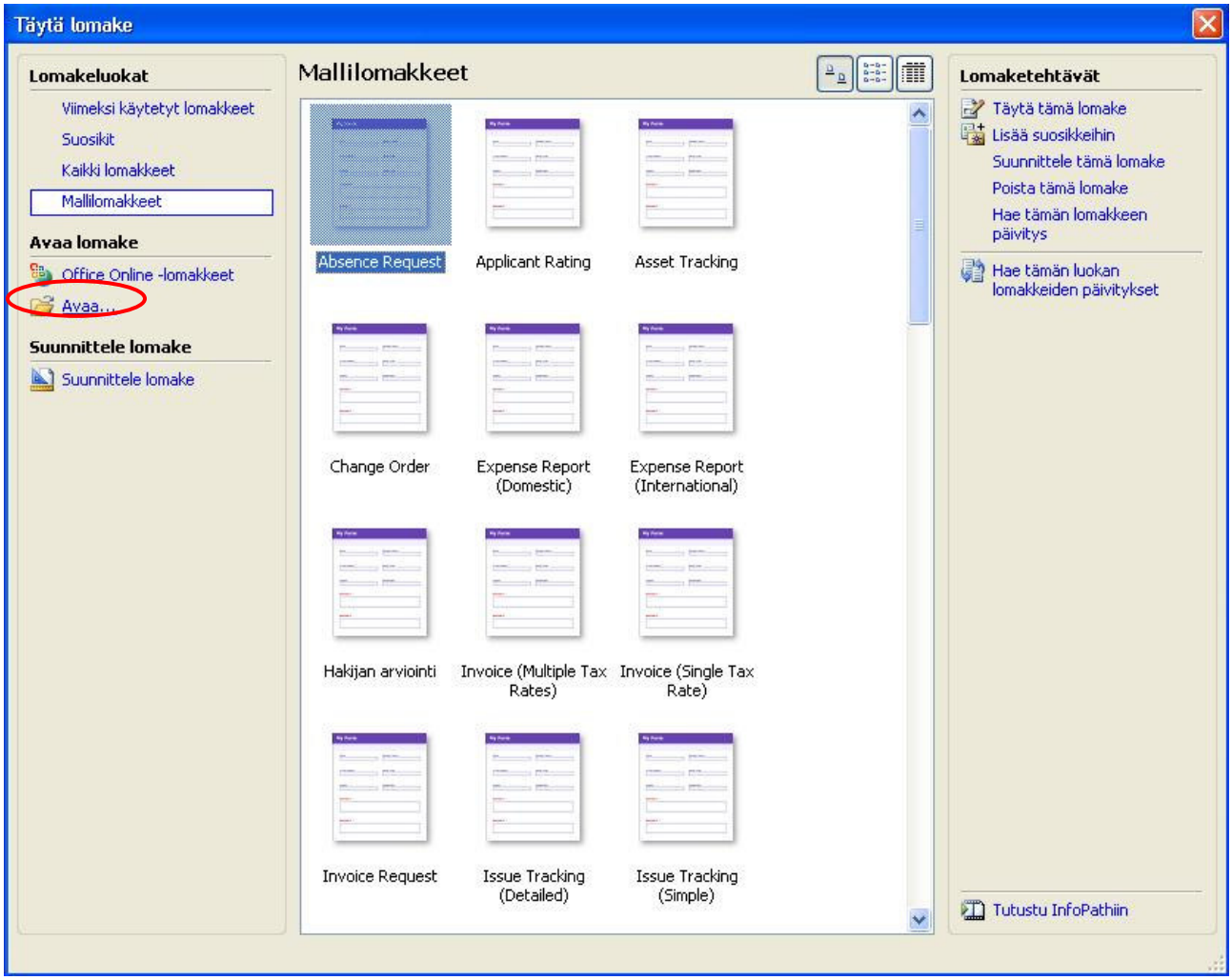
LIITE 3

Käyttöohjeet DDI-dokumentaation syöttämiselle ja selailulle
Juha Sinkkonen
27.6.2005

Uuden DDI-dokumentin luominen:

1) Avaa Mircosoft Office InfoPath 2003 -ohjelma.
Ohjelman saat avattua esimerkiksi Käynnistä -palkissa olevan pikakuvakkeen avulla (*"Käynnistä" → "Ohjelmat" → "Microsoft Office" tai "Start" → "Programs" → "Microsoft Office"*). Mikäli InfoPath -ohjelmaa ei ole asennettu koneellesi, ota yhteyttä mikrotukihenkilöön.

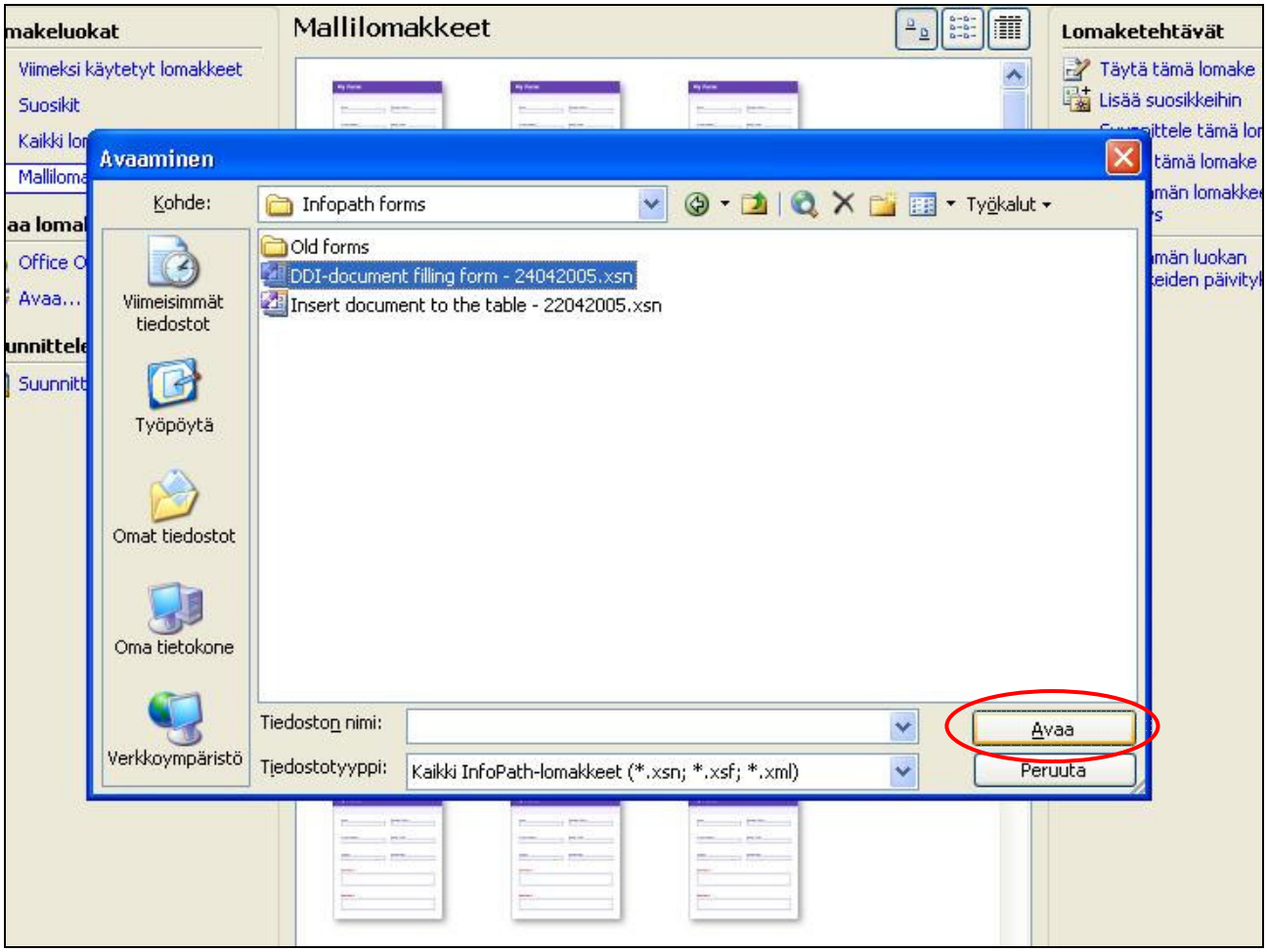
2) Avaa lomakepohja valitsemalla päällimmäiseksi avautuvan ikkunan vasemmasta reunasta kohta *"Avaa..."*



Vaihtoehtoisesti voit avata lomakepohjan myös painamalla InfoPathin pääikkunassa ”Tiedosto” → ”Avaa” (”File” → ”Open”) tai painamalla -nappia. Myös tuplaklikkaaminen resurssien hallinnassa avaa lomakepohjan.

Avattava lomakepohja sijaitsee verkkolevyllä osoitteessa: **…\DDI-documentation\ Infopath forms**.

Valitse uusin versio ”DDI-document filling form” -nimisestä dokumentista. Uusimman version tunnistat tiedostonimessä olevasta päivämäärästä.



3) Syötä tutkimuskertaa koskevat metatiedot avattuun lomakepohjaan.

Selitykset lomakkeessa oleville kentille:

- **Tiedot dokumentaatiosta**
  - **Dokumentaation otsikko.** Otsikko tutkimuskerran metatiedot sisältävälle DDI-dokumentaatiolle. Dokumentaation otsikoksi tulisi valita sellainen ot-

sikko, jonka avulla voidaan tunnistaa eri tutkimuskertoihin liittyvät dokumentaatiot toisistaan. Esimerkiksi tutkimusdatan sisältävä SPSS-tiedoston kahdeksan ensimmäistä merkkiä toimivat monessa tapauksessa hyvänä otsikkona dokumentaatiolle.

- **Dokumentaation syöttäjä.** Dokumentaation syöttäjäksi tulee merkitä sen henkilön nimi, joka on syöttänyt dokumentaation. Nimi tulee merkitä muodossa Sukunimi, Etunimi.
- **Syöttöpäivämäärä.** Syöttöpäivämääräksi tulee merkitä se päivämäärä, jolloin tutkimuskertaan liittyvä dokumentaatio on tallennettu tai kun sitä on muokattu viimeksi.
- **Kontaktihenkilö dokumentaation liittyville asioille.** Kontaktihenkilöksi tulee merkitä se henkilö, johon otetaan yhteyttä, mikäli dokumentaatioon liittyvissä asioissa ilmenee jotain kysyttävää. Oletusarvona kentässä on Kenneth Eklundin nimi ja sähköpostiosoite.

- **Tiedot tutkimuskerrasta**

  Yleiset tiedot. Tutkimuskerran tietojen syöttäminen aloitetaan antamalla yleisiä tietoja kyseisestä tutkimuskerrasta. Nämä tiedot syötetään vain kerran ja ne koskevat kaikki lomakkeen avulla kuvailtuja aihepiirejä.

  - **Tutkimuskerran otsikko.** Tutkimuskerran otsikoksi valitaan sellainen nimi, joka kuvaa tutkimuskertaa mahdollisimman hyvin. Esimerkiksi tutkimusdatan sisältävä SPSS-tiedoston nimi toimii monessa tapauksessa hyvänä tutkimuskerran otsikkona.
  - **Tutkittavan lapsen ikä vuosina.** Tähän kohtaan tulee merkitä joko **A)** tutkittavan lapsen ikä vuosina, kun kyseessä on ennen kouluikää tehtävä tutkimus tai **B)** testausajankohdan numerokoodi (esimerkiksi 8.5), kun kyseessä on kouluiässä tehtävä tutkimus. (Kuukausien muuttaminen desimaaleiksi: 1 kk = 0.1 v, 2 kk = 0.2 v, 3 kk = 0.25 v, 4 kk = 0.3 v, 5 kk = 0.4 v, 6 kk = 0.5 v, 7 kk = 0.6 v, 8 kk = 0.7 v, 9 kk = 0.75 v., 10 kk = 0.8 v, 11 kk = 0.9 v, 12 kk = 1.0 v.)
  - **Linkki tiedostoon, joka sisältää muut tutkimuskertaan liittyvät huomiot.** Tähän kohtaan voi lisätä tiedostopolun tiedostoon, jossa on annettu sellaisia lisätietoja, joita ei ole dokumentoitu lomakkeen muissa kohdissa.

- **Aihepiirit ja niihin liittyvä tutkimus.** Tutkimuskertaan liittyvien tietojen syöttäminen tapahtuu kuvailemalla siinä käytetyt aihepiirit ja niihin liittyvät tiedot. Tutkimuskerran metatiedot syötetään siten, että aluksi lomakkeeseen valitaan alasvetovalikosta tutkimuskertaan liittyvä aihepiiri ja tämän jälkeen tallennetaan aihepiiriin liittyvät tiedot. Tämä toistetaan jokaisen tutkimuskertaan liittyvän aihepiirin kohdalla.
  - **Aihepiirit (Content Areas).** Aihepiireillä tarkoitetaan niitä aihealueita, joita tutkimuskerrassa keskitytään tarkastelemaan.
  - **Tietotyypit (Data Types).** Valitse tietotyyppi alasvetovalikosta. Mikäli aihepiiriin sisältyy useampia tietotyyppejä, paina ”Lisää tietotyyppi” -painiketta. Voit myös poistaa tietotyyppejä valitsemalla alasvetovalikon vasemmalle puolelle ilmestyvää nuolta painamalla avautuvasta valikosta ”Poista data-Kind”.
  - **Mittarit (Content/Measure).** Kirjoita tähän kussakin aihepiirissä käytetyt tutkimusmittarit.

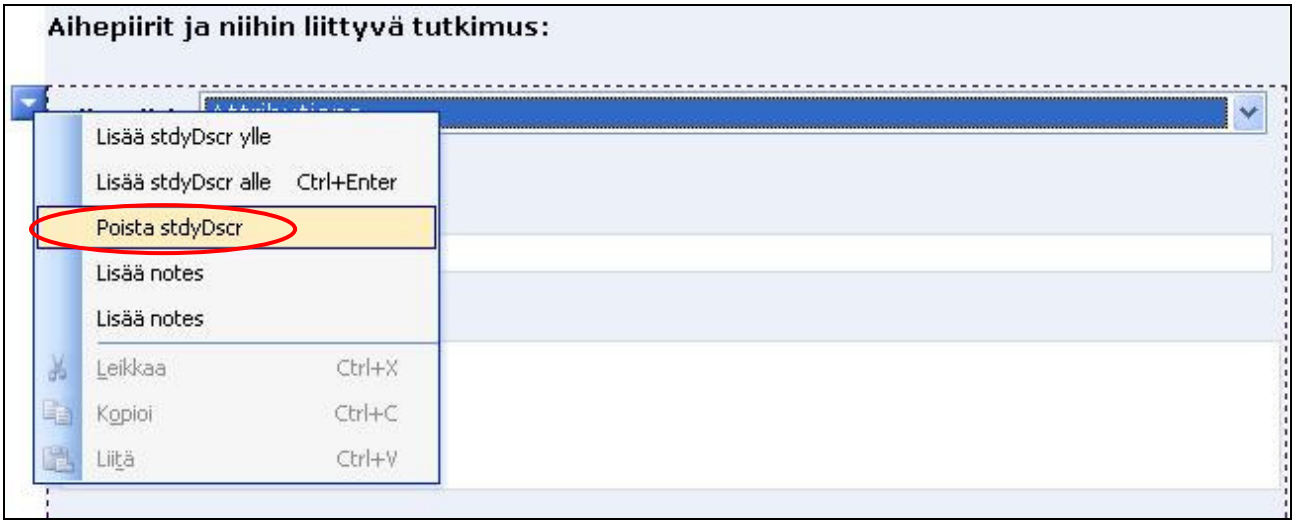
- **Linkki tiedostoon, joka sisältää edellä mainittuihin mittareihin liittyviä muita huomioita.** Tähän kohtaan voi lisätä yhden tai useamman tiedostopolun niihin tiedostoihin, joissa lisätietoja on annettu. Tarkemmat ohjeet tiedostopolkujen lisäämisestä on annettu hieman myöhemmin, tiedostojen kuvailun yhteydessä.

  Tutkimuskerrat saattavat myös sisältää sellaisia tutkimusvaiheita, jotka eivät sisälly mihinkään varsinaiseen tutkimusaihepiiriin. Näihin vaiheisiin liittyvät tiedot (esim. tietotyypit ja mittarit) tulisi kuitenkin kuvailla metatietolomakkeen avulla. Tällöin tulee käyttää aihepiirivalikossa olevia viivoja (-, --, ---, ----) kuvaamaan tyhjiä aihepiirejä. Viivoja on useampi kappale, jotta lomakkeen avulla voitaisiin kuvailla useita sellaisia vaiheita, jotka eivät liity mihinkään tutkimusaihepiiriin. Merkitse ensimmäisen tyhjän aihepiirin kohdalla ”-”, toisen ”--” jne.

  Yhden tutkimuskerran aikana saatetaan myös suorittaa useampia tutkimusvaiheita, jotka käsittelevät samaa tutkimusaihepiiriä. Tällöin tulee käyttää aihepiirivalikossa olevia järjestysnumeroita erottelemaan aihepiirejä toisistaan (esim. Reading ja Reading2).

  Voit aloittaa uuden aihepiirin, sekä siihen liittyvien tietojen kuvailun painamalla ”Tiedot tutkimuksesta” -kohdassa alimpana olevaan ”Lisää aihepiiri” -painiketta.

  Jos haluat poistaa tutkimuskertaan liittyvän aihepiirin ja siihen liittyvät tiedot, paina kyseisen aihepiirin kohdalla vasemmalle ilmestyvää nuolta ja valitse avautuvasta listasta ”*Poista stdyDscr*”.



- **Tiedostojen kuvailu**
  - **Valitse aihepiiri, johon tiedosto liittyy.** Tässä kohdassa olevan alasvetovalikon avulla voidaan kuvailtavat tiedostot liittää ”Tiedot tutkimuksesta” -kohdassa mainittuihin aihepiireihin seuraavanlaisesti: **A)** Liitä tiedosto yhteen aihepiiriin. Mikäli tiedostossa olevat tiedot koskevat esimerkiksi matemaattisia tehtäviä, voidaan se liittää aihepiiriin ”Mathematics”. **B)** Liitä tiedosto kaikkiin lomakkeessa mainittuihin aihepiireihin valitsemalla valikosta kohta ”Kaikki mainitut aihepiirit”. Näin voidaan toimia silloin kun tiedosto sisältää tietoja koko tutkimuskerrasta. **C)** Liitä tiedosto kahteen tai useampaan aihepiiriin. Näin tulee tehdä, jos tiedoston tiedot koskevat joitain mainittuja aihepiirejä (esim. kahta tai kolmea).

    Useamman aihepiirin voit syöttää painamalla valitsemalla alasvetovalikon vasemmalle puolelle ilmestyvää nuolta painamalla avautuvasta valikosta ”Lisää notes ylle/alle”, aihepiirien poistaminen onnistuu puolestaan valitsemalla ”Poista notes”.
  - **Tiedoston nimi.** Lisää tähän kohtaan Windowsin resurssienhallinnassa näkyvä tiedoston nimi. Tiedoston nimeen tulee sisällyttää myös tiedostopääte (esimerkiksi .sav tai .doc).
  - **Tiedoston lyhenne.** Tiedostolle tulee valita myös sopiva lyhenne. Tämä lyhenne näkyy metatietojen pohjalta muodostettavassa taulukossa ja se toimii linkkinä varsinaiseen tiedostoon.
  - **Tiedoston tyyppi.** Valitse tiedoston tallennustyyppi alasvetovalikosta.
    - Microsoft Word -tiedostot = **Word**
    - SPSS-tiedostot = **SPSS**
    - Metoditiedostot = **Method**
    - Julkaisutiedostot = **Publ**
    - SPSS-tiedostot sisältävä kansio = **Kansio-SPSS**
    - Word-tiedostot sisältävä kansio = **Kansio-Word**
  - **Kuvaus tiedoston sisällöstä.** Lyhyt kuvaus tiedoston sisällöstä. Esim. Kerätty tutkimusdata tai testauslomake.
  - **Tiedoston tallennussijainti.** Tiedoston tallennussijainti pitää syöttää tarkasti, koska sen avulla muodostetaan linkki tiedostoon. Tallennussijainnin saa selville esimerkiksi Windowsin resurssienhallinnasta. Seuraavalla sivulla olevasta kuvasta ympyröity polku toimii linkkinä siihen kansioon, joka sisältää kuvassa näkyvät tiedostot, Mikäli linkki halutaan tehdä varsinaiseen tiedostoon, tulee polkuun:

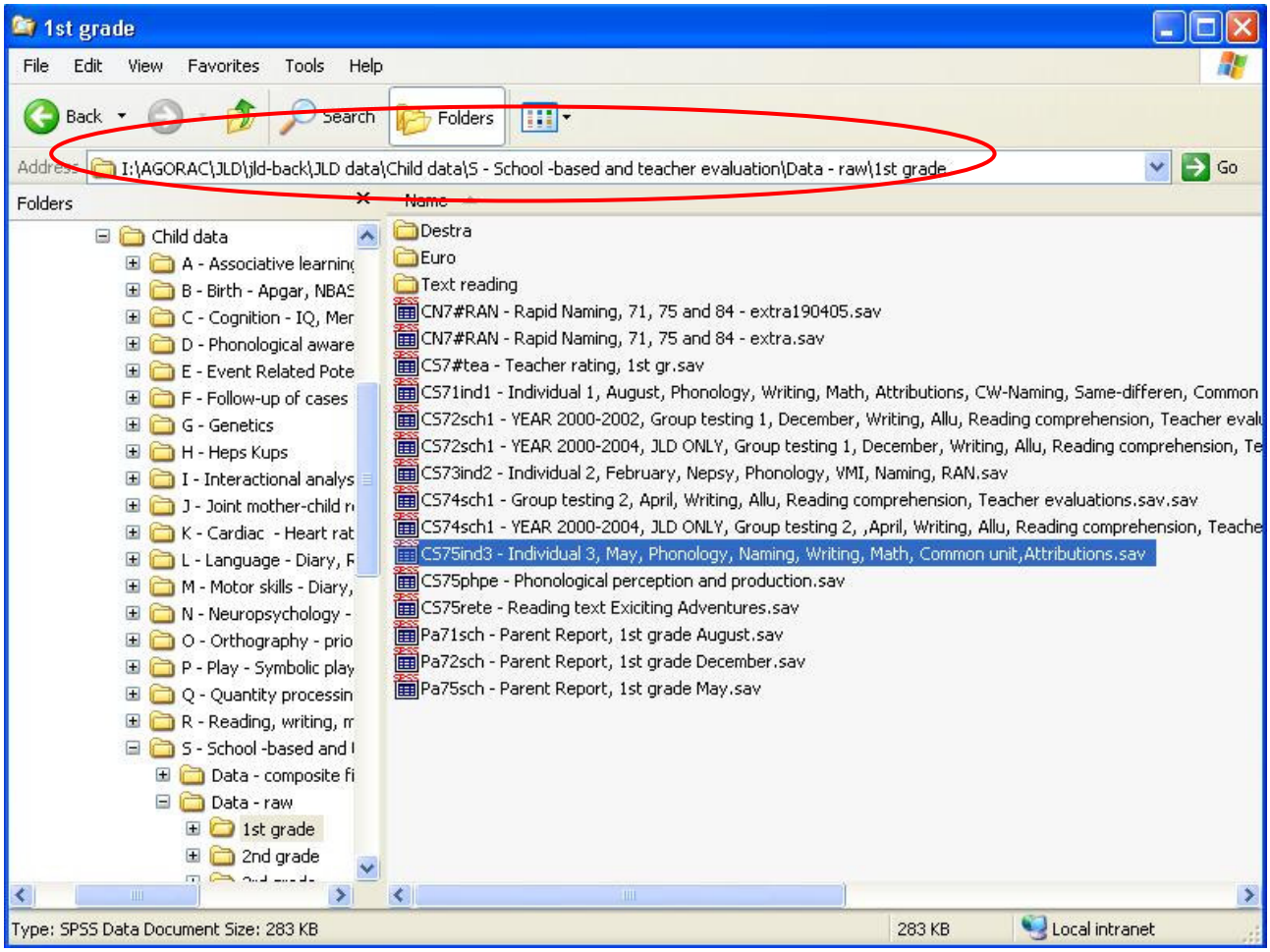
    ”***... \JLD data\Child data\S - School -based and teacher evaluation\Data - raw\1st grade***”

    lisätä vielä tiedoston nimi päätteineen, jolloin tallennussijainti on muotoa:

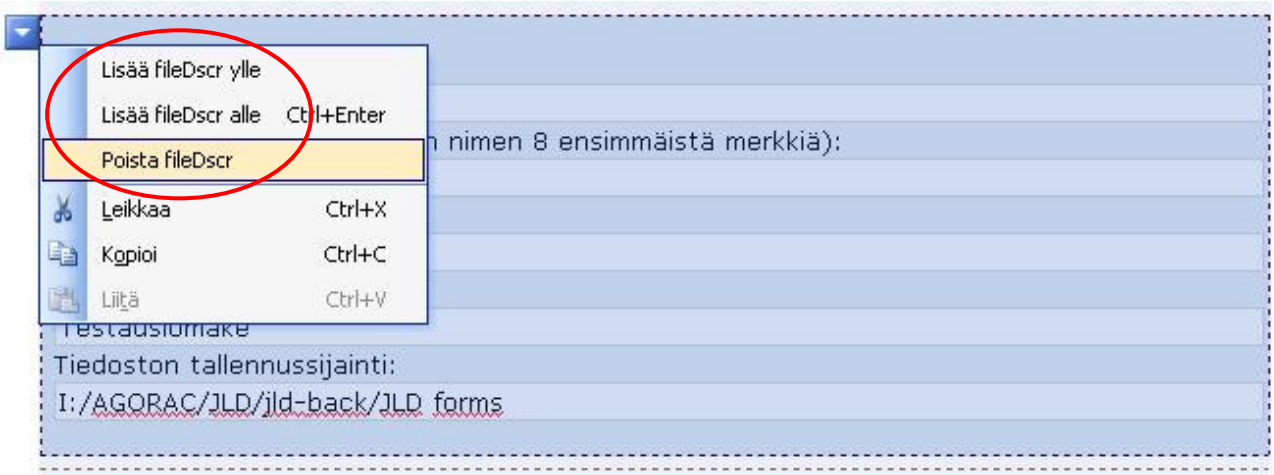
    ”***... \JLD data\Child data\S - School -based and teacher evaluation\Data - raw\1st grade\ CS75ind3 - Individual 3, May, Phonology, Naming, Writing, Math, Common unit,Attributions.sav***”.

Tiedostojen ja kansioiden tallennussijainneissa tulee resurssienhallinnasta poiketen käyttää kenoviivan (”\”) sijasta kauttaviivaa (”/”). Tällöin edelle mainitun tiedoston tallennussijainti merkittäisiin lomakkeeseen:

”.../JLD data/Child data/S - School -based and teacher evaluation/Data - raw/1st grade/ CS75ind3 - Individual 3, May, Phonology, Naming, Writing, Math, Common unit,Attributions.sav”



Tiedostojen kuvailu -kohdassa on oletusarvoisesti kuvattuna kaikkien erityyppisten tiedostojen tiedot (SPSS, Word, Metodi, Artikkeli, SPSS-kansio ja Word-kansio). Todennäköistä kuitenkin on, että tutkimuskerta ei sisällä kaikkia yllä mainituista tiedostotyypeistä tai että jotain tiedostotyyppiä esiintyy enemmän kuin yksi kappale. Tällöin voidaan painaa kunkin tiedostotyypin kohdalla vasemmalle ilmestyvää nuolta ja valita joko ”*Poista fileDscr*” tai ”*Lisää fileDscr ylle/alle*”.



- **Tärkeimpien muuttujien kuvailu**
  - **Aihepiiri, johon muuttajakuvaus liittyy.** Tässä kohdassa voidaan muuttujatiedot liitää lomakkeessa mainittuihin aihepiireihin. Muuttujien liittäminen tapahtuu samalla tavalla kuin tiedostojen liittäminen tiedostojen kuvailun yhteydessä.
  - **Muuttujan lyhenne.** SPSS-dokumenteissa käytettävä lyhenne muuttujalle. Lyhenne näkyy metatietojen pohjalta muodostettavassa taulukossa.
  - **Muuttujan kuvaus (SPSS).** SPSS-dokumenteissa käytettävä kuvaus. Kuvaus näkyy taulukossa kun hiiren osoitin laitetaan muuttujan lyhenteen päälle.

  Mikäli tutkimuskertaan ei liity sellaisia muuttujia, jotka halutaan mainita metatietojen yhteydessä, voidaan painaa muuttujakuvauksen vasemmalle puolella ilmestyvää nuolta ja valita *"Poista var"*. Samoin voidaan tehdä silloin kun lomakkeeseen on vahingossa syötetty sellaisten muuttujien tietoja, jota siinä ei tarvitsisi mainita. Mikäli muuttujia halutaan kuvata enemmän kuin yksi kappale voidaan myös tällöin painaa vasemmalle ilmestyvää nuolta ja valita *"Lisää var ylle/alle"*.

Muita huomioita lomakkeen täyttöön liittyen.

- **Merkit.** Lomakkeen kentät eivät saa sisältää seuraavia merkkejä: ' (heittomerkki), > (suurempi kuin merkki) ja < (pienempi kuin merkki). Mikäli haluat sisällyttää nämä merkin lomakkeen kenttiin voit tehdä se käyttämällä seuraavia vastineita:
  - ' → \'
  - < → <
  - > → >)

4) Tallenna lomakkeen tiedot

Kun tutkimuskertaan liittyvät metatiedot on syötetty, valitse *"Tiedosto"* → *"Tallenna"* (*"File"* → *"Save"*) tai paina  -nappia. Tiedostot tallennetaan muodossa XML-muodossa ja dokumenttien tiedostopäätteeksi tulee .xml.

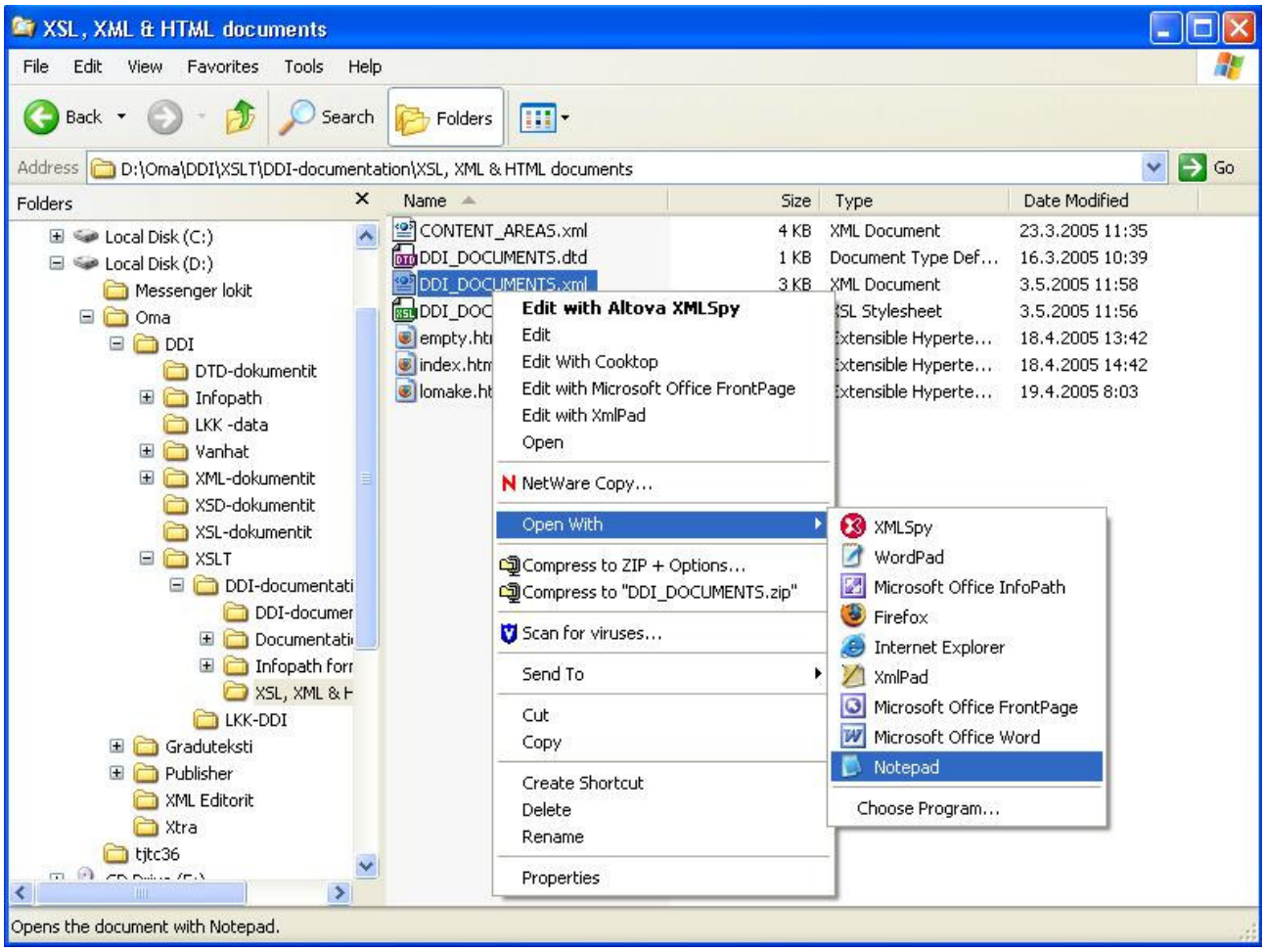
Lomakkeen tiedot tulee tallentaa sijaintiin ***I:\AGORAC\JLD\jld-back\JLD data information\DDI-documentation\DDI-documents***. Lomakkeen nimeksi tulee valita jokin yksilöllinen nimi, joka kuvaa mahdollisimman hyvin kuvattua tutkimuskertaa. Esimerkiksi tutkimusdatan sisältävän SPSS-tiedoston kahdeksan ensimmäistä merkkiä toimivat monessa tapauksessa hyvänä tiedostonimenä DDI-dokumentille. Tiedoston nimestä tulee selvitä minkä ikäistä lasta tutkimuksessa on tarkasteltu (esim. CS72sch1.xml).

Jos lomakkeeseen syötetään paljon tietoa tai jos kaikkea tietoa ei syötetä yhdellä kertaa, voit myös tallentaa lomakkeen tiedot vaikka kaikkia tietoja ei olisi vielä syötetty. Keskeneräisen lomakkeen voi myös avata myöhemmin jatkomuokkaamista varten. Olemassa olevan DDI-dokumentin muokkaamista käsitellään myöhemmin tässä ohjeessa.

Tiedostojen lisääminen DDI_DOCUMENTS.xml -dokumenttiin

Kun kaikki tarvittavat tiedot on syötetty lomakkeeseen ja lomakkeen tiedot on tallennettu voit halutessasi sulkea Microsoft InfoPath -ohjelman. Tämän jälkeen tallennettu tiedosto tulee lisätä DDI_DOCUMENT.xml tiedostossa olevaan listaan, jotta tallennetut tiedot näkyisivät metatietotaulukossa. Tiedoston lisääminen tapahtuu seuraavasti:

1) **Avaa DDI_DOCUMENT.xml -tiedosto Muistioon** (Notepad) painamalla tiedoston kohdalla hiiren oikeaa painiketta ja valitsemalla listasta *"Avaa sovelluksessa"* → *"Muistio"* (*"Open With"* → *"Notepad"*). (Katso kuva alla) Kyseinen tiedosto sijaitsee kansiossa ***I:\AGORAC\JLD\jld-back\JLD data information\DDI-documentation\XSL, XML & HTML documents***.

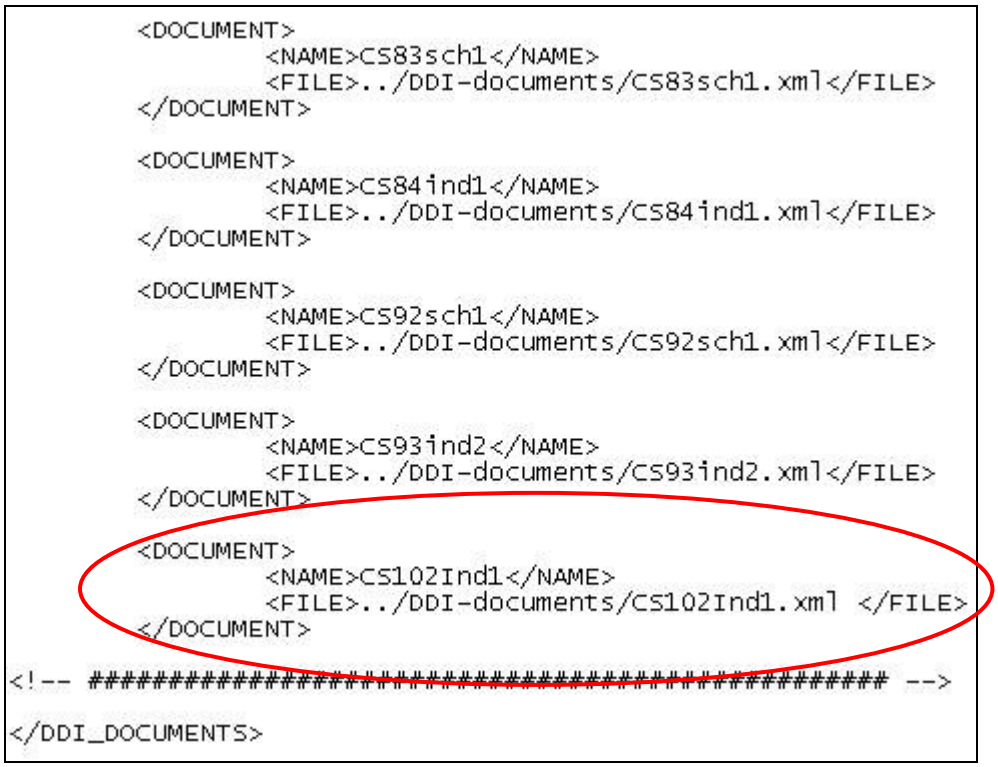


2) Lisää avattuun tiedostoon tutkimuskerran metatiedot sisältävän tiedoston nimi seuraavanlaisesti:

<DOCUMENT>
<NAME>[Tiedoston nimi]</NAME>
<FILE>../DDI-documents/[Tiedoston nimi].xml</FILE>
</DOCUMENT>

Esimerkiksi jos InfoPath -lomakkeeseen syötetyt tiedot tallennettiin tiedostoon nimeltä CS102Ind1.xml, tulee listaa lisätä seuraavanlainen merkintä.

<DOCUMENT>
<NAME>CS102Ind1</NAME>
<FILE>../DDI-documents/CS102Ind1.xml </FILE>
</DOCUMENT>



Tiedostot tulee syöttää listaan tutkittujen lasten syntymäaikajärjestyksestä. Syntymäajan voi päätellä listassa olevien tiedostojen nimissä olevista numeroista.

3) Kun tiedosto on lisätty listaan, valitse *"Tiedosto" → "Tallenna"* (*"File" → "Save"*).

Tämän jälkeen voit sulkea Muistion. Listaan lisätyn tiedoston tiedot tulisi nyt näkyä metatietotaulukossa.

Olemassa olevan DDI-dokumentin muokkaaminen

1) Avaa Microsoft Office InfoPath -ohjelma.

Tarkemmat ohjeet ohjelman avaamiseen löydät tämän ohjeen alusta.

2) Avaa DDI-dokumentti, jota haluat muokata

Kun olemassa olevaa dokumenttia halutaan muokata, avataan syöttölomakkeen sijasta se DDI-dokumentti, joka sisältää tietyn tutkimuskerran tiedot. Dokumentti avataan sijainnista ***... \DDI-documentation\DDI-documents***.

3) Tee dokumenttiin tarvittavat muutokset

Avattu dokumentti sisältää kaikki ne tiedot, jotka siihen on syötetty kyseisen dokumentin luomisen yhteydessä. Muokatessasi dokumenttia voit muuttaa tai poistaa olemassa olevia tietoja tai vaihtoehtoisesti voit myös lisätä uutta tietoa.

4) Tallenna dokumenttiin tehdyt muutokset

Kun muutokset on syötetty, valitse *"Tiedosto" → "Tallenna"* (*"File" → "Save"*) tai paina  -nappia. Tiedostoa ei tarvitse enää uudestaan lisätä DDI_DOCUMENT.xml -dokumenttiin, vaan tekemäsi muutokset näkyvät metatietotaulukossa kun se ladataan uudelleen.

Metatietotaulukon avaaminen

Kun haluat tarkastella InfoPath -lomakkeen avulla tallennettua metatietoa taulukossa toimi seuraavanlaisesti:

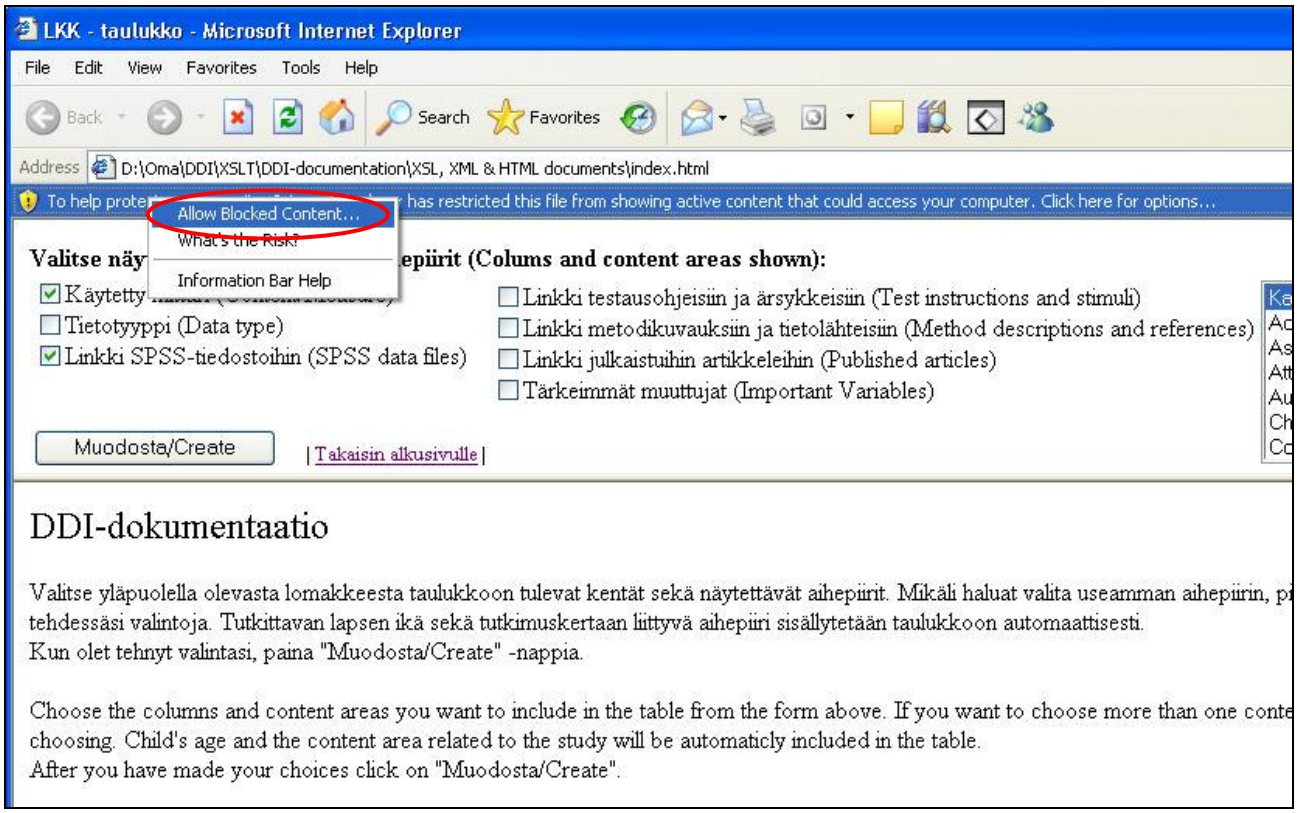
1) Avaa tiedosto index.html -tiedosto Internet Explorer -selaimessa

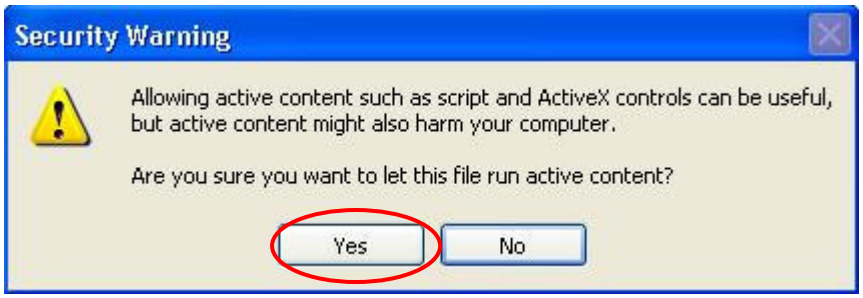
Mikäli Internet Explorer on asetettu tietokoneesi oletusselaimeksi, voit avata index.html -tiedoston tuplaklikkaamalla tiedostoa Windowsin resurssienhallinnassa. Internet Explorer on oletusselain mikäli tiedoston edessä on  -kuvake. Jos Internet Explorer ei ole oletusselain, paina index.html -tiedostoa hiirin oikealla napilla ja valitse avautuvasta valikosta *"Avaa sovelluksessa" → "Internet Explorer"* (*"Open With" → "Internet Explorer"*).

Tiedosto index.html sijaitsee kansiossa ***I:\AGORAC\JLD\jld-back\JLD data information\DDI-documentation* \XSL, XML & HTML documents\.**

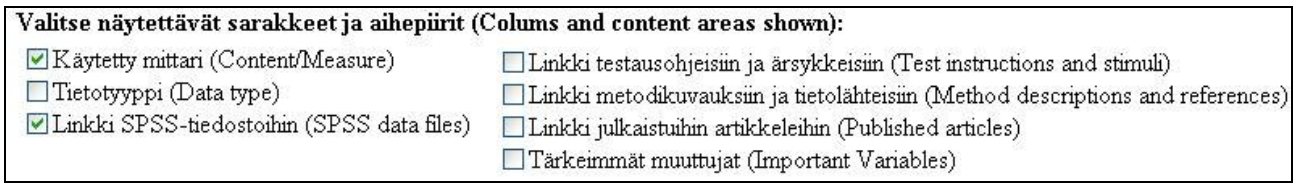
2) Salli JavaScript -osien toiminen sivuilla

Taulukko sisältää JavaScript -kielellä toteutettuja toimintoja, joista Internet Explorer varoittaa turvallisuusriskinä. Tällä sivulla olevat JavaScriptit eivät kuitenkaan sisällä turvallisuutta vaarantavaa toimintaa ja ne tuleekin sallia, jotta sivut toimisivat tarkoitetulla tavalla (katso kuvat alla). Mikäli turvallisuusvaroitusta ei tule, voit jatkaa metatietojen selailua normaalisti.





3) Rastita haluamasi kohdat ja/tai valitse luettelosta haluamasi aihepiirit



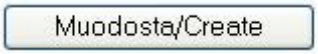
Kuvassa näkyvien rastien avulla voidaan valita, mitkä sarakkeet näkyvät muodostettavassa taulukossa. Oletusarvoisesti rasteista on valittuna ”Käytetty mittari” ja ”Linkki SPSS-tiedostoihin”. Valittavien rastien lisäksi taulukkoon lisätään automaattisesti tutkittavan lapsen ikä sekä tutkimuksen aihepiiri. Luettelo sarakkeista:

- **Age in Years** = Tutkittavan lapsen ikä vuosina (Näytetään aina)
- **Content Areas** = Tutkimuksen aihepiiri (Näytetään aina)
- **Content/Measure** = Käytetty mittari (Valittavissa, oletusarvoisesti valittuna)

- **Data Types** = Tutkimuksessa tallennetun tiedon tyypit (Valittavissa, oletusarvoisesti ei valittuna)
- **SPSS Data Files** = Linkit tutkimuksen SPSS-tiedostoihin (Valittavissa, oletusarvoisesti valittuna)
- **Test Instructions and Stimuli** = Linkki tutkimuksen testauslomakkeisiin (Valittavissa, oletusarvoisesti ei valittuna)
- **Method Descriptions and Sources =** Linkki tiedostoon, jossa kuvaillaan tutkimuskerrassa käytetty metodi ja tietolähteet (Valittavissa, oletusarvoisesti ei valittuna)
- **Published articles =** Linkki tutkimuskerran pohjalta mahdollisesti julkaistuihin artikkeleihin (Valittavissa, oletusarvoisesti ei valittuna)
- **Important variables =** Lista tutkimuskerran tärkeimmistä muuttajista (Valittavissa, oletusarvoisesti ei valittuna)

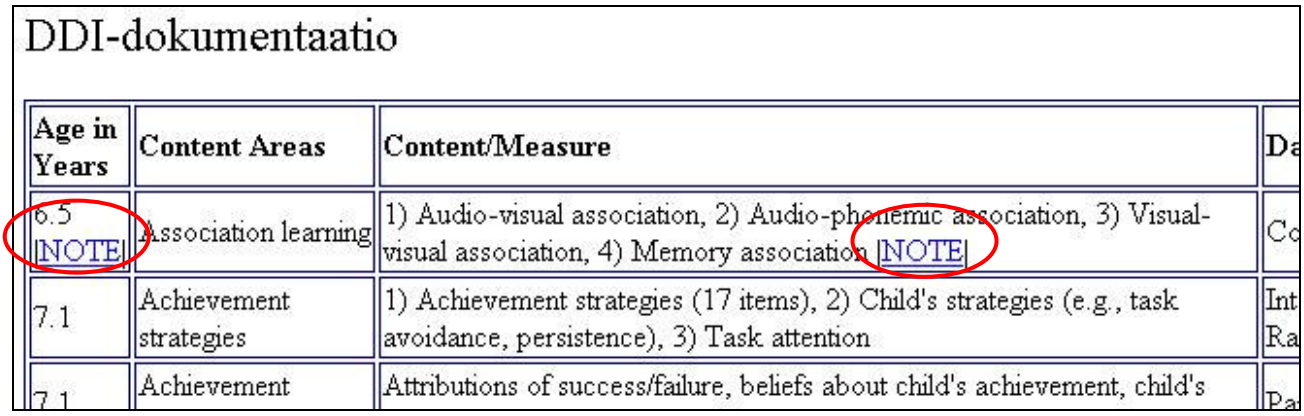
Osa tiedostojen linkkejä sisältävistä sarakkeista sisältää myös linkin ”Avaa Kansio”, jota painamalla voi avata sen tiedostokansion, jossa kyseisen tutkimuskerran tiedostot sijaitsevat.

Yläpuolella esitetyssä kuvassa olevan listan avulla voi rajata tarkastelun koskemaan vain haluttuja aihepiirejä. Oletusarvoisesti listasta on valittuna ylin arvo ”Kaikki (all)”, jolloin kaikki aihepiirit näytetään taulukossa. Voit valita yhden tai useamman aihepiirin luettelosta painamalla hiiren vasenta painiketta kyseisen aihepiirin kohdalla. Mikäli haluat valita useamman aihepiirin, pidä Ctrl-näppäin alhaalla tehdessäsi valintoja.

4) **Kun olet tehnyt valintasi, paina** 

-nappia, jolloin taulukko muodostuu valintaruudun alapuolelle.

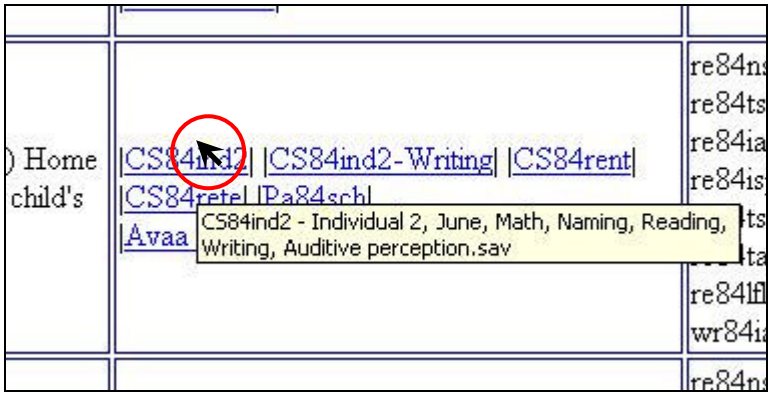
Mikäli tutkimuskertaan liittyen on tallennettu InfoPath lomakkeen avulla lisätietoja, näkyy kyseisen tutkimuskerran ”Age in Years” -sarakkeessa teksti ”NOTE”. Vastaava teksti näkyy myös ”Content/Measure” -kentässä, mikäli kyseistä mittari koskevaa lisätietoa on tallennettu.


DDI-dokumentaatio

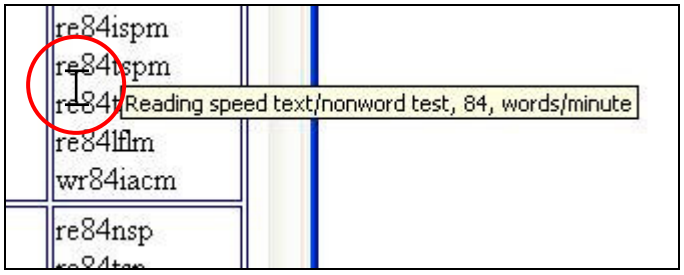
Age in Years	Content Areas	Content/Measure	Da
6.5 [NOTE]	Association learning	1) Audio-visual association, 2) Audio-phonemic association, 3) Visual-visual association, 4) Memory association [NOTE]	Co
7.1	Achievement strategies	1) Achievement strategies (17 items), 2) Child's strategies (e.g., task avoidance, persistence), 3) Task attention	Int Ra
7.1	Achievement	Attributions of success/failure, beliefs about child's achievement, child's	Pa

”NOTE” -linkkiä painamalla avataan tiedosto, johon on tallennettu kyseiseen tutkimuskertaan tai mittariin liittyvät lisätiedot.

Taulukosta on myös hyvä huomioida se, että tiedostolinkkien kohdalla, linkin kohteena olevan tiedoston koko nimi tulee näkyviin kun hiiren osoitin viedään linkin päälle.



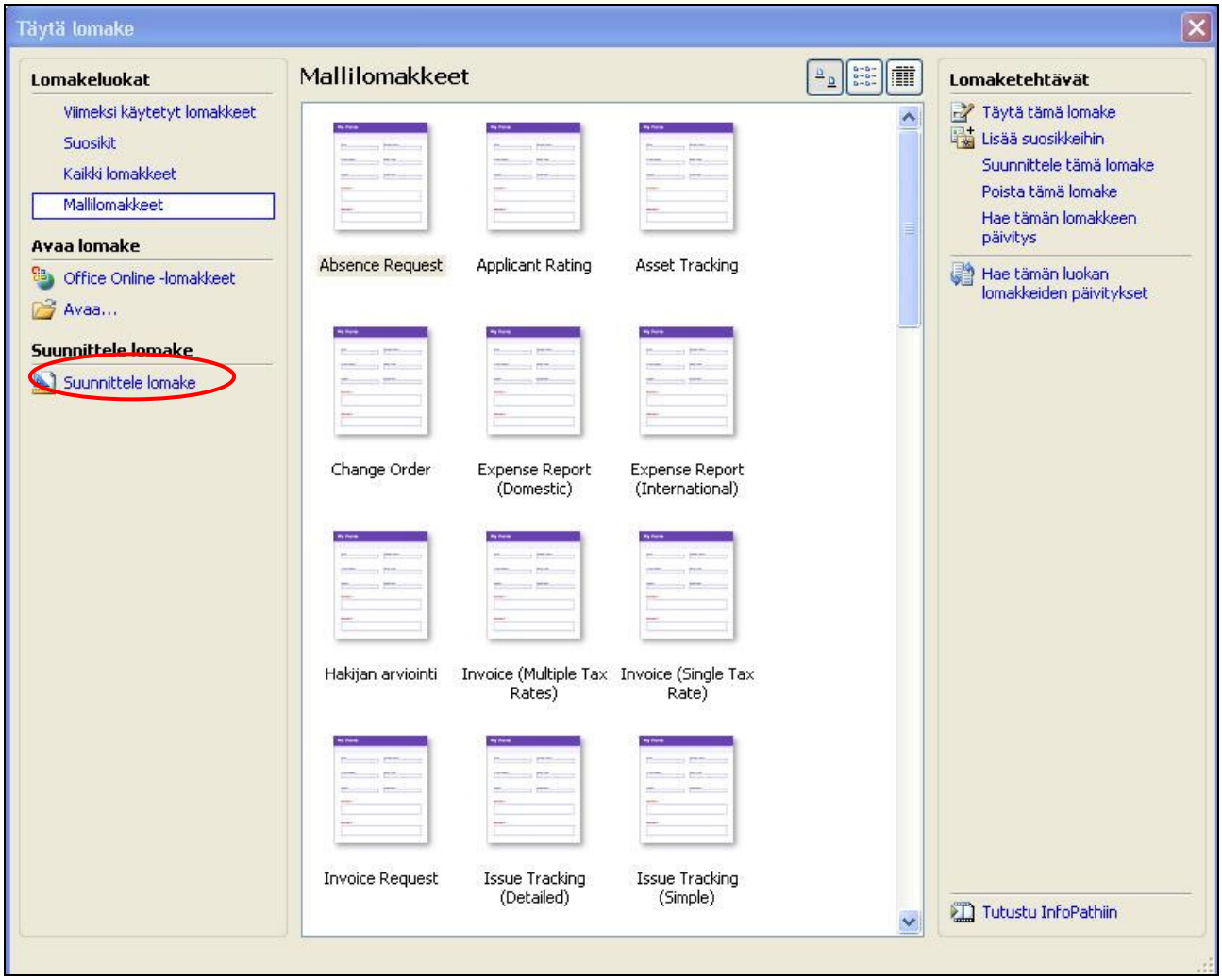
Myös tärkeimpien muuttujien kohdalla, muuttujan koko nimen saa näkyviin asettamalla hiiren kohdistimen muuttujan lyhenteen päälle.



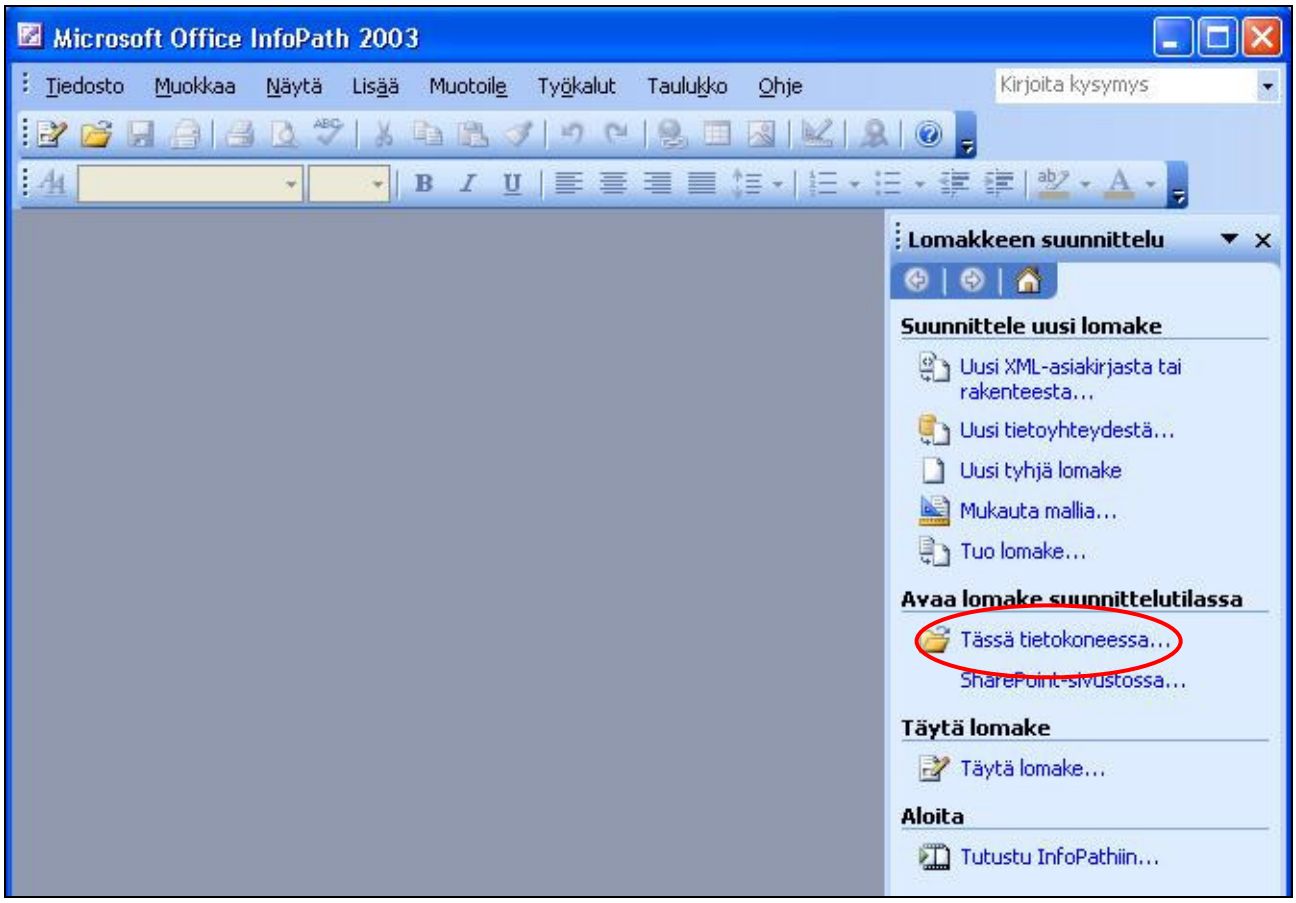
Lomakepohjan muokkaaminen

1) **Avaa Mircosoft Office InfoPath 2003 -ohjelma.** Tarkempi ohjeistus tämän ohjeen alussa.

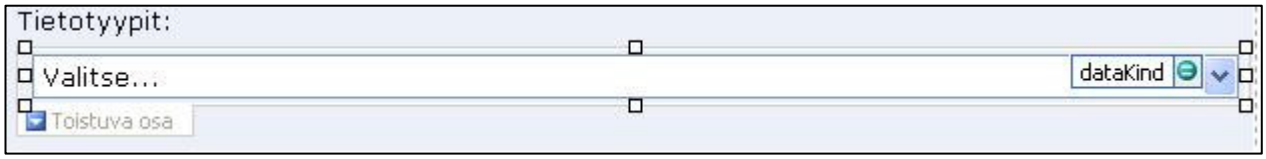
2) **Valitse päällimmäiseksi avautuvan ikkunan vasemmasta reunasta kohta *"Suunnittele lomake".***



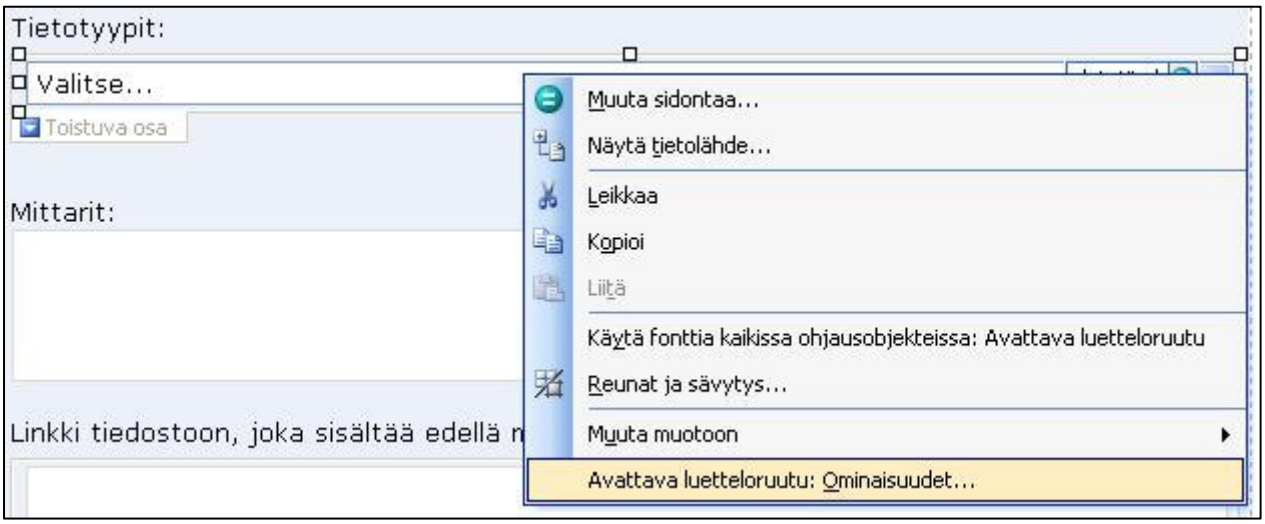
3) **Valitse oikeasta reunasta "Tässä tietokoneessa…".**



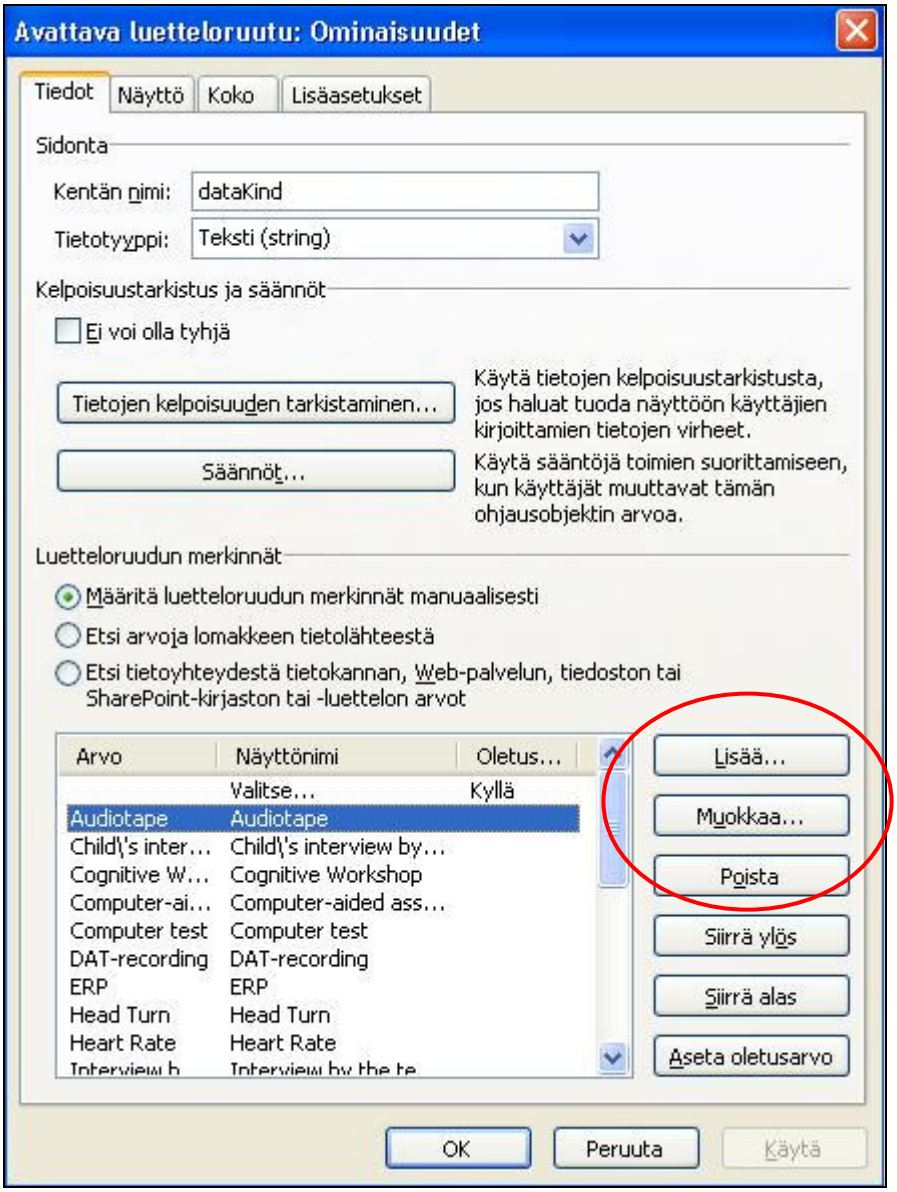
4) **Valitse lomakepohja, jota haluat muokata.**
5) **Lomakepohja on nyt avattu muokkaustilassa.** Tässä tilassa voit muokata esimerkiksi lomakkeen kenttiin liittyviä ohjetekstejä, alasvetovalikkojen sisältöjä tai lomakkeen yleistä ulkoasua. Varsinaisia kenttiä (ohjausobjekteja) tai niiden sidontaa lomakepohjan tietolähteeseen ei tule muuttaa.
6) **Alasvetovalikon muokkaaminen**:
   - Valitse alasvetovalikko aktiiviseksi painamalla sitä hiiren vasemmalla napilla.



   - Paina nyt alasvetovalikkoa hiiren oikealla napilla ja valitse avautuvasta valikosta ”Avattava luetteloruutu: Ominaisuudet”.



- Muokkausikkunan avulla voit lisätä, poistaa ja muokata valikossa olevia vaihtoehtoja.



- Vaihtoehtoja muokatessa voit asettaa erikseen arvo ja näyttönimen. Arvolla tarkoitetaan sitä arvoa, joka tallentuu XML-dokumenttiin ja näyttönimellä taas sitä arvoa, joka näkyy alasvetovalikossa. Pääsääntöisesti näiden tulee olla samoja.

7) Tallenna tehdyt muutokset.

LIITE 4

LKK-tutkimuksessa syntyvän tutkimusaineiston kuvailuun tarkoitetun InfoPath-lomakkeen tarkempi kuvailu.

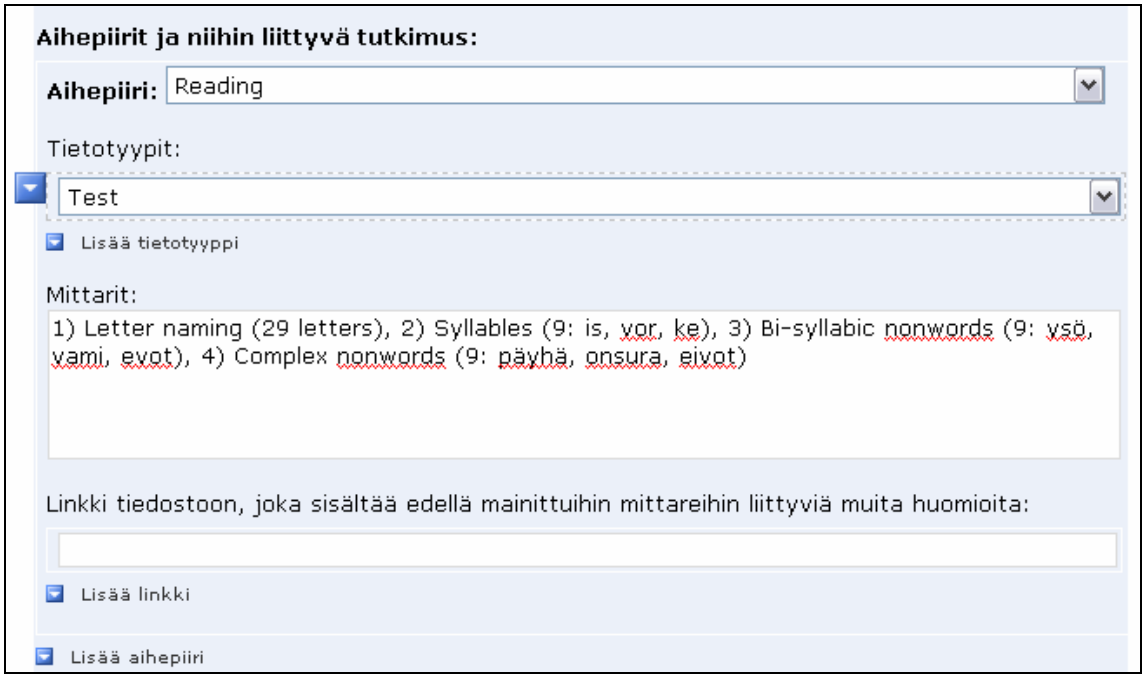
Tiedot dokumentaatiosta

- **Dokumentaation otsikko.** Otsikko tutkimuskerran metatiedot sisältävälle DDI-kuvailulle. Dokumentaation otsikoksi tulisi valita sellainen otsikko, jonka avulla voidaan tunnistaa eri tutkimuskertoihin liittyvät dokumentaatiot toisistaan.
- **Dokumentaation syöttäjä.** Dokumentaation syöttäneiden henkilöiden nimet. Tässä tulee mainita myös dokumentaatiota muokanneiden henkilöiden nimet.
- **Syöttöpäivämäärä.** Syöttöpäivämääräksi merkitään dokumentaation syöttöpäivämäärä tai päivämäärä, jolloin sitä muokattu viimeksi.
- **Kontaktihenkilö dokumentaation liittyville asioille.** Kontaktihenkilöksi merkitään se henkilö, johon otetaan yhteyttä, mikäli dokumentaatioon liittyvissä asioissa ilmenee jotain kysyttävää.

Tiedot tutkimuskerrasta

- **Yleiset tiedot.** Tämä kohta sisältää tietoja, jotka koskevat koko tutkimuskertaa.
    - **Tutkimuskerran otsikko.** Tutkimuskerran otsikoksi valitaan sellainen nimi, joka kuvaa tutkimuskertaa mahdollisimman hyvin. Otsikosta tulee ilmetä ainakin tutkittavan lapsen ikä ja mielellään myös tutkimuskerran käsittelemät aihepiirit ja tutkimusajankohta.
    - **Tutkittavan lapsen ikä vuosina.** Tähän kohtaa merkataan joko tutkittavan lapsen ikä vuosina (ennen kouluikää tehty tutkimus) tai tutkimusajankohdan numerokoodi (kouluiässä tehty tutkimus). Metadatataulukon rivit järjestetään laskevaan järjestykseen tämän kentän perusteella.
    - **Muut tutkimuskertaan liittyvät huomiot.** Tähän kohtaan voidaan tallentaa yksi tai useampi tutkimuskertaa koskeva huomio, jota ei ole dokumentoitu lomakkeen muissa kohdissa. Kohtaan voidaan syöttää myös linkki tiedostoon, josta tutkimuskertaa koskevat lisätiedot löytyvät. Tämä osio on valinnainen.
- **Tutkimusaihepiirit ja niihin liittyvät tiedot.** Tämä kohta sisältää toistettavan osion, jonka avulla kuvaillaan kaikki tutkimuskertaan liittyvät aihepiirit sekä niihin liittyvät tiedot. Kohta toistetaan jokaisen aihepiirin kohdalla erikseen.
    - **Aihepiiri.** Aihepiireillä tarkoitetaan niitä aihealueita, joita tutkimuskerrassa keskitytään tarkastelemaan. Kuvailtavat tutkimusaihepiirit valitaan alasvetovalikosta, joka sisältää listan kaikista LKK-tutkimuksen tutkimusalueista. Alasvetovalikon avulla voidaan myös varmistaa, että kaikki metatietoja syöttävät henkilöt merkitsevät aihepiirit samalla tavalla, eikä erilaisia variaatioita ilmene. Aihepiirit ovat dokumentaation tärkein osa siinä mielessä, että tutkimuskertaan liittyvät tietotyypit ja mittarit sekä muuttujatiedot liitetään aina tiettyyn aihepiiriin. Myös metadatasovelluksen avulla muodostettavan taulukon luominen perustuu aihepiireihin, koska taulukon rivit muodostetaan niiden mukaan.

- **Aihepiiriin liittyvät tietotyypit.** Tähän kohtaan merkitään yksi tai useampi edellä mainittuun aihepiiriin liittyvä tietotyyppi. Myös tietotyypit valitaan alasvetovalikosta.
- **Aihepiiriin liittyvät mittarit.** Tähän kohtaan kirjoitetaan edellä mainittuun aihepiiriin liittyvät tutkimusmittarit. Tutkimusmittarien avulla kuvaillaan niitä menetelmiä, joiden avulla aihepiiriin liittyvä tutkimusaineisto on kerätty.
  - **Edellä mainittuihin mittareihin liittyvät muut huomiot.** Tähän voidaan tallentaa kuhunkin tutkimusmittariin liittyviä lisätietoja. Myös tiedostolinkin lisääminen on mahdollista.



KUVIO 22. Tutkimusaihepiiriin liittyvien metatietojen tallennus InfoPath-lomakkeen avulla.

Tiedostojen kuvailu. Koko tiedostojen kuvailu -osio on toistuva, minkä ansiosta lomakkeen avulla voidaan kuvailla useita tutkimuskertaan liittyviä tiedostoja käyttäen seuraavia kenttiä.

- **Aihepiiri tai aihepiirit, johon tiedoston sisältämät tiedot liittyvät.** Tässä kohdassa olevan alasvetovalikon avulla voidaan määritellä mihin edellisessä kohdassa kuvailluista aihepiireistä kyseinen tiedosto liittyy. Tiedostot voidaan liittää joko kaikkiin lomakkeessa kuvailtuihin aihepiireihin tai vaihtoehtoisesti voidaan erikseen valita yksi tai useampi aihepiiri, johon tiedosto liitetään.
- **Tiedoston nimi.** Lisää tähän kohtaan syötetään tiedoston nimi. Myös tiedostopääte sisällytään tiedostonimeen.
- **Tiedoston lyhenne.** Tiedostolle tulee valita myös sopiva lyhenne. Tämä lyhenne näkyy metatietojen pohjalta muodostettavassa taulukossa ja se toimii linkkinä varsinaiseen tiedostoon.

- **Tiedoston tyyppi.** Tähän kohtaan merkitään tiedoston tallennustyyppi (esim. SPSS tai Word).
- **Kuvaus tiedoston sisällöstä.** Tähän kirjoitetaan lyhyt kuvaus tiedoston sisällöstä (esim. kerätty tutkimusdata tai testauslomake).
- **Tiedoston tallennussijainti.** Tiedoston tallennussijainti syötetään, jotta muodostettavaan metadatataulukkoon voitaisiin sisällyttää linkki kuhunkin tutkimuskertaan liittyvään tiedostoon.

Tärkeimpien muuttujien kuvailu (Muuttujien kuvailu -osio on valinnainen)

- **Aihepiiri tai aihepiirit, johon muuttujan tiedot liittyvät.** Samoin kuin tiedostojen kuvailun yhteydessä, myös muuttujatiedot voidaan liittää alasvetovalikon avulla kuuluvaksi joko kaikkiin tai yhteen tai useampaan lomakkeessa kuvailtuun aihepiiriin.
- **Muuttujan lyhenne.** Tähän merkataan SPSS-dokumenteissa käytettävä lyhenne muuttujalle. Lyhenne näkyy metatietojen pohjalta muodostettavassa taulukossa.
- **Muuttujan kuvaus (SPSS).** SPSS-dokumenteissa käytettävä kuvaus. Kuvaus näkyy taulukossa kun hiiren osoitin laitetaan muuttujan lyhenteen päälle.