Ekoinformatiikka tutkimusalana:

ekologisen datan hallinnan tarkastelua

informaatiotieteiden näkökulmasta

Sanna Lamminen

Pro gradu -tutkielma
Informaatiotutkimuksen laitos
Tampereen yliopisto
Huhtikuu 2008

TIIVISTELMÄ

TAMPEREEN YLIOPISTO
Informaatiotutkimuksen laitos
LAMMINEN, SANNA: Ekoinformatiikka tutkimusalana: ekologisen datan hallinnan tarkastelua informaatiotieteiden näkökulmasta

Pro gradu -tutkielma, 95s.
Informaatiotutkimus
Huhtikuu 2008

Tutkielmassa kartoitetaan ekoinformatiikan, eli ekologisen datan ja informaation hallinnan keskeisiä piirteitä. Tutkimusmenetelmänä on systemaattinen kirjallisuuskatsaus, jossa aineiston läpikäynnissä on käytetty sisällönanalyysin ja -erittelyn keinoja. Varsinaisena tutkimusaineistona on käytetty ekoinformatiikan alan tutkimusartikkeleita ja tukimateriaalina muuta ekologisen datan hallintaan liittyvää kirjallisuutta sekä alaan liittyvien tutkimuslaitosten internet-sivustoja. Tutkielman tarkoituksena on ekoinformatiikan alan määritelmien, sisällön, tavoitteiden ja haasteiden kuvailu ja analyysi informaatiotieteiden näkökulmasta.

Ekoinformatiikka on erittäin nuori ja määritelmällisesti vakiintumaton tutkimusala. Sen taustalla vaikuttavat ekologiassa ja teknologiassa tapahtuneiden muutosten heijastuminen ekologien ja muiden ekologista dataa tarvitsevien tahojen tiedonhallinnan käytäntöihin. Tiedonhallintaan sisältyy datanhallinnan lisäksi informaation- ja tietämyksenhallintaa, mutta ekoinformatiikassa datan asema on korostunein johtuen sekä ekologian dataintensiivisestä luonteesta että ekologisen datan ominaispiirteistä.

Ekoinformatiikka-tutkimuksen painopiste on tähän mennessä ollut datanhallinnassa ja tutkimustoimintaa on harjoitettu lähinnä ekologien ja tietojenkäsittelytieteilijöiden yhteistyössä. Ekologisen datanhallinnan muuttuminen pitkäkestoisemmaksi, datan uudelleenkäyttöön tähtääväksi osoittaa myös informaatiotutkijoiden panoksen tarpeellisuuden esimerkiksi dokumentointi- ja arkistointikäytäntöjen kehittämisessä.

Avainsanat: ekoinformatiikka, tiedonhallinta, ekologia, data, metadata, arkistointi

SISÄLLYS

1 JOHDANTO

Mitä informaatiotutkija voi tehdä ilmastonmuutoksen torjumiseksi? - Auttaa parantamaan ilmastonmuutostutkimuksen tiedonhallintaa! Muun muassa tästä tarpeesta on saanut alkunsa uusi tutkimusala, ekoinformatiikka. Termi ekologinen assosioituu arkikielessä useimmiten kierrätykseen ja muuhun ympäristöystävälliseen toimintaan. Tässä tutkielmassa termeillä ekologia ja ekologinen sen sijaan viitataan ekologian tieteenalaan ja siinä harjoitettavaan tutkimustoimintaan liittyviin seikkoihin. Kierrätys metaforana kuvastaa kuitenkin osuvasti ekologisen datan ekoinformatiikan kehityskulun myötä muuttuvaa luonnetta kertakäyttöisestä uudelleenkäytettäväksi.

Tutkielma on luonteeltaan kartoittava, sillä ekoinformatiikka ei ole entuudestaan informaatiotutkimuksen alalla tunnettu tutkimusalue. Se on kuitenkin luontevasti sijoitettavissa tieto- ja asiakirjahallinnon piiriin, kuten tutkielmasta käy ilmi. Tutkielma on toteutettu systemaattisena kirjallisuuskatsauksena (Tuomi & Sarajärvi 2002 119-121), jossa aineistoa on käyty läpi sisällönerittelyn ja -analyysin keinoin tavoitteena saada abstrahoiduksi aineistosta mahdollisimman kattava ja selkeä kuvaus ekoinformatiikasta tutkimusalana. Tutkielman lähtökohtana oli oletus alasta saatavilla olevan informaation hajanaisuudesta ja suomenkielisen aineiston puuttumisesta. Tutkielman tavoitteena on näin ollen sekä koota yhteen informaatiota että herättää tutkimuskiinnostusta ekoinformatiikkaa kohtaan informaatiotutkijoiden piirissä.

Tutkielman pääongelmana on ymmärtää mitä ekoinformatiikka on. Tarkemmin ottaen haluttiin selvittää sitä, kuinka ekoinformatiikka on määriteltävissä, minkälaisiin osa-alueisiin ala on jaettavissa ja minkälaisiin tavoitteisiin ja haasteisiin tutkimusala

pyrkii vastaamaan. Pohjimmaisena tavoitteena oli saada selville, minkälainen rooli informaatiotutkijoilla tämän alan tutkimustoiminnassa voisi olla. [1]

Tutkielman aineistona on sekä ekoinformatiikkaa käsitteleviä tutkimusartikkeleita, että yleisemmin ekologista dataa ja sen hallintaa käsittelevää kirjallisuutta. Varsinaisessa sisällönerittelyssä eli alan keskeisten piirteiden selvittelyssä käytettiin kuitenkin vain sellaista aineistoa, jossa alan nimikin esiintyi joko otsikossa, asiasanoissa, itse tekstissä tai useammassa näistä paikoista. Muuta tutkimusaineistoa käytettiin lisätiedon saamiseksi sisällönerittelyssä merkittäviksi havaittuihin asioihin. Tutkimusaineisto kerättiin etsimällä aluksi käsiin muutama ekoinformatiikkaa käsittelevä artikkeli, ja laajentamalla aineiston kokoa näistä saatuja lähdeviitteitä seuraamalla. Kattavan otoksen takaamisessa käytettiin apuna saturaatioperiaatetta. Perimmäisenä ajatuksena saturaatiossa on, että tietty määrä aineistoa riittää tuomaan esiin tutkimuskohteen perusluonteen (Tuomi & Sarajärvi 2002, 89). Artikkeleiden alettua toistuvasti viittaamaan toisiinsa ja sisällöllisesti samojen teemojen tullessa vastaan, arvioitiin oleellisen aineiston olevan kasassa. Tutkielmassa ei siis pyritty kartoittamaan kaikkia ekoinformatiikan alalla tehtyjä tutkimuksia ja kaikkia mahdollisia piirteitä, vaan löytämään ekoinformatiikan keskeisin olemus ja keskeisimmät tutkimusteemat. Muun muassa erikseen esiteltyjen ekoinformatiikka-projektien osalta taustoja ja tuoreinta tietoa on selvitetty kirjallisuuden lisäksi projektien ja tutkimuslaitosten verkkosivuilta.

[1] Kiinnostukseni aiheeseen kumpuaa koulutustaustastani, johon sisältyy informaatiotutkimuksen lisäksi insinööri (AMK) –tutkinto ympäristöteknologiasta (HAMK 2001) sekä ympäristötieteiden yliopisto-opintoja.

Aineiston analyysimenetelmäksi valittiin sisällönerittely, koska se soveltuu erinomaisesti nimenomaan tutkimuksiin, joissa pyritään löytämään keskeisin aines kirjallisesta aineistosta. Pietilän (1973) mukaan sisällönerittelyä käyttävissä tutkimuksissa pyritään joko tilastollisesti tai sanallisesti kuvailemaan joko dokumenttien sisältöä ilmiönä sinänsä tai niitä ulkopuolisia ilmiöitä, joita dokumenttien sisällön ajatellaan ilmaisevan. Sisällönerittely voidaan käsittää joukoksi menettelytapoja, joiden avulla dokumenttien sisällöstä tieteellisiä pelisääntöjä noudattaen tehdään havaintoja ja kerätään tietoja. Tutkimusongelma kulloinkin määrittelee mistä ilmiöstä ja missä muodossa tietoja kerätään ja mihin tarkoitukseen niitä käytetään – ilmiön kuvailuun vai selittämiseen. (Pietilä 1973, 52-55) Tässä tutkielmassa tutkimusaineiston sisältöä on kuvailtu vain sanallisesti, joka voidaan mieltää myös sisällönanalyysiksi, mutta havaintoja tehtäessä on pyritty käyttämään hyväksi käsitteiden ja termien esiintymismääriä artikkeleissa.

Tarkimman sisällönerittelyn kohteeksi valikoitui laajin ja kattavin yleisartikkeli ekoinformatiikan alasta; Jones et al. (2006) The new bioinformatics: integrating ecological data from the gene to the biosphere. Artikkelista muodostui lopulta tutkielman selkäranka sikäli, että muista lähteistä tehtyjä havaintoja ja tulkintoja peilattiin sitä vasten. Toinen merkittävä lähde tutkielman kannalta oli Michenerin ja Bruntin (2000) toimittama kokoelmajulkaisu Ecological Data: Design, Management and Processing, joka toimi aloituslähteenä ekologisen datan problematiikan selvittämiseen.

Tutkimusaineistoksi valittujen artikkeleiden läpikäynnissä kiinnitettiin erityisesti huomiota tiettyihin avainsanoihin kuten tavoite, päämäärä, haaste, ongelma, este,

ratkaisu tai avain. Näitä sanoja sisältävien ilmausten avulla pyrittiin löytämään vastaus ekoinformatiikan tavoitteita ja haasteita koskeviin tutkimuskysymyksiin sekä erittelemään keskeiset tutkimukselliset osa-alueet. Artikkeleista poimittiin yleisimmin toistuvat teemat, jotka tulkittiin keskeisimmin ekoinformatiikkaan liittyviksi.

Tutkielmassa lähdetään liikkeelle ekoinformatiikan alan määrittelystä tutkimus-kirjallisuudessa sekä sen suhteesta muihin aloihin. Tämän jälkeen tarkastellaan ekoinformatiikan jakautumista kolmeen eri ulottuvuuteen – ekologiaan, teknologiaan ja informaatioon, jotka muodostavat alan peruselementit. Seuraavaksi valotetaan ekoinformatiikan tutkimuksen keskeisimpiä toimijoita ja tutkimushankkeita. Tämän jälkeen eritellään tarkemmin ekologisen datanhallinnan vaiheita ja teemoja. Lopuksi esitetään johtopäätöksiä.

2 EKOINFORMATIIKAN MÄÄRITELMÄ

Englanninkielessä ekoinformatiikasta käytetään yleisimmin termiä ecoinformatics tai sen pidempää muotoa ecological informatics. Joskus näkee käytettävän myös nimitystä ecosystem informatics. Suomenkielessä termiä ekoinformatiikka ei yleisesti tunneta. Termi olisi kuitenkin luonteva valinta, sillä se on muodostettu samalla logiikalla kuin lähitieteenalalle ’bioinformatics’ suomenkieleen vakiintunut nimitys. Ekologia on osa biologiaa ja jos biologisesta informatiikasta käytetään suomenkielessä nimitystä bioinformatiikka on ekologisen informatiikan luonnollisesti oltava ekoinformatiikkaa.

Jørgensenin (2002) mukaan ekoinformatiikka on ekologian alatiede, joka on virallisesti hyväksytty erilliseksi tutkimusalaksi vuonna 2000. Tämä on ainoa löydetty lähde, jossa alan syntyajankohta määritellään. Asian paikkansa pitävyyden puolesta puhuu kyllä vahvasti se, että vaikka tämänkin tutkimuksen lähdeaineistona on ennen vuotta 2000 julkaistua aineistoa, joka on sekä sisältönsä puolesta että myöhemmin julkaistujen artikkeleiden lähdeviittausten perusteella tunnistettavissa ekoinformatiikan piiriin kuuluviksi, ei varsinaisia ekoinformatiikan nimissä tehtyjä tutkimuksia esiinny kuin vasta 2000-luvulla. Tätä ennenkin ekoinformatiikaksi luokiteltavaa tutkimustoimintaa on jo harjoitettu ainakin laskennallinen ekologia (computational ecology) –nimikkeen alla (esim. Helly et al. 1999).

Ekoinformatiikka on näin ollen nuori tutkimusala, mutta sen taustalta löytyvä informatiikka on jo hieman vanhempi käsite. Informatiikan alkujuuret on löydettävissä 1960-luvulta, jolloin se yhdistettiin enemmän informaatiotieteisiin kuin tietojen-käsittelytieteisiin. 1970-luvun lopulta lähtien termi informatiikka on kuitenkin

enenevässä määrin omaksuttu kuvaamaan nimenomaan informaatioteknologian soveltamista eri tieteenaloilla (He 2003, 117-118).

2.1 Ekoinformatiikan määritelmiä kirjallisuudessa

Ekoinformatiikalle ei ole olemassa vakiintunutta määritelmää, vaan erilaisia määritelmiä lienee yhtä monta kuin määrittelijöitäkin. Etsittäessä määritelmiä tutkimuskirjallisuudesta, pitäydyttiin niissä määritelmissä, joissa ekoinformatiikan kirjoitusasuna käytettiin muotoja eco(-)informatics tai ecological informatics. Artikkelien läpikäynnissä huomattiin, etteivät ekoinformatiikan alan tutkimuksia valottavat artikkelit usein sisällä ekoinformatiikan suoraa määritelmää lainkaan, vaikka termiä ekoinformatiikka niissä käytettiinkin. Sen sijaan määritelmiä on jonkin verran löydettävissä erilaisista konferenssiesitelmistä ja ekoinformatiikan alaan liittyvien yhdistysten tai tutkimuslaitosten internet-sivuilta.

Löydetyt määritelmät eroavat muun muassa sen suhteen mihin vaiheeseen ekologisen datan elinkaarta ne keskittyvät. Tutkimusaineiston etsintävaiheessa syntyi vaikutelma, että Euroopassa ja Amerikassa vallitsee tässä suhteessa eroa sikäli, että Euroopassa on ensisijainen mielenkiinto kohdistunut tietoteknisten ratkaisujen kehittämiseen nimenomaan datanhallinnan niin sanottuun aktiivivaiheeseen, eli keräämisessä, analysoinnissa, mallinnuksessa ja visualisoinnissa tarvittaviin teknologisiin välineisiin, kun taas Amerikassa on kiinnostuttu ennen kaikkea kerran kerätyn datan säilymiseen ja uudelleenkäytön mahdollisuuksien edistämiseen liittyvästä tutkimuksesta. Tämän

kokonaisvaltaisemman otteen vuoksi tämän tutkimuksenkin aineisto muodostui amerikkalaispainotteiseksi.

Eurooppalaista ekoinformatiikkaa edustava Recknagel (2002) on määritellyt alan perusteoksen esipuheessa ekoinformatiikan tieteidenväliseksi viitekehykseksi, joka edistää kehittyneiden laskennallisten teknologioiden (esim. sumea logiikka) käyttöä minkä tahansa ekosysteemin kompleksisuustasoa koskevan informaation (geeneistä ekologisiin verkostoihin) käsittelyn periaatteiden selventämiseksi sekä ekologiseen vakauteen, biodiversiteettiin ja ilmastonmuutokseen liittyvän päätöksenteon tukemiseksi. Erillisiksi ekoinformatiikan osa-alueiksi Recknagel luettelee vielä datan integroinnin eri ekosysteemikategorioiden ja kompleksisuustasojen välillä, päättelyn datamalleista ekologisiin prosesseihin sekä ekosysteemien simuloinnin ja ennustamisen (Recknagel 2002). Analysointikeskeisyytensä lisäksi määritelmä on teknologiapainotteinen.

Toisenlaisen mutta myöskin datanhallinnan aktiivivaiheeseen keskittyvän eurooppalaisen näkemyksen esittää Jørgensen (2002), jonka mukaan ekoinformatiikka tulisi määritellä tieteeksi, joka tutkii tapoja tuottaa ekologista informaatiota. Täten ekoinformatiikan piiriin kuuluisivat muun muassa internetissä oleva ekologinen informaatio, ekologiset tietokannat sekä niiden luominen ja kehittäminen, ekologinen tilastotiede, mallien käyttäminen ekologisen informaation tuottamiseen, parametrien estimointimallit, ekologiassa sovellettava tietotekniikka sekä epävarmuus ja ekologinen data.

Niin sanottua amerikkalaista koulukuntaa edustavan määritelmän mukaan (SEEK 2004) ekoinformatiikka on ekologisen informaation luontaisen rakenteen tutkimista, tähtäimenä tämän informaation hallintaan ja analysointiin tarvittavan tietotekniikan luominen ja soveltaminen. Erityisesti ekoinformatiikassa on saman lähteen mukaan kyse tietokantojen ja algoritmien kehittämisestä laajan mittakaavan ekologisen tutkimuksen helpottamiseksi ja tehostamiseksi. Keskeisenä erona edellisiin tässä määritelmässä puhutaan analysoinnin ohella tiedonhallinnasta, jota toteutetaan teknologiaa apuna käyttäen ja tähtäimenä ekologisen tutkimuksenteon parantaminen.

Selkeämmin amerikkalaisen ekoinformatiikan eurooppalaisesta ekoinformatiikasta poikkeava luonne tulee ilmi alan nimeä kantavan lehden esittelyssä (Ecoinformatics), jossa sanotaan ekoinformatiikkaan kuuluvan ekologisen datan hallintaan, arkistointiin, ylläpitoon, löytämiseen, hakemiseen, integrointiin, analysointiin, syntetisointiin ja ennustamiseen sopivien järjestelmien kehittäminen, soveltaminen ja koordinointi.

Edellä esiteltyjen määritelmien taipumuksena on ekoinformatiikan sisällön laajuuden korostaminen. Toinen yleinen trendi esitetyissä määritelmissä on selittää ekoinformatiikkaa siinä mukana olevien tutkimusalojen avulla. Tutkimusaloihin tukeutuvat määritelmät ovat tyypillisesti edellä mainittuja suppeampia ja tähän kategoriaan voidaan sijoittaa muun muassa Michenerin (1997) antama määritelmä, jonka mukaan ekoinformatiikka on yksinkertaisesti tietokoneiden soveltamista ekologiseen tietohallintoon.

Jones et al.:n (2006, 520) mukaan ekoinformatiikka on ekologian, tietojenkäsittelytieteen ja informaatioteknologian yhtymäkohtaan sijoittuva

tutkimusala. Toisin sanoen kyseessä on informaatioteknologian ja tietojenkäsittely-tieteiden soveltamista ekologiaan (LTER 2005a, 9). Yleisimmin määritelmissä puhutaan tähän tapaan ekologiasta ja teknisistä tieteistä, mutta myös laajempia valikoimia mukana olevista tieteenaloista on esitetty.

Wilson (2007b) esimerkiksi kokee, että ekoinformatiikassa informaatioteknologia ja tietojenkäsittelytiede (eli informatiikka) yhdistyy matematiikkaan ja tilastotieteeseen, joiden avulla kehitetään innovatiivisia tapoja kerätä, järjestää, asettaa saataville, analysoida ja tulkita ekologista dataa. On myös väitetty (Wyoming 2008), että ekoinformatiikka on ekologian, tietojenkäsittelytieteiden, paikkatietotekniikan ja kvantitatiivisten menetelmätieteiden välinen tutkimusala.

2.2 Ekoinformatiikan määritelmä tässä tutkielmassa

Ekoinformatiikasta esitetyistä määritelmistä käy selville, että alan tutkimus- ja kehitystyötä voidaan tehdä monienkin eri alojen yhteistyönä. Alojen kattava luetteleminen alan määritelmissä ei kuitenkaan ole välttämättä järkevää. Määritelmistä puuttuu yleensä esimerkiksi sosiaalitieteet, joilta kuitenkin koetaan saatavan apua muun muassa ekologien asennemuutosten toteuttamiseksi. Toisaalta liiallinen teknologian painottaminen on sikäli turhaa, että teknisiä apuvälineitä käytetään nykypäivänä jo niin yleisesti alalla kuin alalla ja informaatioteknologian mukanaolo tulee myös ilmi jo nimen informatiikka-osasta.

Ekoinformatiikassa mukana olevien tutkimusalojen tai teknologisten ratkaisujen kehittämisen painottamisen sijasta olisi houkuttelevampaa nostaa selkeämmin määritelmän keskiöön ekologinen data ja informaatio ja nimenomaan näiden hallinta, joka on keskeisin erottava tekijä lähitieteenaloihin nähden. Tässä tutkielmassa ekoinformatiikka onkin määritelty yksinkertaisesti ekologisen datan ja informaation hallinnan edistämiseen tähtääväksi tieteidenväliseksi tutkimusalaksi.

2.3 Ekoinformatiikan suhde lähitieteenaloihin

Koska ekoinformatiikka terminä samoin kuin alana on määritelmällisesti vielä vakiintumaton, saatetaan se herkästi sekoittaa esimerkiksi ympäristöinformatiikkaan tai alana jo hieman vakiintuneempaan bioinformatiikkaan. Toisaalta on hyvä huomata, että lähialoilla on myös paljon yhtäläisyyksiä ja päällekkäisyyksiäkin ekoinformatiikan kanssa, eivätkä alojen rajat näin ollen ole tarkat. Alanimitysten käytössä onkin havaittavissa horjuvuutta esimerkiksi siinä, että ekologisesta datasta saatetaan puhua minkä tahansa alan yhteydessä ja vastaavasti harhaanjohtavasti käyttää ekoinformatiikka-nimitystä esimerkiksi ympäristödatan yhteydessä. Ekoinformatiikka-tutkimusta ei tulisikaan tehdä täysin eristyksissä vaan lähialoilla tehdyistä yrityksistä ja erehdyksistä voi ottaa oppia ja voimavaroja voidaan ainakin jossain määrin yhdistää. Esiteltävät lähitieteenalat on valittu sillä perusteella, että ne esiintyvät useimmin ekologisen datan ja / tai ekoinformatiikan yhteydessä.

2.3.1 Bioinformatiikka

Ekoinformatiikan lähialoista tunnetuin lienee bioinformatiikka (biological informatics tai bioinformatics). Bioinformatiikka on terminä suomen kielessäkin jo suhteellisen tuttu. Alan tutkimusta ja opetusta harjoitetaan myös Suomessa, esimerkiksi Helsingin ja Turun yliopistoissa. Helsingin yliopiston opintoesitteessä bioinformatiikka esitellään monitieteiseksi tutkimusalaksi, joka kehittää laskennallisia malleja ja tietojenkäsittelymenetelmiä biologisten sovellusten tarpeisiin. Tarkemmin määriteltynä bioinformatiikkaa voidaan kuvata biologisen ja lääketieteellisen informaation tietokoneavusteiseksi keräämiseksi, prosessoinniksi ja analysoinniksi (Helsingin yliopisto 2005).

Bioinformatiikka voidaan myös käsittää eräänlaiseksi kattomääritteeksi, jonka alle ekoinformatiikkakin kuuluu; onhan ekologia eräs biologian haara ja näin ollen ekologinen data myös tietyntyyppistä biologista dataa (Kalra 2005, 335). Käytännössä bioinformatiikka kuitenkin mielletään ensisijaisesti geneettistä tai muuta lääketieteeseen liittyvää biologista dataa koskevaksi.

2.3.2 Ympäristöinformatiikka

Ympäristöinformatiikka (environmental informatics) on myös alanimityksenä suomen kielessä käytössä ja bioinformatiikan tavoin ympäristöinformatiikkaan liittyvää opetus- ja tutkimustoimintaakin on Suomessa jo jonkin verran. Ympäristöinformatiikassa yhdistetään ympäristötieteet ja informaatioteknologia erilaisten ympäristökysymysten ratkaisemiseksi. Alan tutkimusta ja opetusta harjoittavassa Kuopion yliopistossa

ympäristöinformatiikka määritellään alaksi, joka kehittää menetelmiä ympäristöön liittyvien suurten tietomassojen analysointiin ja jalostamiseen eri loppukäyttäjille soveltuvaan muotoon (Kuopion yliopisto). Ympäristöinformatiikka muistuttaa datan keräämisen analysoinnin tehostamiseen keskittymisessään eurooppalaista lähestymistapaa ekoinformatiikkaan sillä erotuksella, että ekologisen datan sijaan käsittelyn kohteena on esimerkiksi ilmanlaatuun tai jätehuoltoon liittyvää dataa.

2.3.3 Biodiversiteetti-informatiikka

Biodiversiteetti-informatiikka (biodiversity informatics) on myös suomen kielessä jo vähäisessä määrin käytetty alanimitys. Termi sisältyy esimerkiksi vuonna 2007 ilmestyneeseen luonnon monimuotoisuuden käsikirjaan (Salo & Sääksjärvi), jossa biodiversiteetti-informatiikan tavoitteeksi mainitaan biodiversiteettitiedon järjestäminen ja kokoaminen helpommin käytettävissä olevaan muotoon tietoverkkoon. Biodiversiteetti-informatiikka on hyvin lähellä ekoinformatiikkaa, onhan biodiversiteetti oleellinen osa ekologiaa ja keskeinen ekologisen tutkimuksen kohde. ‘Biodiversity Informatics’ -lehden mukaan biodiversiteetti-informatiikalla tarkoitetaan biologiseen monimuotoisuuteen liittyvän informaation luomista, integroimista, analysointia ja ymmärtämistä. Biodiversiteetti-informatiikka ei siten olisi yhtä kokonaisvaltaista datan ja informaation hallintaa kuin ekoinformatiikka ja myös keskittyy suppeampaan data-ainekseen. Biodiversiteetti-informatiikka voidaankin mieltää eräänlaiseksi ekoinformatiikan alalajiksi.

Biodiversiteetti-informatiikasta saatetaan toisinaan käyttää nimitystä ´ecoinformatics´, koska alan oma nimi on niin pitkä ja lyhenne ´bioinformatics´ on jo varattu biologiselle

informatiikalle (Wilson 2007b). Ekoinformatiikan ja biodiversiteetti-informatiikan erottaminen ei aina välttämättä olekaan tarpeen, mikäli ei ole syytä painottaa, että ollaan tekemisissä nimenomaan biodiversiteettidatan eikä minkään muun ekologisen datatyypin kanssa.

Biodiversiteetti-informatiikkaan liittyvää tutkimusta tehdään Suomessakin jossain määrin. Esimerkiksi Turun yliopiston biologian laitos on mukana eurooppalaisessa ENBI-verkostossa (European Network for Biodiversity Information), kokoamassa laajaa havaintoaineistojen tietopankkia (Turun yliopisto). Biodiversiteetti- ja ekosysteemi-informatiikka (biodiversity and ecosystem informatics, BDEI) on myös kirjallisuudessa usein näkyvä alanimitys, joka saatetaan lähteestä tai käyttäjästä riippuen rinnastaa yhtä hyvin biodiversiteetti-informatiikkaan kuin ekoinformatiikkaan.

2.3.4 Meri-informatiikka

Meri-informatiikka (ocean informatics, OI) on keskittynyt nimenomaan oseanografisen datan hallintaan ja on siten biodiversiteetti-informatiikan tavoin ekoinformatiikkaa suppeampi ja ekoinformatiikan alaisuuteen katsottavissa oleva ala, onhan meri yksi ekologian tutkimista ekosysteemeistä. Satunnaisesti puhutaan myös nimenomaan meren biodiversiteetti-informatiikasta (ocean biodiversity informatics, OBI), joka voidaan määritellä tietoteknologian käyttämiseksi nimenomaan merta koskevaan biodiversiteetti-informaatioon, kuten datan taltioimiseen, tallettamiseen, hakemiseen, visualisointiin, kartoitukseen, mallinnukseen, analysointiin ja julkaisemiseen (esim.

Costello & Berghe 2006, 203). Tutkimusalan toimintorepertuaari on varsin laaja, mutta käsiteltävä datatyyppi on huomattavasti rajallisempi kuin ekoinformatiikassa.

3 EKOINFORMATIIKAN KOLMIJAKOINEN TAUSTA

Käsitys ekoinformatiikan kolmijakoisesta taustasta syntyi sekä ekoinformatiikasta esitettyjen määritelmien pohjalta että tutkimusaineiston tarkemman sisällönerittelyn myötä. Tutkimuksen puitteissa tarkastelluissa ekoinformatiikan määritelmissä esiintyi vahvimmin kolme teemaa, joihin tutkimukseen valikoituneissa artikkeleissa käytetty termistökin oli karkeasti jaettavissa, nimittäin ekologinen tutkimus, informaatio ja teknologia. Nämä kolme ulottuvuutta muodostavat siten ne tukijalat, joiden päälle ekoinformatiikka on kehittynyt. Yksittäisissä ekoinformatiikkatutkimuksissa painottuu toisinaan jokin puoli toisia enemmän, riippuen todennäköisimmin siitä minkä alan lähtökohdista tyypillisesti tieteidenvälistä ekoinformatiikkatutkimusta kulloinkin tehdään, mutta kaikki puolet ovat ekoinformatiikassa aina väistämättä mukana. Seuraavassa tarkastellaan ekoinformatiikan eri lähtökohtia ja tutkimusnäkökulmia.

3.1 Ekologia

Ekologia on biologian alatieteisiin kuuluva varsin laaja luonnontiede, joka eriytyi virallisesti omaksi tieteenalakseen 1800-luvun lopulla (Hanski et al. 1998, 34 ja 472). Ekologian alasta voidaan esittää monenlaisia määritelmiä. Paitsi eliöiden levinneisyyden ja runsauden tutkimiseksi, ekologia voidaan määritellä myös luonnon rakenteen toiminnan tutkimiseksi tai opiksi eliöistä ja niiden suhteista elolliseen ja elottomaan ympäristöönsä. Ekologian tutkimustraditiot ovat olleet yhtä moninaisia kuin tutkimuksen kohteena oleva elävä luonto (Hanski et al. 1998, 21).

Ekologian ensisijainen tehtävä on tuottaa tietoa, jota yhteiskunta voi halutessaan käyttää ratkaisuja tehdessään. Ekologian tehtävä voidaan ymmärtää monella biologisen hierarkian tasolla (yksilö, populaatio, eliöyhteisö, ekosysteemi) ja ekologia voidaan jakaa moniin eri osa-alueisiin esimerkiksi juuri tämän hierarkiatason mukaan ja yksilötason tutkimuksissa tutkimuskohteena olevan eliölajin mukaisesti muun muassa kasvi- ja eläinekologiaan. (Hanski et al. 1998, 13 ja 34) Korkeammilla hierarkiatasoilla voidaan puolestaan puhua esimerkiksi järvi- tai suoekologiasta.

Koko ekologian kentässä jokainen osa-ala on saanut jossain määrin vaikutteita fysiikasta ja matematiikasta. Matemaattisten mallien käyttö on välttämätöntä silloin, kun tavoitteena on tutkittavien ilmiöiden käsitteellistäminen ja mallintaminen. Matemaattisia malleja tarvitaan ekologiassa apuna esimerkiksi monimutkaisia vuorovaikutussuhteita, kuten ravintoverkostoja tutkittaessa. Ekologia on läheisessä vuorovaikutuksessa myös monien muiden lähitieteidensä, kuten taksonomian eli luokitteluopin kanssa. (Hanski et al. 1998, 38-39)

Ekologian historiassa on jo yli 200 vuotta ollut vallitsevana havainnoimalla tapahtuva perustietojen kerääminen yksittäisistä kasvi- ja eläinlajeista (Hanski et al. 1998, 40). Toinen 1960-luvulla yleistynyt tutkimustyyppi ekologiassa on kokeellinen tutkimus (Jones et al. 2006, 520; Wilson 2007a, 1). Kokemusperäisen tiedon ohella luonnontiede tarvitsee myös teoriaa. Teoreettisella ekologialla tarkoitetaan ekologisten vuorovaikutussuhteiden käsitteellistä ja matemaattista tutkimusta. Teoreettisen tarkastelun perustana on pitkälle kehitetty hypoteesi, teoria ilmiöiden välisistä suhteista, joille on jo olemassa kokemusperäistä tai aikaisempaan tutkimukseen perustuvaa tukea. Luonteeltaan soveltavaksi ekologia muuttuu, kun ekologisista

tutkimuksista saatujen tulosten perusteella tehdään ihmisten toimintaa muuttavia päätöksiä. Soveltavan ekologian pyrkimyksenä on laajasti ottaen luonnonvarojen kestävä käyttö. (Hanski et al. 1998, 471-472)

3.1.1 Ekologinen tutkimus

Ekologit ovat omaksuneet monia tutkimuksellisia lähestymistapoja ja olleet osittain jopa edelläkävijöitä niiden harjoittamisessa. Jotkut lähestymistavoista sopivat parhaiten tietyn tyyppisten ongelmien ratkaisemiseen ja toiset taas ovat yleisemmin käytettäviä. Ekologialla on esimerkiksi vahva menneisyys innovatiivisten laboratorio- ja kenttäkokeiden käyttämisessä. (Michener 2000c, 7) Erityisen sopivia moniin ekologien esittämiin kysymyksiin vastaamisessa ovat myös pitkäkestoiset tutkimukset, sillä ekologiassa tutkitaan usein erilaisia hitaita, herkkiä tai suurta vaihtelua sisältäviä prosesseja, harvinaisia tapahtumia sekä jaksottaisia tai kompleksisia ilmiöitä. Muun muassa ekologisten tutkimusten tavanomaisista rahoituskäytännöistä johtuen kovin pitkäkestoisia tutkimusprojekteja ei ole yleisesti mahdollista toteuttaa, vaan ekologista ymmärrystä tuotetaan tavallisemmin useampien lyhytkestoisten projektien tulosten synteesitutkimuksissa. Uusia näkemyksiä ja uutta ymmärrystä sekä parempaa ennustettavuutta ekologiassa voidaan saavuttaa laaja-alaisten vertailututkimusten avulla. Fokusoidumpien kokeellisten tutkimusten tuottaman mekanistisen ymmärryksen sijaan vertailututkimuksissa tavoitellaan yleisten mallien tunnistamista. (Michener 2000c, 10-11) Ylipäätään ekologisessa tutkimuksessa historialliset muutokset ovat avain pyrittäessä ymmärtämään nykyistä tilannetta ja ennakoimaan tulevaa (Karasti & Baker 2004, 1).

Perinteinen ekologinen tutkimusprojekti on käytännössä käsittänyt tutkimus-kysymyksen muuntamisen yhdeksi tai useammaksi testattavaksi hypoteesiksi, sopivan kenttä- tai laboratoriokokeen suunnittelun, datan keräämisen, analysoinnin ja tulkinnan sekä tulosten julkistamisen. Tämän projektin on usein suunnitellut ja toteuttanut yksi tutkija tai pieni tutkimusryhmä ja tutkimuskohteena on tyypillisesti ollut yksi tai muutama eliölaji. (Michener 2000c, 2-3) Ekologiset kenttätutkimukset ovat tyypillisesti paljon työvoimaa vaativia, mikä on tällaisissa pienissä tutkimus-projekteissa rajoittanut tutkittavan alueen kokoa.

3.1.2 Muutokset ekologisessa tutkimuksessa

Yhden tutkijan tai pienen tutkimusryhmän tekemät tutkimukset eivät ole menettäneet merkitystään, mutta ekologisessa tutkimuksenteossa on meneillään muutoksia, jotka vaativat enenevässä määrin myös laajemman ekologiyhteisön ja tieteenalarajat ylittävässä yhteistyöverkostoissa tehtyjä tutkimuksia.

Ensinnäkin ekologinen tutkimus on alkanut laajentua temaattiselta fokukseltaan, ekologien kiinnostuttua osaltaan yhteiskunnassa ja tiedemaailmassa paljon huomiota saaneisiin aiheisiin kuten ilmastonmuutos, biodiversiteetin väheneminen ja ympäristön vakaus. Tämän seurauksena ekologien esittämät tutkimuskysymykset ovat laajentuneet ja siten myös tutkimusprojektit ovat kasvaneet. Aiemmin projektit kestivät rahoitus-käytännöistä johtuen korkeintaan kolme vuotta, mutta jo nyt joitakin ilmiöitä tutkitaan pitkäkestoisissa ekologisissa tutkimusohjelmissa, jotka on suunniteltu jatkumaan vuosikymmeniä. Ekologit ovat alkaneet etsiä vastauksia uusiin tutkimuskysymyksiin myös maantieteellisesti aiempaa laajemmalla tasolla. Totutun korkeintaan sadan

neliömetrin kenttätutkimukset ovat kasvaneet alueellisiin, mannerkohtaisiin ja globaalisiin mittoihin. (Michener 2000c, 1)

Ekologit eivät pysty suoriutumaan laajentuneista tutkimusongelmista yksinään (Brown 1994, 23). Etusijalle ekologiassa tulee siis väistämättä nousemaan monen tutkijan suorittamat, monitieteenalaiset tutkimukset ja sen myötä myös datan jakamisen sekä korkealaatuisen ja hyvin ylläpidetyn datan saavutettavuuden merkitys korostuu. Muutos ekologisen tutkimuksen teossa vahvistaa näin ollen datan roolia ja aiheuttaa muutoksia datan hallinnassa (tästä tarkemmin luvussa 5). (Michener 2000c, 3)

Esimerkkeinä uudenlaisesta tavasta tehdä ja organisoida ekologista tutkimusta on yhdysvaltalainen, koko maan kattava tutkimusverkosto NEON (the National Ecological Observatory Network) sekä globaali järvien tutkimusverkosto GLEON (the Global Lake Ecological Observatory Network). NEONissa tutkitaan ilmaston-muutoksen, maankäytön muutosten sekä tulokaslajien aiheuttamia ekologisia vaikutuksia. Tarkoituksena on kerätä ja säilyttää ekologista dataa ainakin 30 vuoden ajan. (Zimmerman & Nardi 2006; NEON) GLEON puolestaan on limnologeista, informaatioteknologian asiantuntijoista ja insinööreistä koostuva verkosto, johon kuului vuonna 2006 12 tutkimuslaitosta eri puolilta maailmaa. Suomesta verkostossa on mukana Helsingin yliopistolle kuuluva Lammin biologinen tutkimusasema. (Kratz et al. 2006)

3.2 Teknologia

Tärkeimpiin teknologisiin edistysaskeleisiin ekologiassa ovat kuuluneet mallien ja prosessien analysointiin ja mallinnukseen saatavilla olevan tietoteknisen tehokkuuden suunnaton lisäys, edistyneemmät keinot mitata ja merkitä muistiin tapahtumia ja havaintoja sekä uudet menetelmät informaation vaihtoon (Jones et al. 2006, 538). Teknologian kehittyminen on mahdollistanut ekologisissa tutkimuksissa kerättävän ja analysoitavan datan määrän merkittävän lisäämisen sekä tutkimusasetelmien ja tutkimuskysymysten laajentamisen. Kehittyneempää teknologiaa tarvitaan myös tämän kasvaneen ja samalla monimuotoistuneen datamäärän hallintaan sekä helpottamaan datan jakamista.

Tutkimusaineistosta esille tulleesta teknologia-termistöstä käsitellään seuraavassa hieman keinotekoisesti erikseen datan keruu- ja analysointivaiheessa käytettävää teknologiaa sekä datan hallintaan laajemmassa mielessä käytettävää teknologiaa, vaikka molemmat informaatioteknologiaa ja osittain päällekkäisiä ratkaisuja ovatkin. Datan keruu- ja analysointiteknologiaan on tässä luettu kuuluvaksi yksittäisen tutkimusprojektin käytössä olevat, datanhallinnan aktiivivaiheeseen liittyvät teknologiset ratkaisut. Datanhallintateknologialla puolestaan tarkoitetaan tässä yhteydessä datan laajemman käytettävyyden, kuten pitkäkestoisen säilyttämisen mahdollistavaa datanhallintateknologiaa.

3.2.1 Datan keruu- ja analysointiteknologia

Datan keruu- ja analysointiteknologiaan liittyen artikkeleissa sivutaan erilaisten välineiden kehittämistyön yksityiskohtaisen kuvailun lisäksi enimmäkseen sitä, minkälaisia ohjelmistoja ekologeilla on käytössä. Keruu- ja analysointivaiheen teknologisen kehityksen luomista mahdollisuuksista ja haasteista ei puhuta paljoakaan. Tässä pyrin kuitenkin nostamaan esiin joitakin ekologisen datan hallintaan yleisesti vaikuttavia teknologisia muutoksia.

Datamäärän lisääntymiseen liittyvä merkittävin teknologinen kehityssuunta on ollut automaattisen- ja kaukokartoitusteknologian tulo ekologiseen tutkimukseen. Tämä on mahdollistanut datan keräämisen aiempaa tiheämmin väliajoin ja laajemmilta alueilta. Erilaiset perusmittaukset, kuten lämpötila, voidaan ohjelmoida tapahtuvaksi tarvittaessa jopa sekuntien tai minuuttien välein. Lisäksi dataa on tällä tavoin mahdollista kerätä turvallisesti myös vaarallisemmista paikoista. Erilaisista automaattisista datankeruuteknologioista on ekologiassa myös se hyöty, että eläinlajeja havainnoivat tutkijat voivat jäädä kauemmaksi tutkimuskohteistaan, mikä vähentää ihmisen läsnäolon vaikutusta eläinten luontaiseen käyttäytymiseen (Porter et al. 2005; Michener 2000d, 143).

Toinen varsin uusi teknologinen vaikutus ekologisen datan keräämiseen on elektronisten kenttäoppaiden yleistyminen. Niiden avulla pyritään parantamaan toisinaan varsin haastavaa lajitunnistusta erityisesti ekologiassa usein hyödynnettyjen amatöörivoimien osalta, ja näin ollen parantamaan datan laatua ja tutkimusten luotettavuutta. Elektronisiin oppaisiin liittyy monia parannuksia perinteisiin

painettuihin oppaisiin verrattuna. Esimerkiksi Yu et al.:n kehittämä EcoPod on kämmenmikrossa toimiva kenttäopas, joka vaatii käyttäjältään mahdollisimman vähän informaatiota ja on muutenkin kirjamuotoista opasta joustavampi ja dynaamisempi käyttää. Se muun muassa ottaa havainnointikontekstin (aika ja paikka) huomioon tarjoamalla vain vuodenaikaan ja kyseiseen maantieteelliseen alueeseen sopivia eliöiden värivaihtoehtoja. (Yu et al. 2006)

Analyysiteknologiaan liittyen ekoinformatiikassa on käynnissä monia projekteja, joissa pyritään parantamaan ekologisen datan analysointia mallintamalla datan kulku koko analysointiprosessin läpi. Nämä tieteelliset työnkulkujärjestelmät (workflow systems) tukevat tavallisesti useita analyyttisiä puitteita ja komponentteja ja niitä on käytetty menestyksekkäästi monilla eri aloilla, muun muassa ekologiassa, joissa datan saavutettavuus, mallintaminen ja visualisointi ovat kompleksisia ja monivaiheisia. Tieteellisiin työnkulkujärjestelmiin liittyy monia etuja. Ensinnäkin ne tuottavat formaalin kuvauksen analyysiprosessissa suoritetuista vaiheista. Toiseksi ne tarjoavat usein suoran pääsyn datalähteisiin sekä monia välineitä datanhallintaan. Kolmanneksi tieteellisissä työnkulkujärjestelmissä on tavallisesti korkealuokkaiset graafiset käyttöliittymät analyyttisten prosessien laadintaan. Tieteelliset työnkulkujärjestelmät voidaan myös mieltää eräänlaiseksi dokumentaatiomuodoksi, joka on helposti arkistoitavissa ja jaettavissa kollegojen kesken. (Jones et al. 2006, 535-536)

Tunnetuin esimerkki ekologian alalle kehitetyistä työnkulkujärjestelmistä lienee tutkimusaineistossakin usein viitattu Kepler (http://kepler-project.org). Keplerin avulla tutkijat voivat tallentaa työnkulkuja helposti vaihdettavissa, arkistoitavissa, versioitavissa ja toteutettavissa olevassa muodossa (Altintas et al. 2004). Kepler myös

tarjoaa suoran pääsyn satojen tutkimusalueiden kenttädataan, eri luonnonmuseoiden ylläpitämään kokoelmadataan sekä GenBank-palvelun sisältämään molekyylibiologia-dataan (Jones et al. 2006, 535).

3.2.2. Datanhallintateknologia

Datanhallintateknologioiksi on tässä yhteydessä mielletty sellaiset artikkeleissa paljon käsitellyt ratkaisut kuin tietokannat, informaatiojärjestelmät sekä laajemmat teknologiset infrastruktuurit. Myös tietokannanhallintajärjestelmät (DBMS) mainitaan tutkimusaineistossa usein. Informaatiojärjestelmillä tarkoitetaan tässä tutkielmassa lähinnä erilaisia projektikohtaisia ratkaisuja datanhallintaan. Kyberinfrastruktuurit taasen ovat laajemmissa tutkimusverkostoissa käytettäviä systeemejä.

Perinteiset lähestymistavat ekologisen datan hallintaan ovat yleensä tuottaneet projektikohtaisia ohjelmistoratkaisuja, jotka ovat olleet käyttökelpoisia vain rajallisessa, tietyn tutkimusprojektin kontekstissa (Jones 2007, 193). Tässä työssä ei tästä syystä olekaan kiinnitetty erityisemmin huomiota ekologian alalla käytettyihin erilaisiin ohjelmistoihin, vaan edellä mainitunkaltaisiin laajempiin teknologisiin teemoihin. Ohjelmistotasolla huomioitava asia on lähinnä se, että olemassa olevia ratkaisuja on paljon ja niiden yhteensovittaminen voi olla ongelmallista.

3.2.2.1 Tietokannat

Tieteellinen tietokanta on Porterin (2000) esittämän määritelmän mukaan tietokoneella ylläpidetty kokoelma toisiinsa liittyvää dataa, organisoituna siten, että se on tieteellisen

tutkimuksen saavutettavissa ja pitkäkestoisesti hoidettu. Tieteelliset tietokannat mahdollistavat erilaisen data-aineksen integroinnin ja datan uudet käyttötavat, usein yli tieteenalarajojen. Tieteellisten tietokantojen kehittämiseen ja käyttöön liittyy useita etuja. Ensinnäkin tietokannat aikaansaavat datan yleislaadun parantumisen, sillä useammat käyttäjät tarkoittavat myös useampia mahdollisuuksia havaita ja korjata datassa olevia virheitä. Toinen etu liittyy kustannuksiin. Datan tallettaminen maksaa nimittäin yleensä vähemmän kuin kerääminen uudelleen. Ekologisen datan kohdalla uudelleenkerääminen ei usein ole käytännössä edes mahdollista, johtuen kompleksiselle luonnolle ominaisista huonosti kontrolloitavista tekijöistä, kuten säästä, jotka vaikuttavat tutkittaviin prosesseihin. (Porter 2000, 48)

Ensisijaisena syynä tieteellisten tietokantojen kehittämiselle tulisi olla niiden mahdollistamat uudenlaiset tieteelliset tutkimukset. Erityisiin tietokantoja tarvitseviin ekologiassa yleisiin tutkimustyyppeihin kuuluvat kohdassa 3.1.1 mainitut pitkäkestoiset tutkimukset, jotka turvaavat tietokantoihin projektihistorian säilyttämisessä; synteesitutkimukset, joissa usein yhdistetään dataa muussa tarkoituksessa kuin mihin data on alun perin kerätty sekä integroidut monitieteenalaiset projektit, jotka tarvitsevat tietokantoja helpottamaan datan jakamista. (Porter 2000, 48)

Suurin haaste hyödyllisten tieteellisten tietokantojen kehittämisessä on datan moninaisuuden kanssa toimeen tuleminen. Moninaisuuden haaste ulottuu myös tietokannan käyttäjiin siinä mielessä, että tietokannan tulee pystyä tukemaan eri taustoista lähtöisin olevien käyttäjien erilaisia tavoitteita. Tämän päivän tieteellisillä tietokannoilla tulisi olla pitkän tähtäimen tavoitteita, mikä on vierasta monille tietokantatyypeille. Kirjallisuudessa usein mainittu tavoite ekologiselle tietokannalle

on, että tietokantaan talletettu data olisi saavutettavissa ja tulkittavissa 20 vuoden kuluttua tallettamisesta. (Porter 2000, 49-51)

Eräs ongelma tietokantojen käytössä ekologiassa on siinä, että harvoilla ekologeilla on tarvittavaa asiantuntemusta tietokantateknologian käyttämiseen, eikä yksittäisillä ekologeilla tai pienillä tutkimusryhmillä ole varaa palkata ohjelmoijia. Useimmat ekologit säilyttävätkin tutkimusdataansa taulukkolaskentaohjelmissa (Cushing et al. 2007, 7-8). Taulukkolaskentaohjelmien hyvinä puolina voidaan pitää niiden helppoa saatavuutta sekä sitä, että niissä pystyy suorittamaan myös jonkin verran erilaisia muokkaus- ja analysointitoimintoja (Brunt 2000, 34). Lisäksi ne ovat varsin joustavia ja helppokäyttöisiä (Jones et al. 2006, 522). Taulukkolaskentaohjelmat eivät kuitenkaan välttämättä ylläpidä tiedoston sisäistä johdonmukaisuutta, sillä jokaista saraketta voidaan muokata muista rivin sarakkeista irrallaan (Brunt 2000, 34). Ohjelmien joustavuus kostautuu myös vaikeutena kehittää automaattisia datan käsittelytapoja, kun datan voi tiedostoissa järjestää niin monella tapaa (Jones et al. 2006, 522).

Hyvä esimerkki tietokantateknologisesta kehitystyöstä ekologian alalla on puiden latvustotutkimuksen tarpeisiin keskittynyt Canopy Database Project (CDP). Projektissa on suunniteltu tietokantaprototyyppi (Canopy DataBank), jonka avulla ekologit itse pystyvät toimimaan omina tietokantaohjelmoijinaan. Keskeisenä ideana on käyttää alakohtaisia tietokantakomponentteja, tietynlaisia mallineita (templates), jotka kuvaavat jonkin fyysisen objektin (kuten puun tai oksan) mittaamisesta saatua dataa. Lähtökohtana on käsitys siitä, että ekologiset tutkimukset sisältävät usein havaintojen tekemistä rakenteellisista elementeistä, jotka eivät yleensä muutu ajan kuluessa tai eroa

eri tutkimuksissa ja voivat näin ollen toimia yhtymäkohtina eri tutkimusten välillä. Tällaiset tietokannat helpottaisivat tiedostojen vertailua aiemmin käytettyjen taulukkolaskentaohjelmien sekalaiseen järjestykseen nähden sekä datan visualisointia. (Cushing et al. 2007)

Paikkatietojärjestelmät (Geographic Information Systems, GIS) puolestaan edustavat ekologiankin alalla paljon käytettyä erityislaatuista tietokannanhallintajärjestelmää (DBMS). Paikkatietojärjestelmissä on tietokantatoimintoihin yhdistetty spatiaalinen kartoitus ja analyyttiset valmiudet. (Brunt 2000, 34)

3.2.2.2 Informaatiojärjestelmät

Ekologisten informaatiojärjestelmien tarkoituksena on tukea tieteellistä tutkimusta, mutta niiden suunnittelua, käyttöönottoa ja toimintaa ei vielä ymmärretä riittävän hyvin (Strebel et al. 1994, 59). Artikkeleissa puhutaan sekä informaatiojärjestelmistä (IS) että informaationhallintajärjestelmistä (IMS). Informaationhallintajärjestelmien yleisenä tavoitteena on jakaa informaatiota käyttäjien ja tuottajien kesken. Informaation-hallintajärjestelmät muodostavat infrastruktuurin, jonka tarkoituksena on palvella tietyn tutkimuspaikan tiedeyhteisön yleistä etua tarjoamalla välineitä synteeseihin ja tutkimusalueiden välisiin toimintoihin (Mélendez-Colon & Baker 2002).

Strebel et al. (1994, 59) ovat kehittäneet käsitteellisen viitekehyksen tieteellisten informaatiojärjestelmien suunnittelulle, käyttöönotolle ja toiminnalle. Viitekehys perustuu fokusoiduissa kenttäkokeissa, pitkäkestoisessa data-arkistoinnissa ja datan julkaisemisessa saatuihin kokemuksiin. Viitekehys koostuu hallinnallisista ja

organisatorisista rajoitteista, tiedeyhteisön vaatimuksista, datan kulusta alkulähteestä arkistoon sekä resurssivaatimuksista. Erotuksena yritysmaailman informaatio-järjestelmiin, tieteellisten informaatiojärjestelmien tulee kyetä käsittelemään monipuolisempaa data-ainesta ja sopeutumaan monenlaisiin, esimerkiksi mittaus-tavoissa ja datatiedostojen välisissä suhteissa yhtä hyvin kuin tutkimushankkeessa laajemminkin tapahtuviin muutoksiin. (Stebel et al. 1994, 59)

Yleisesti ottaen tieteellisiltä informaatiojärjestelmiltä vaaditaan joustavuutta ja tasapainottelua tutkijakohtaisten ja yleisempien käyttäjätarpeiden, lyhyentähtäimen ja pitkäkestoisemman datan käsittelyn sekä paikallisten ja yleisempien suunnittelu-menetelmien välillä. Joustavuuden ja tasapainon puuttuminen voivat aiheuttaa datatiedostojen ja informaatiojärjestelmien jäämisen käyttämättömiksi (Baker et al. 2000, 964-965).

Esimerkkinä ekologian alalla olevista informaatiojärjestelmistä voisi mainita Hollannissa kehitetyn SynBioSys-järjestelmän, josta on tehty sekä Hollannin käyttöön että koko Euroopan laajuiseen käyttöön tarkoitetut versiot (SynBioSys NL ja SynBioSys Europe). Molemmat järjestelmäversiot toimivat verkkopalvelimen kautta toisiinsa liitettyjen hajanaisten tietokantojen verkostona ja niissä yhdistellään monen tasoista kasvillisuusinformaatiota. (Schaminée et al. 2007, 464)

Pitkän aikavälin perspektiivi ja siihen liittyvät haasteet eivät ole tähän mennessä juurikaan saaneet huomiota informaatiojärjestelmätutkijoiden keskuudessa (Karasti 2007, 1). Ekoinformatiikassa pitkän aikavälin perspektiivi on kuitenkin merkittävä ongelma ekologisten tutkimusprojektien keston venyessä ja teknologian jatkaessa

kehittymistään kiihtyvällä tahdilla. Toinen merkittävä teknologinen haaste on eri sovellusten yhteensovittaminen, esimerkiksi jouduttaessa siirtämään dataa järjestelmästä toiseen. Kummassakin tilanteessa on ensiarvoisen tärkeää huolehtia datan säilymisestä ymmärrettävänä ja laadukkaana. Monikielisissä yhteisöissä lisäongelmia aiheuttavat myös erikielisten dokumenttien hallinnan tarve (Lin et al. 2006).

3.2.2.3 Kyberinfrastruktuurit

Kyberinfrastruktuurit ovat edistyneitä informaatioteknologioita, jotka tekevät jaetuista resursseista, kuten tietokonelaitteista ja -palveluista, välineistä, datasta ja ihmisistä helpommin saavutettavia ja tukevat näin tieteellisten löydösten tekemistä. Tällä hetkellä kehitteillä olevat kyberinfrastruktuurit on tarkoitettu mahdollistamaan erilaisten käyttäjien pitkäaikainen yhteistyö. Suurin haaste tässä kehitystyössä on erilaisten monilla eri tutkimuspaikoilla sijaitsevien käyttäjien tukeminen nopeasti muuttuvissa olosuhteissa. (Zimmerman & Nardi 2006, 1601-1602). Ekologisiin tutkimusverkostoihin liittyvää infrastruktuuritutkimusta ovat tutkimusaineiston perusteella tähän mennessä tehneet etenkin Baker, Bowker ja Karasti (esim. 2002), mutta terminä kyberinfrastruktuuri näkyy jossain määrin myös muussa aineistossa.

Haasteena erilaisten teknologisten ratkaisujen kehittämisessä ylipäätään on muun muassa tutkijoiden haluttomuus oppia käyttämään vieraita välineitä datan hallintaansa (Jones et al. 2006, 536). Zimmerman (2007, 5-6) huomauttaakin, että mikäli ekologien nykyisiä työskentelytapoja, heidän tarpeitaan, järjestelmien käytettävyyttä ja ekologiyhteisön sosiaalisia aspekteja ei huomioida, saattaa uudesta teknologiasta

koitua vain vähän käytännön hyötyä ja esimerkiksi kyberinfrastruktuureihin tehdyt valtavat sijoitukset mennä hukkaan. Teknologiaa tuleekin aina viime kädessä arvioida suhteessa sen ekologiselle tutkimukselle tuottamaan arvoon (Karasti & Baker 2004, 7).

3.3 Informaatio

Informaationäkökulmasta ekologisessa tutkimuksessa on kyse luonnon objektien muuttamisesta tieteellisen tietämyksen objekteiksi. Toisin sanoen kohdeilmiö täytyy digitoida numeroiksi ja biteiksi siten, että muu tiedemaailma pystyy niiden perusteella ymmärtämään ekologien esittämiä väitteitä ilmiöistä ja eliöistä, joita he eivät välttämättä ole koskaan konkreettisesti kohdanneet. (Roth & Bowen 1999, 721) Tästä syystä ekologiaa voidaan luonnehtia hyvin dataintensiiviseksi alaksi.

Datan merkityksellisyys tuli ilmi myös artikkeleiden sisällönerittelyn myötä, valtaosan informaatio-termistöstä viitatessa nimenomaan dataan, sen käsittelyyn tai erilaisiin datatuotteisiin. Edellä käsitellyt ekologisen tutkimuksen laajentuminen ja verkostoituminen sekä teknologisen kehityksen aikaansaama ekologisen datan määrän lisääntyminen ja monimuotoistuminen ovat nostaneet datan asemaa ekologiassa entisestään.

Tässä luvussa käydään läpi datan, informaation ja tietämyksen perusmääritelmät ja käsitys data-informaatio-tietämys-jatkumosta ja sen soveltuvuudesta ekologiaan ja ekoinformatiikkaan. Lopuksi tarkastellaan ekologisen datan erityispiirteitä.

3.3.1 Data-informaatio-tietämys

Tutkimusaineiston sisältämä informaatio-kategoriaan luokiteltavissa oleva termistö on jaettavissa dataan, informaatioon ja tietämykseen liittyviin ilmauksiin. Datan voidaan määritellä koostuvan kokonaan merkeistä ja numeroista, joilla on vain vähän tai ei lainkaan sisäistä merkitystä. Informaatio puolestaan on korkeamman tason esitys datasta, eli datalle on annettu muoto tai olemus ja sovittu merkitys (Michener 2000a, 163). Datasta tulee siis informaatiota mikäli datalle on tunnistettavissa tietty käyttötarkoitus ja kun sille muodostetaan sellainen rakenne, että se on mahdollisimman helposti käytettävissä (Blair 2002, 1019). Tietämys taasen on informaation tutkimisesta, käsittämisestä ja sisäistämisestä muodostuvaa ymmärrystä (Michener 2000a, 163), jota tarvitaan informaation saamiseksi datasta (Blair 2002, 1021).

Datan, informaation ja tietämyksen lisäksi oleellinen tiedonhallinnan käsite tässä työssä on metadata. Metadata voidaan määritellä datan ymmärtämiseen ja käyttämiseen tarvittavaksi kontekstuaaliseksi informaatioksi, tai lyhyemmin sanottuna dataa kuvailevaksi dataksi (Jones et al. 2006, 524). Michener et al. (1997, 331) ovat määritelleet kaavan, jonka mukaan yhdistettäessä ekologiseen raakadataan metadataa saadaan informaatiota, tietyn käsitteellisen viitekehyksen puitteissa. Raakadatalla tarkoitetaan suoraan laboratoriosta tai kentältä talletettua, käsittelemätöntä dataa (Baker et al. 2000, 966). Ekologista informaatiota voi näin ollen hävitä yhtä hyvin raakadatan kuin metadatan turmeltumisen myötä (Michener et al. 1997, 331). Michener on toisaalla täsmentänyt, että metadatan liittämisen lisäksi ekologinen raakadata vaatii yleensä myös muuta käsittelyä ennen kuin siitä saadaan informaatiota (Michener 2000d, 143).

Tieteellinen data käsittää NRC:n (National Research Council) määritelmän mukaisesti erilaisia tieteellisiä tai teknisiä mittauksia, näistä laskettuja arvoja sekä havaintoja ja faktoja, jotka voidaan esittää numeroina, taulukoina, graafisina esityksinä, malleina, tekstinä tai symboleina, ja joita käytetään päättelyn perustana tai laskennassa (NRC 1997, 198). Tieteelliselle datalle on ominaista suuri vaihtelevuus volyymin ja kompleksisuuden suhteen. On suurivolyymistä suhteellisen homogeenista dataa (esim. satelliittikuvat), pienivolyymistä erittäin kompleksista dataa, kuten taulukkomuotoinen monia analyyseja kuvaava ekologinen data, joka vaatii paljon metadataa. Lisäksi on dataa, joka on sekä suurivolyymistä että kompleksista, kuten paikkatietojärjestelmien (GIS) datakerrokset. (Porter 2000, 49) Näistä tavallisin primaaridatan muoto ekologian alalla on taulukkodata (Brunt et al. 2002, 2).

3.3.2 Data-informaatio-tietämys ekologiassa

Ekologista tietämystä ei useinkaan synny yksittäisen tutkimuksen tuloksista vaan tietämys kasvaa ja kehittyy etsimällä ja tunnistamalla yleisiä malleja, jotka tulevat usein näkyviksi vasta lukuisten tutkimustulosten tarkastelun jälkeen. Tietämyksen aikaansaaminen edellyttää siten ekologisen datan hallintaa ja käsittelyä suuressa mittakaavassa. (Michener 2000a, 163)

Ekologiaan voisi hyvin sopia Tuomen (1999) esittämä perinteiselle data-informaatio-tietämys -jatkumolle käänteinen hierarkia, jonka mukaan dataa ei olisi ilman informaatiota, jota puolestaan syntyy vain jos meillä on tietämystä. Toisin sanoen esimerkiksi ekologista raakadataa ei ole sellaisenaan luonnossa valmiina olemassa, vaan se tulee osata havainnoida tai mitata erilaisen ohjeistuksen ja alalla

vallitsevan tietämyksen avulla. Perinteinen näkemys siitä, että data on vain informaation ja tietämyksen raaka-ainetta, saa datan vaikuttamaan informaatiota ja tietämystä arvottomammalta, ja siitä huolehtimisen informaatioksi jalostamisen jälkeen toissijaiselta ja epäkiinnostavalta asialta. Puutteellinen tai huolimaton datanhallinta informaatioksi rikastamisen jälkeen puolestaan haittaa uuden monista tutkimus-tuloksista muodostettavan ekologisen tietämyksen syntymistä.

Ekologinen data, informaatio ja tietämys kuuluvat kaikki osaltaan ekoinformatiikan piiriin. Ekologista informaatiota tai tietämystä ja sen hallintaa ei kuitenkaan juurikaan käsitellä tämän tutkimuksen aineistoon kuuluvissa tutkimusartikkeleissa vaan valtaosa informaatio-kategoriaan kuuluvasta termistöstä on nimenomaan dataan liittyvää. Informaatiota ei ekologiassa nähtävästi koeta yhtä hankalana hallita kuin dataa, eikä informaationhallinta näin ollen vaikuta kovin houkuttelevalta tutkimuskohteelta ekoinformatiikalle. Tietämyksenhallinnasta puolestaan ei liene vielä ehtinyt muodostua kovin merkittävä tutkimusalue ekoinformatiikassa, mutta mielenkiinto tietämystä kohtaan datanhallinnan rinnalla on jo herännyt. Esimerkiksi Saksassa ollaan viime vuosina kehitetty tietämyksenhallintajärjestelmää ekologian alalle (ILMAX) (Neumann et al. 2003). Lisäksi tietämyksen olemassaolo ekologisen datan aikaansaamisessa ja toisaalta jonkun toisen keräämän datan tulkitsemisessa on havaittu oleelliseksi tekijäksi (esim. Zimmerman 2003).

3.3.3 Ekologinen data

Tässä alaluvussa selvennetään hieman sitä, mistä on kyse kun puhutaan ekologisesta datasta, sekä tuodaan esiin tutkimusaineistosta esiin nousseita näkemyksiä ekologisen

datan erityispiirteistä datanhallinnan kannalta. Ekologisella datalla voidaan tarkoittaa joko ekologisissa tutkimuksissa kerättyä ja tuotettua dataa tai ekologisissa tutkimuksissa tarvittavaa dataa, joka voi olla osittain peräisin monilta muilta tieteenaloilta. Erään määritelmän mukaan (Michener 2006, 3) ekologisissa tutkimuksissa tarvittava data kattaa biologian, kemian, fysiikan ja yhteiskuntatieteet sekä monet niiden alatieteistä. Tässä tutkielmassa ei yleisesti ottaen rajata määritelmää tiukasti vain tutkimuksissa tuotettavaksi dataksi, koska toisaalta artikkeleista ei aina käy yksiselitteisesti ilmi kummassa mielessä datasta puhutaan ja toisaalta ekologisessa tutkimuksessa syntyvissä datatiedostoissa on yleensä yhdistettynä monenlaista dataa, jolloin nimenomaan kaikkea ekologiassa tarvittavaa dataa voidaan pitää ekoinformatiikan tutkimuksen kohteena. Toisinaan artikkeleissa puhutaan ekologisen datan rinnalla esimerkiksi biologisesta datasta, ympäristödatasta tai biodiversiteetti-datasta, mikä kuvastanee osittain tutkimuksissa tarvittavan datan kirjoa, mutta ennen kaikkea sitä, ettei ekoinformatiikka ole määritelmällisesti vielä vakiintunut eikä selvärajainen.

Ekologista dataa luonnehditaan useimmiten heterogeeniseksi. Data voi olla monessa eri muodossa (mm. tekstinä, numeroina tai kuvina), monella tapaa loogisesti organisoituna ja monenlaisilla näytteenottomenetelmillä aikaansaatua (Jones et al. 2006, 519). Ekologisessa datassa näkyvät erilaiset syyt datan keräämiselle, erilaiset havainnoidut muuttujat ja erilaiset ajalliset ja alueelliset näytteenottoasetelmat (Fegraus et al. 2005, 158-159). Tämän lisäksi ekologinen data on sisältönsä puolesta monipuolista, sillä se voi periaatteessa koskea mitä tahansa geeneistä biosfääriin mukaan lukien erilaisiin prosesseihin liittyvät seikat, kuten arviot kasvillisuuden lehtivahingoista. (Jones et al. 2006, 519/521).

Ekologista dataa kuvaillaan usein myös kompleksiseksi ja hajanaiseksi. Kompleksisuudella voidaan tarkoittaa puuttuvien arvojen, kesken kaiken muuttuvien näytteenotto- ja laboratoriomenettelyjen, tutkimusparametrien lisäilyjen ja poistojen, henkilöstövaihdosten, muuttuneiden ympäristöolosuhteiden sekä monien muiden seikkojen aiheuttamia poikkeamia datatiedostoissa (Michener et al. 1997, 332). Hajanaisuudella puolestaan voidaan toisaalta niin ikään viitata suuresti vaihteleviin ajallisiin ja alueellisiin mittasuhteisiin ja toisaalta taas siihen, että ekologinen data on sijoitettu tavallisesti erillisiin pieniin projektikohtaisiin datatiedostoihin, joita ei useinkaan ole koottu keskitetysti mihinkään tietokantaan.

Edellä esitellyt ekologisen datan kuvaukset ovat vain eräitä mahdollisia tulkintoja, sillä artikkeleista ei aina käy selkeästi ilmi, mitä luonnehdinnat heterogeenisyydestä, monipuolisuudesta, kompleksisuudesta ja hajanaisuudesta tarkalleen ottaen tarkoittavat ja kuinka ne eroavat toisistaan. Kaiken kaikkiaan tilanne on ekologisen informaation kannalta kuitenkin se, että tietämystä luonnonympäristöstämme ei rajoita vain luonnon ilmiöiden ja prosessien kompleksisuus vaan myös niitä kuvailevan datan kompleksisuus (Michener et al. 2007, 112).

Ekologisesta datasta annetut konkreettisemmat esimerkit ovat hyvin moninaisia. Tiivistäen voidaan sanoa, että ekologinen data voi kuvata eri eliölajien esiintymistä tai esiintymättömyyttä tietyllä alueella, eliöiden ja niiden esiintymisalueen ominaisuuksia (fyysisiä, fysiologisia, käyttäytymistä) sekä kuvauksia monenlaisista vuorovaikutus-suhteista (kuten sijoittumisesta ravintoverkostoihin). Käytännössä niin sanottu raakadata on tulosta monenlaisesta mittaus- tai havainnointitoiminnasta, johon liittyy paljon epävarmuutta (ks. 5.1).

Vaikka kaikella ekologisella datalla voidaan ajatella olevan tilallinen ulottuvuutensa, on ekologinen data yleensä luokiteltavissa joko geospatiaaliseksi tai ei-geospatiaaliseksi. Geospatiaalinen data kiinnittyy suoraan johonkin maantieteelliseen paikkaan, kun taas ei-geospatiaalinen ekologinen data saattaa olla peräisin esimerkiksi erilaisista laboratoriokokeista. (Michener 1998, 47)

Yksi ekologisen datan merkittävä ominaispiirre datanhallinnan kannalta on sen kuvaamien luonnonilmiöiden yksiselitteisen nimeämisen ja luokittelemisen vaikeus. Esimerkiksi yksinomaan eliölajien taksonomiset nimet ja luokitukset ovat epästabiileja. (Bowker 2000) Lajinimistön käsitteellinen tulkinta muuttuu systemaatikkojen joutuessa aika ajoin tarkistamaan organismien luokituksia uuden datan valossa. Tästä seuraa se, että sama kaksiosainen tieteellinen nimi saatetaan eri taksonomistien mukaan yhdistää moniin eri lajeihin, mikä luonnollisesti aiheuttaa monitulkintaisuutta ekologisten havaintodatatiedostojen kohdalla. (Jones 2007) Lisäksi luonnonympäristöön kuuluu asioita, joita on vaikeata nimetä. Tällaisia ovat muun muassa kokonaisuudet, jotka eivät ole selvärajaisia, kuten maaperät tai maisematyypit. Luokiteltavien kohteiden epämääräisyys ja nimeämisessä ja luokituksissa jatkuvasti tapahtuvat muutokset ovat varsin ongelmallisia datanhallinnan kannalta. Nämä seikat ovat muun muassa johtaneet monien eri luokitusjärjestelmien laatimiseen eri maissa ja myös eri organisaatioissa saman maan sisälläkin, mikä puolestaan tekee eri järjestelmillä luokitellun datan yhdistelemisestä ja tulkinnasta hankalaa. (Bowker 2000, 652-653)

Tärkeä teema ekologiseen dataan liittyen on arvostus, sen osittainen puuttuminen ja toisaalta datan arvon ymmärtämisen oleellisuus ekoinformatiikan pyrkimysten

mukaisten muutosten aikaansaamiseksi ekologisen datan hallinnassa. Tieteellisestä näkökulmasta tarkasteltuna datan todellinen arvo liittyy suoraan kykyymme aikaansaada datasta korkeamman tason tietämystä, eli datassa aluillaan olevaan informaatiosisältöön (Michener 2000a, 162). Ekologiselle datalle on ominaista, että tämän informaatiosisällön hyödyllisyys säilyy aikojen kuluessa, eli toisin kuin monilla muilla aloilla datan arvo kasvaa vanhetessaan, mikäli datasta vain pidetään asianmukaisesti huolta.

Perinteisessä ekologisessa tutkimusprojektissa data on kuitenkin tavallisesti nähty vain keinona projektin loppuunsaattamiseksi, eli julkaisun aikaansaamiseksi ja valitettavan usein on data julkaisun valmistuttua heitetty menemään tai hylätty arkistolaatikoston pohjalle tai vanhentuneeseen tiedostomuotoon (Michener 2000c, 3). Dataa itsessään ei siis olla ekologiassa niinkään aiemmin arvostettu, eikä sen säilyttämistä olla yleisesti ottaen nähty tarpeellisena tutkimuksen päätyttyä. Ekologisen havaintodatan säilyttämistä tulisi kuitenkin pitää ensiarvoisen tärkeänä, koska tällainen data on aina tallenne tapahtumasta, joka ei tule toistumaan, ja näin ollen hävitessään korvaamaton (Zimmerman 2003, 4).

3.4 Ekologia + teknologia + informaatio = ekoinformatiikka

Edellä käsitellyt kolme ulottuvuutta muodostavat lähtökohdan ekoinformatiikka-tutkimukselle, jossa ekologia toimii kasvualustana, rakennusaineina on teknologia ja varsinaisena kohteena informaatio. Ulottuvuudet voidaan tulkita myös ekoinformatiikassa välttämättä mukana tarvittaviksi tutkimusaloiksi, eli ekologiaksi,

tietojenkäsittelytieteiksi ja informaatiotieteiksi. Luvussa 2 esiin nostettu sosiaalitiede voidaan sijoittaa näiden välimaastoon, ikään kuin sitomaan muita tieteitä yhteen.

Ekoinformatiikan määritelmän tavoin tutkimusaineistoon kuuluvista artikkeleista ei tyypillisesti myöskään ole löydettävissä selkeää yksittäistä yleistä päämäärää tai tavoitetta näistä aineksista muotoutuvalle ekoinformatiikalle. Artikkeleissa annetut tavoitteet liittyvät yleensä joihinkin yksittäisiin osa-alueisiin. Samoin kuin määritelmien osalta myös tavoitteita kartoitettiin ainoastaan sellaisista artikkeleista, joissa ekoinformatiikasta käytettiin kirjoitusasua eco(-)informatics tai ecological informatics ja tämän lisäksi puhuttiin nimenomaan ekologisesta datasta.

Sim et al.:n (2004) näkemyksessä korostetaan tämän tutkielman tavoin ekologista dataa ja jätetään tutkimusalat ja keinot tavallaan avoimeksi. Heidän mukaansa ekoinformatiikan tavoitteena on suunnitella välineitä datan jakamis-, hallinta- ja integrointitoimien tukemiseen sekä tekemään mahdolliseksi tutkijoiden käyttää yhtä tarkoitusta varten kerättyä dataa uuden ongelman tutkimisessa. McCartney ja Jones (2002, 379) ovat samoilla linjoilla todeten ekoinformatiikan tavoitteena olevan ekologisen datan pitkäaikaisen saatavuuden takaaminen sekä datan käytettävyyden parantaminen tietämyksen saavuttamiseksi ympäristöstämme.

Kinemanin ja Kumarin (2006, 367) hieman edellisistä poikkeavan näkemyksen mukaan ekoinformatiikan päämääränä voidaan pitää luonnon monimuotoisuuden ja ekosysteemi-ilmiöiden kuvaamista sekä tällaisen informaation välittämistä yhteiskunnalle. He siis korostavat yleisemmin painotettujen datanhallinnallisten seikkojen sijaan viestinnällistä puolta ekoinformatiikassa.

Kaiken kaikkiaan tutkimusartikkeleista on yleisesti nähtävissä, että ekologisessa tutkimuksessa käynnissä olevan ’lokaalista globaaliksi’ -suuntauksen tavoin myös ekoinformatiikassa tavoitellaan kokonaisvaltaisen datanhallinnan kehittämistä laajemman ekologisen tutkimusyhteisön eikä niinkään yksittäisen tutkijan tai tutkimusryhmän näkökulmasta. Käytännössä tutkimus- ja kehitystyötä tehdään pienemmissä konteksteissa, mutta kehitystyön perustaksi pyritään löytämään yleisesti ekologiseen tutkimukseen kuuluvia piirteitä.

Ekoinformatiikan tutkimus- ja kehitystyön tavoitteiden saavuttamiseen liittyy valtava määrä haasteita. Jokaisessa artikkelissa mainitaan monia erilaisia haasteita, ongelmia tai esteitä ekoinformatiivisten ratkaisujen kehittämisen tiellä. Jones et al. (2006, 520) julistavat kaikkein merkittävimmäksi haasteeksi ekoinformatiikassa ekologisissa tutkimuksissa tarvittavalle datalle luontaisen kompleksisuuden ja laajuuden kanssa pärjäämisen. Haaste on hyvin perustavanlaatuinen, kaikkeen tutkimustoimintaan liittyvä, onhan juuri ekologinen data ekoinformatiikan ydin.

Muut artikkeleissa esitetyt haasteet liittyvät selkeämmin ekoinformatiikan osa-alueisiin tai muutoin suppeampaan kontekstiin, ja niitä käsitellään tarkemmin luvussa 5 omissa yhteyksissään. Kiteytetysti voidaan sanoa, että melkoisia haasteita liittyy vielä jokaiseen ekoinformatiikan osa-alueeseen, katsottiinpa asioita sitten ekologian, teknologian tai informaation näkökulmasta. Sosiaalisessa mielessä keskeisimpiä haasteita lienee Bakerin, Bowkerin ja Karastin (2002, 3) esille tuoma mielekkään pitkäkestoisen datan luomisen ja säilyttämisen tavoitteen siirtäminen ja juurruttaminen tutkijoiden päivittäiseen työhön.

4 EKOINFORMATIIKKA TUTKIMUSALANA

Tässä luvussa hahmotellaan ekoinformatiikkatutkimuksen puitteita eli sitä, ketkä tutkimusta tyypillisesti tekevät ja minkälaisia julkaisu- ja yhteistyöfoorumeita ekoinformatiikkatutkijoilla on käytössään.

4.1 Ekoinformatiikkatutkijat

Ammattinimikkeeltään bioinformaatikkoja olevia henkilöitä on jo olemassa, mutta ekoinformaatikkoja ei liene vielä ainuttakaan, vaikka alan koulutusta ainakin Yhdysvalloissa jossain määrin onkin jo tarjolla. Kuvaa tyypillisistä ekoinformatiikan tutkijoista pyrittiin näin ollen selvittämään niistä tutkimusaineistoon kuuluvista artikkeleista, joissa ekoinformatiikka oli terminä mukana joko asiasanana tai otsikossa. Näiden artikkeleiden kirjoittajien taustaorganisaatioiden perusteella näyttäisi siltä, että ekoinformatiikkatutkimusta todella tehdään määritelmissä (luku 2) esiin tulleen kuvan mukaisesti pääasiassa ekologian ja tietojenkäsittelytieteiden yhteistyönä.

Organisaatioiden joukossa on vain yksi ekoinformatiikan nimeä kantava tutkimuslaitos, nimittäin vuonna 2004 perustettu yhdysvaltalainen Pacific Ecoinformatics and Computational Ecology Lab. Sen henkilöstö koostuu pääasiassa ekologeista, mutta mukana on myös tietojenkäsittelytieteiden edustaja (PEaCE Lab). Huomattavimman poikkeuksen ekoinformatiikkatutkimuksen tähänastisessa kahtia jakautuneisuudessa ekologian ja teknologian edustajiin muodostavat tämän selvityksen perusteella Sim, Zimmerman ja Nardi (2004), jotka edustavat informaatioalaa.

Tutkielman aineiston sisällössäkin kirjasto- ja informaatiotieteet mainitaan satunnaisesti ja tiettyjen teemojen yhteys tunnistetaan. Esimerkiksi Helly et al. (1999, 6) toteavat kirjastotieteen alalla tapahtuneen kehityksen voivan toimia mallina sille, kuinka standardointi, datan jakaminen, luettelointi ja arkistointiyritykset voivat kehittyä ja hyödyttää tieteentekoa yleisemminkin. Myös Kalra (2005, 335) peräänkuuluttaa kirjasto- ja informaatioalan ammattilaisia mukaan bioinformatiikkaan, jonka hän siis mieltää yhden esittämänsä määritelmän mukaan eräänlaiseksi sateenvarjotermiksi eli kaikenlaisen biologisen informaation tietotekniseksi käsittelyksi, käsittäen näin myös ekologisen datan ja informaation. Informaatioalan ammattilaisilla olisi hänen mukaansa paljon annettavaa tälle tutkimusalalle, koska tietämyksen järjestämisen ja tiedonhaun soveltamisesta, kuten luokittelusta, metadatasta ja sanastokontrollista, on tullut tärkeitä valtavien biologisen datan määrien käsittelyssä. Arkistotieteitä tutkimusaineistossa ei erikseen mainita, vaikka arkistoinnista paljon puhutaankin. Data-arkistoinnin ajatellaan ilmeisesti eroavan niin merkittävästi perinteisestä arkistotieteen fokuksesta, ettei mielenkiintoa tieteellisen datan arkistoinnin kehittämiseen uskota arkistoammattilaisissa heräävän (esim. Jones et al. 2006, 530).

Luvussa kolme käsiteltyjä ekoinformatiikan lähtökohtia ajatellen ei riitä, että ekoinformatiikkatutkimusta tehdään vain tietoteknisten alojen ja ekologien yhteistyönä. Tällöin tulevat näkökulmina huomioiduiksi vasta kaksi ekoinformatiikan perustavaa tukijalkaa - ekologia ja teknologia - informaationäkökulman jäädessä toissijaiseksi. Vaikka ekoinformatiikan ensisijaisena tutkimuskohteena voidaankin pitää dataa eikä informaatiota, tulisi myös data mieltää erityislaatuiseksi informaatioresurssiksi, sillä myös ekologiseen dataan kohdistuu muun muassa tarpeita,

hakua, hankintaa ja arkistointia. Näiden ekologisen datanhallinnan uusien osa-alueiden kehittämisessä tulisikin ekologien ja tietojenkäsittelytieteiden edustajien rinnalle saada lisää nimenomaan informaatioalan ammattilaisia. Zimmermanin yksin tai tutkimusryhmänsä kanssa tekemien tutkimusten lisäksi oikeastaan vain laajoissa ekologisissa tutkimusprojekteissa mukana oleva tietohallintohenkilöstö on osaltaan ollut mukana etsimässä parhaita mahdollisia datanhallintatapoja, joskin usein suppeammasta näkökulmasta kuin ekoinformatiikassa yleisesti.

Kirjasto- ja informaatiotieteiden lisäksi myös sosiaalitieteet saavat jossain määrin huomiota tutkimusaineistossa. Esimerkiksi Baker et al. (2002) ovat eri yhteyksissä tuoneet esille, että myös ekosysteemi-informatiikan (eli ekoinformatiikan) sosiaaliset ja organisatoriset ulottuvuudet tulisi ottaa huomioon parempien tietokantojen kehittämiseksi. Sosiaalitieteilijät ovat myös tahollaan tutkineet eräitä ekoinformatiikkaakin kiinnostavia teemoja, kuten datan jakamista (Zimmerman 2003, 5). Tämän tutkielman aineistoonkin kuuluu kaksi osittain ekoinformatiikkaan liittyvää, sosiaalitieteellisessä julkaisussa julkaistua artikkelia (Bowker 2000 ja Roth & Bowen 1999), joten kiinnostusta yhteistyöhön varmasti löytyisi yleisemminkin.

Myös ekologiassa yleistyvien laajojen tutkimusverkostojen toimintaan liittyy paljon sosiaalitieteellisesti lähestyttäviä haasteita. Ekologisiin tutkimusverkostoihin onkin jo jossain määrin kohdistettu sosiaalitieteellistä etnografista tutkimusta ainakin seuraavassa esitellyn LTER-verkoston puitteissa. Nämä tutkimukset (esim. Baker et. al. 2002) ovat osaltaan auttaneet ymmärtämään sosiaalitieteiden merkitystä ekoinformatiikan pyrkimysten saavuttamisessa.

4.2 Tutkimustoiminta

Tähän mennessä suurin osa ekoinformatiikkatutkimuksesta on keskittynyt kehittämään menetelmiä ekologisen tutkimuksen tuottavuuden parantamiseksi sekä lisäämään tutkimusdatan julkista saatavuutta (Cushing & Wilson 2005, 5). Käytännössä suuri osa ekoinformatiikan tämän hetkisestä tutkimustoiminnasta näyttäisi olevan keskittynyt LTER - verkoston toimiston ja NCEASin (National Center for Ecological Analysis and Synthesis) käynnistämiin projekteihin.

Yhdysvaltalainen Long Term Ecological Research (LTER) -ohjelma (www.lternet.edu) aloitettiin vuonna 1980, jolloin siinä oli mukana kuusi tutkimusaluetta. Sittemmin LTER-ohjelma on laajentunut käsittämään peräti 26 tutkimusaluetta Yhdysvalloissa. LTER- verkostoon kuuluu yli 1100 tutkijaa sekä yli 700 opiskelijaa arviolta 140 eri laitokselta (Baker et al. 2000, 975). Tutkimusalueiden lisäksi LTER-ohjelmaan kuuluu verkoston toimintaa ja tutkimusalueiden välistä viestintää koordinoiva toimisto (Michener et al. 2002, 1). LTER-ohjelman keskeisenä tavoitteena on oppia ymmärtämään ekologisten järjestelmien pitkänaikavälin malleja ja prosesseja monilla tilallisilla laajuuksilla (Hobbie et al. 2003). Tämän tekee mahdolliseksi ohjelman kuuden vuoden sykleissä uusittava rahoitusmalli, joka luo vakaan ympäristön yhteistyön edistämiselle ja mahdollistaa erilaiset innovatiiviset kokeilut (Baker & Bowker 2007, 4).

LTER on pisimpään yhtäjaksoisesti jatkunut ekologisen datan keräyshanke (Karasti 2007, 3) ja datanhallinta on ohjelman alusta alkaen ollut vaadittu osa jokaisen verkostoon kuuluvan alueen tutkimusohjelmaa (Baker & Bowker 2007, 4). LTER:n

tietohallintohenkilöstö onkin ollut pioneeriasemassa tutkimuspaikkakohtaisen ekologisen datan hallinnan kehittämisessä (LTER 2005a, 9). Alusta lähtien LTER:ssä on myös korostettu datan pitkäaikaista säilyttämistä (Karasti & Baker 2004, 2) ja pyritään muutenkin rakentamaan hyvin pitkäkestoisia, tutkijan eliniän sijasta ekosysteemin elämän kestäviä lähtökohtia ekologiselle datalle (Baker & Bowker, 2007, 1).

LTER-verkostoon kytkeytyy myös kansainvälinen ILTER (www.ilternet.edu) -ohjelma (International Long Term Ecological Research), joka sai alkunsa vuonna 1993. Vuoden 2006 toukokuuhun mennessä verkostoon on liittynyt 32 jäsentä. Myös Suomi on perustanut oman FinLTSER-verkostonsa (Finnish Long-Term Socio-Ecological Research Network), joka pyrkii osaksi kansainvälistä ILTER-verkostoa (FinLTSER 2007). ILTER:n tarkoituksena on kehittyä maailmanlaajuiseksi ohjelmaksi ja kehittää tarvittava infrastruktuuri kommunikointia ja tietohallintoa helpottamaan. ILTER tarjoaa oivan mahdollisuuden arvioida erilaisia lähestymistapoja tieteidenväliseen tutkimukseen (Hobbie et al.2003, 29).

Vuonna 1995 perustettu NCEAS puolestaan on johtavassa asemassa yhteistyöhankkeiden ja teknisten ratkaisujen kehittämisessä ekologisen datan käyttöön liittyviin ongelmiin. Koska NCEASin toiminta perustuu olemassa olevan datan käyttöön, on se myös ollut tukemassa ekoinformatiikkatutkimusta. NCEAS:n ekoinformatiikkaohjelman projektit liittyvät neljään avainhaasteeseen: ympäristödatan tallennukseen ja hallintaan, datan löytämiseen ja valmisteluun jatkoanalyyseja ja synteesiä varten, edistyneeseen automatisoituun järjestelmäprosessointiin sekä näiden voimavarojen asettamiseen tutkijoiden saataville. NCEAS myös ylläpitää julkista data-arkistoa.

Ekoinformatiikkatutkimusta helpottaakseen NCEAS on vuonna 2003 perustanut Ekoinformatiikkakeskuksen (EcoInformatics Center, EIC), jonka tarkoituksena on parantaa ekologisen informaation saavutettavuutta. (NCEAS 2008)

NCEASin osallistumista projekteihin näkyvimpiin kuuluvat KNB (The Knowledge Network for Biocomplexity) ja SEEK (Science Environment for Ecological Knowledge), joissa molemmissa on mukana muitakin tahoja kuten edellä esitelty LTER. Kohdassa 3.2.1 mainittu Kepler on kehitetty osana SEEK-hanketta ja kuuluu näin ollen myös NCEASin aikaansaannoksiin.

KNB on yhdysvaltalainen tutkimusasemista, laboratorioista ja tutkijoista koostuva verkosto, joka tukee pitkäkestoista ja laaja-alaista ekologisen informaation synteesiä kehittämällä ohjelmistoja helpottamaan ekologisen datan paikantamista, hankintaa ja integrointia (KNB). Keskeisessä asemassa KNB:ssä on metadatastandardi EML (ks. 5.2.3.1), jonka ympärille on kehitetty datan ja metadatan hallintaväline Morpho (ks. 5.2.3.2) sekä Metacat-palvelin, joka tallettaa metadatadokumentteja ja tekee niistä saavutettavia internetin välityksellä (Andelman et al. 2004, 243).

SEEK taasen on hanke, jonka tarkoituksena on luoda kyberinfrastruktuuri ekologian, ympäristöalan ja biodiversiteetin tutkimukselle sekä antaa ekologiyhteisölle koulutusta ekoinformatiikasta. SEEK-kyberinfrastruktuurin avulla pyritään parantamaan ekologisen datan ja informaation saavutettavuutta, hajautettujen tietokonepalveluiden paikannettavuutta ja käytettävyyttä sekä mahdollisuutta käyttää uusia tehokkaita menetelmiä datan haltuun saamiseksi, toistamiseksi ja analysoimiseksi. Infrastruktuuri koostuu rakenteellisesti kolmesta osasta: datan tallettamista, jakamista,

saavutettavuutta ja analysointia tukevasta internet-arkkitehtuurista (EcoGrid), datan löytämis- ja integrointijärjestelmästä (Semantic Mediation System) sekä visuaalisesta automatisoidusta ympäristöstä, jossa ekologit voivat suunnitella, muuttaa ja yhdistellä analyyseja muodostaakseen uusia työnkulkuja (workflows) ja malleja. SEEK kattaa monia kyberinfrastruktuurivälineitä kompleksisen ekologisen datan integroimiseen ja mahdollistamaan tieteellisten analyysien nopean kehittämisen ja uudelleenkäytön. (SEEK 2008; Michener et al. 2007)

Muista ekoinformatiikan projekteista voisi vielä mainita SPiREn (Semantic Prototypes in Research Ecoinformatics), joka on viiden yhdysvaltalaisen tutkimustahon yhteistyöprojekti. SPiRE on vuonna 2003 aloitettu projekti, jonka tarkoituksena on kehittää prototyyppejä semanttisen webin luomien mahdollisuuksien tutkimiseen datan löytämiseen, tietämyksen jakamiseen ja yhteistyön toteuttamiseen liittyvissä ongelmissa. Projektin ekologisena tutkimusalueena ovat olleet tulokaslajit, joiden tutkiminen vaatii suuren määrän laaja-alaista havainnointia. Yksi projektissa aikaan saaduista prototyypeistä on ELVIS (the Ecosystem Location Visualization and Information System), joka on ravintoverkkojen ennustamiseen tarkoitettu välineistö. (SPiRE; Sachs et al. 2006)

4.3 Ekoinformatiikan julkaisu- ja yhteistyöfoorumit

´Ecological Informatics´ on kansainvälinen ekoinformatiikkaa ja laskennallista ekologiaa käsittelevä lehti, joka on ilmestynyt vasta vuodesta 2006 alkaen. Lehden sisältö on varsin teknologiapainotteinen ja edustaa niin sanottua eurooppalaista

näkemystä ekoinformatiikasta. Lehti on tarkoitettu ekologeille, biologeille, matemaatikoille sekä tietojenkäsittely- ja informaatiotieteilijöille ekologisia analyyseja, synteesiä ja ennustusta koskevien uudenlaisten lähestymistapojen edistämistä ja jakamista varten (Recknagel 2006). Periaatteessa alalla on olemassa myös toinen lehti ´Ecoinformatics´, josta tosin ei ole vielä ilmestynyt ensimmäistäkään numeroa (Ecoinformatics).

Ekoinformatiikka-artikkeleita on löydettävissä myös monista ekologian ja / tai biologian, tietoteknologian ja jonkin verran mahdollisesti myös informaatioalan lehdistä (esim. BioScience, Journal of Intelligent Information Systems ja International Journal on Digital Libraries). Ekoiformatiikka-informaation löytymistä näistä kuitenkin hankaloittaa se, ettei ekoinformatiikkaa läheskään aina mainita artikkelin otsikossa tai asiasanana.

Ecoinformatics.org (http://www.ecoinformatics.org/) on suunnittelijoiden ja tutkijoiden avoin hanke, jonka tarkoituksena on tuottaa ohjelmistoja, järjestelmiä, julkaisuja ja palveluja, jotka ovat hyödyllisiä ekologialle ja ympäristötieteille. Sivusto toimii hyvänä ekoinformatiikan alan tiedonlähteenä, sillä sinne on koottu linkit useimpiin LTER:n, NCEAS:n ja muiden tahojen projekteihin sekä moniin muihin ekoinformatiikka-lähteisiin. Sivustolla on myös pienivolyyminen postituslista, jonka avulla voi muun muassa seurata eri ohjelmistojen kehittymistä sekä LTER:n tietohallinnon tapahtumia LTER DataBits -tiedotuslehtien kautta.

Alan perustietoa sekä tietoa eri projektien etenemisestä on saatavilla myös erilaisten tapahtumien yhteydessä pidetyistä esitelmistä ja niistä tehdyistä erillisjulkaisuista.

Esimerkiksi Amerikan ekologinen yhdistys (Ecological Society of America, ESA) järjestää joka vuosi monipäiväisen kokoontumisen, jossa käsitellään ekologiaan liittyviä asioita erittäin monipuolisesti, siis myös ekoinformatiikkaa. Toinen mainitsemisen arvoinen tapahtuma on vuonna 2000 järjestäytyneen toimintansa aloittaneen kansainvälisen ekoinformatiikkayhdistyksen (International Society of Ecological Informatics, ISEI) säännöllisesti järjestämä konferenssi 'International Conference on Ecological Informatics' (ISEI). Näiden lisäksi ekoinformatiikkaan kuuluvia aiheita käsitellään enemmän tai vähemmän myös monien muiden ympäristö-, informaatio- ja tietojenkäsittelytieteiden alojen tapahtumien yhteydessä.

5 EKOLOGISEN DATAN HALLINTA

Tässä luvussa käsitellään tarkemmin ekoinformatiikan keskeisintä tutkimusaluetta, eli ekologisen datan hallintaa ja sen eri osa-alueisiin liittyviä erityispiirteitä ja haasteita.

Datanhallinta on prosessi, joka alkaa tutkimusprojektin suunnitteluvaiheessa ja voi ulottua reilusti yli datan analysointi- ja tulosten julkaisuvaiheiden (Michener 2000d, 143). Perinteisesti ekologisen tutkimusdatan hallinta on päättynyt tutkimuksen päättyessä ja tutkimustulosten valmistuttua, mutta ekoinformatiikassa tavoitellaan datan elinkaaren pidentämistä mahdollisimman pitkälle. Datanhallinta on tässä tutkielmassa käsitetty kattamaan kaikki datan elinkaaren aikana vastaan tulevat toiminnot.

Tehokas datanhallinta helpottaa varsinaista tutkimuksentekoa. Pitkäkestoiset ekologiset tutkimukset edellyttävät, että data taltioidaan yhdenmukaisesti, dokumentoidaan asiaankuuluvasti, arkistoidaan digitaalisesti ja on saavutettavissa elektronisesti. Käytetyt välineet, teknologia ja analysointimenetelmät voivat tutkimuksen aikana muuttua, jolloin datanhallinnan painopisteeksi muodostuu havaintojen taltioinnin yhtenäisyyden ja jatkuvuuden takaaminen. (Karasti 2007, 3)

Ekologisen datan hallinta voidaan ekoinformatiikan pyrkimykset huomioon ottaen jakaa kolmeen erilliseen vaiheeseen. Ensimmäisessä vaiheessa dataa kerätään, analysoidaan ja tuotetaan osana meneillään olevaa tutkimusprojektia. Toisessa vaiheessa dataa valmistellaan laajempaa jakamista ja uudelleenkäyttöä varten. Kolmannessa vaiheessa datanhallinta keskittyy datan mahdollisen uudelleenkäytön

tukemiseen. Ekologisen datanhallinnan malli voidaan lisäksi ajatella kehämäiseksi, sillä datan mahdollinen uudelleenkäyttö toisessa tutkimuksessa on jälleen uuden tutkimuksen puitteissa datanhallinnan ensimmäinen vaihe, josta syntyy uudenlaista jaettavaksi ja säilytettäväksi tarkoitettua tutkimusdataa.

Datanhallinnan eri vaiheissa tehdyillä toimilla on merkittävä vaikutus myöhempiin vaiheisiin. Esimerkiksi kahdenkymmenen vuoden kuluttua datan keräämisestä on liian myöhäistä havahtua siihen, että tietty data ei jostain syystä olekaan enää saatavilla ja / tai tulkittavissa tällöin käsillä olevaa tehtävää varten (Brunt 2000, 26). Merkittävä vaikutus datanhallinnan ensimmäisen vaiheen toimilla myöhempiin vaiheisiin on siinä, että ekologista dataa kerätään tyypillisesti vastaamaan yksittäisen tutkimusprojektin spesifejä tarpeita (Helly et al. 1999, 5), mikä vaikeuttaa sen valmistelua ja soveltamista uudelleenkäyttöön. Toisin sanoen, suurinta osaa ekologisesta datasta ei tällä hetkellä kerätä tulevaisuudessa mahdollisesti tapahtuvaa jakamista ja uudelleenkäyttöä huomioon ottaen (Sim et al. 2004).

Kaikki ekologit osallistuvat datanhallintaan osana normaalia tutkimustoimintaansa. Monesti yksittäisen tutkijan suorittama datanhallinta kuitenkin koostuu ainoastaan niistä toiminnoista, jotka ovat välttämättömiä datan valmistelemiseksi analysointia varten (kuten datan syöttämisestä, laadunvarmistuksesta ja käsittelystä) ja kiinnostus datanhallintaan loppuu usein heti tulosten julkaisemisen jälkeen. (Brunt 2000, 28) Yksittäisten ekologien osallisuus datanhallinnasta rajoittuu siten yleensä ainoastaan datanhallinnan ensimmäiseen vaiheeseen ja on luonteeltaan lyhytnäköistä.

Tässä luvussa puhutaan nimenomaan datanhallinnasta, vaikka yhtä hyvin monin paikoin voitaisiin myös käyttää termiä informaationhallinta, sillä dataan on aina liitettävä metadataa, jonka myötä ollaankin jo tekemisissä informaation kanssa (ks. 3.3.1). Termiä datanhallinta käytetään kuitenkin tutkimusaineistossa yleisesti myös näissä yhteyksissä. Selkeyden vuoksi sekä datan keskeisen aseman korostamiseksi myös tässä työssä päädyttiin tämän saman termin käyttöön viittaamaan datanhallinnan elinkaareen. Kannattaa kuitenkin pitää mielessä, että ekologisen datan hallintaan liittyy varsin kiinteästi myös informaationhallintaa ja lisääntyvässä määrin myös tietämyksenhallintaa.

Hyvän datanhallinnan yleisenä tavoitteena voidaan pitää datan laadun maksimointia, joka puolestaan voidaan saavuttaa vain kokonaisvaltaisella lähestymistavalla datanhallintaan (Wilson 2007a). Laatukysymykset nousivat merkittävinä esille tämänkin tutkielman aineistosta ja niitä tarkastellaan pääpiirteittäin osana eri datanhallintavaiheita, joita esitellä seuraavaksi hieman tarkemmin.

5.1 Datanhallinnan I vaihe

Datanhallinnan ensimmäinen vaihe käsittää datan osana sen synnyttänyttä tutkimusta. Ensimmäinen vaihe kattaa datan keräämisen suunnittelun ja toteutuksen, datan valmistelun ja analysoinnin sekä informaatioksi muovaamisen, päättyen tutkimuksen loppumiseen ja tulosten julkaisemiseen. Tämän jälkeen tutkimuksessa muodostettujen datatiedostojen tulisi olla valmiita siirtymään datanhallinnan toiseen vaiheeseen. Tutkimusdataa koskevan datanhallinnan tulisi alkaa jo tutkimusprojektin suunnittelu-

vaiheessa, jolloin on tärkeää sopia hankkeen yhteiset näytteenottotavat ja talteenottomenetelmät. Tällä pyritään takaamaan, että jokainen näytteenottaja taltioi tutkittavasta ilmiötä tai eliöstä samat muuttujat samalla tarkkuudella (Zuur et al. 2007, 8).

Ekoinformatiikan laajempien tavoitteiden, koko ekologiyhteisön ja muiden ekologista dataa tarvitsevien kannalta datanhallinnan I vaiheessa on ennen kaikkea ongelmana se, että käytännössä jokainen tutkija ja tutkimusryhmä toteuttaa datanhallintaa miten itse parhaaksi ja tarpeelliseksi katsoo. Käytännön kokemusten mukaan ekologisesta raakadatasta huolehtiminen onkin usein toteutettu varsin huonosti (Zuur et al. 2007, 7). Ekologisen datan keräämis- ja säilyttämistavat ovat lähestulkoon yhtä vaihtelevia kuin datan dokumentoima luonto itse (Borgman, et al. 2007, 22). Tästä heijastuu luonnollisesti huomattavia ongelmia datanhallinnan myöhempiin vaiheisiin, joissa hallittavia datatiedostoja pyritään yhtenäistämään niiden käytettävyyden parantamiseksi. Ratkaisuina ongelmaan on tutkimusaineistossa yleisesti esitetty standardoitujen menetelmien käyttöä, dokumentointikäytännön lisäämistä ja parantamista sekä asennemuutoksen aikaansaamista tutkimuksentekoa ja siihen kuuluvaa datanhallintaa kohtaan.

Syynä epävirallisen datanhallintaan voidaan nähdä ainakin kaksi seikkaa. Ensinnäkin raaka/primaaridatalla nähdään yleensä arvoa vain sekundaaridatan ja tutkimustulosten tuottamisessa, jolloin siitä huolehtimista ei ekologien keskuudessa pidetä yleisesti ottaen niin tärkeänä. Toiseksi datanhallinnan taidot on yleensä opeteltava itse, sillä datanhallinnan opetusta ei pääsääntöisesti sisälly ekologien akateemiseen koulutukseen (Jones et al. 2006, 522), eikä niihin paneuduta ekologian, sen paremmin kuin

tilastotieteen tai tietokantoja käsittelevissä oppikirjoissakaan lainkaan (Michener 2000c, 3).

Ongelma ei paradoksaalisesti ole niin merkittävä mittavimmissa ekologisissa tutkimusprojekteissa, joissa datanhallinnan toimivuuden merkittävyys projektin menestykselle on suurin. Niissä on yleensä mukana myös tietohallintohenkilöstöä ja näin ollen jonkinlaista opastustakin datanhallintaan on saatavilla. Esimerkiksi Cook et al. (2001) ovat kenttätyöntekijöiden pyynnöstä kirjoittaneet ohjeistuksen niihin datanhallintatoimiin, joita ekologien tulisi tehdä kerätessään dataansa, jotta datan myöhempi käytettävyys paranisi. Ohjeen ovat sittemmin päivittäneet Hook et al. (2007). Ohjeessa esitetyt seitsemän tärkeintä datanhallintatointa ovat (1) tiedostojen kuvaileva nimeäminen, (2) johdonmukaisten ja vakaiden tiedostomuotojen käyttö, (3) datatiedostojen sisällön määrittely, (4) datan johdonmukainen järjestäminen, (5) peruslaadunvarmistuksen toteuttaminen, (6) datatiedostokokonaisuuksien kuvaileva nimeäminen sekä (7) dokumentaation tuottaminen. Ohjeistus antaa jonkinlaisen käsityksen siitä, minkälaista variaatiota datanhallintakäytännöissä eri ekologien välillä voi olla ja minkälaisia arkistointi-, tulkinta- ja käyttövaikeuksia tällainen vaihtelevuus voi aiheuttaa.

5.1.1 Datan kerääminen

Ekologista datankeruuta, tai tutkimustoimintaa yleisemminkään, eivät ohjaa mitkään fysiikan lakien kaltaiset suuret teoriat, vaan käsitykset siitä mitä ylipäätään on havainnoitavissa ja mitattavissa. Myöskään ekologiassa käytettyihin mittaustapoihin ei olemassa yleispäteviä ohjeita, vaan tyypillisesti samankin ulottuvuuden (esim. pituus

tai etäisyys) mittaustavoissa esiintyy vaihtelua yksittäiselläkin tutkijalla tilanteesta riippuen. Kokeneemmista tutkijoista voi usein olla apua mittausten suorittamistapaa päätettäessä. Ekologisen datan muodostukseen liittyy siis paljon epävarmuutta. (Roth & Bowen 1999, 719-721)

Ekologit joutuvat usein myös työskentelemään vähemmän ihanteellisissa ja varsin kontrolloimattomissa olosuhteissa; tutkimuksen koeyksiköt ovat heterogeenisiä ja kokeiden toistaminen voi olla vaikeaa tai mahdotonta. Tehtävät toimenpiteet puolestaan saattavat olla standardoimattomia ja koostua kertasuorituksen sijaan sarjasta toistuvia käsittelyitä, kuten ekologisissa ennallistamisprojekteissa. (Michener 2000d, 142) Eräs ongelma piilee myös siinä, että eri kenttä- ja laboratoriomenetelmillä saadaan samaa ilmiötä mitattaessa usein eri tuloksia. Mikäli vakiintunut menetelmä on olemassa ja hyvin dokumentoituna löydettävissä alan kirjallisuudesta, voivat ekologit hyötyä monin tavoin tällaisten menetelmien omaksumisesta. Muun muassa vertailu muihin tutkimuksiin helpottuu, menetelmien dokumentointi julkaisuja ja metadataa varten yksinkertaistuu ja kustannuksetkin ovat usein alhaisemmat. Standardoituja menetelmiä ei kuitenkaan kaikkiin tapauksiin ole välttämättä edelleenkään olemassa tai ne voivat olla syystä tai toisesta kyseiseen tutkimukseen epäsopivia. (Michener 2000c, 16)

Ekologinen data kerätään suurimmaksi osaksi edelleen manuaalisesti, jolloin se otetaan talteen ensin paperille joko erityisille kaavakkeille tai muistivihkoon, tai vaihtoehtoisesti esimerkiksi ääninauhurille, ja siirretään myöhemmin tietokoneelle analysoitavaksi ja säilytettäväksi. Paperille tallettamisen hyvinä puolina voidaan pitää halpuutta, pitkäikäisyyttä ja käytön vaivattomuutta, mutta paperi ei kuitenkaan sovellu

hyvin kaikkiin ympäristöolosuhteisiin eikä salli kovinkaan paljoa muokkaamista ennen tietokoneelle siirtoa. Ääninauhurin kanssa ongelmaksi voi muodostua erilaiset toimintahäiriöt ja tarvittavat huoltotoimenpiteet. Enenevässä määrin tutkijat keräävät dataa kentällä myös suoraan kämmenmikroihin, mikäli ympäristöolosuhteet vain sen sallivat. Myös automaattisten datankeräysvälineiden käyttö on lisääntymässä, jolloin data myös tallentuu suoraan tietokoneelle. (Brunt 2000, 34-35; Michener et al. 2002, 2)

5.1.2 Datan valmistelu

Datan valmistelu pitää sisällään erilaiset datalle tehtävät toimenpiteet ennen kuin se voidaan analysoida. Esimerkiksi tiedostoissa, joihin on kerätty dataa pitkiltä ajanjaksoilta, voi olla talletettuna dataa eri muodoissa eri vuosina muuttujien ja tarkkuuden vaihdellessa kerääjän mukaan. Lisäksi osa datasta saattaa olla ajan saatossa kadonnut ja osa muuttujista nimetty vaikeaselkoisilla koodeilla (Zuur et al. 2007, 7). Tällaista dataa on vaikea analysoida ennen perusteellista esikäsittelyä. Valmistelu-vaiheeseen kuuluu esimerkiksi datatiedostojen yhdisteleminen tai jakaminen osiin sekä datan määrän vähentäminen, jota saatetaan tarvita automaattisia datankeräys-menetelmiä käytettäessä. (Michener 2000d, 143)

Ekoinformatiikan kannalta tärkein tähän vaiheeseen luettava toiminto on integrointi, jolla tarkoitetaan eri lähteistä peräisin olevan datan yhteensovittamista ja –liittämistä siten, että lopputulos on mielekäs ja käyttökelpoinen. Aihetta on tutkittu paljon tietojenkäsittelytieteissä ja näin ollen monenlaisia yleisiä lähestymistapoja ja järjestelmiä on jo olemassa. (Jones et al. 2006, 523 ja 532) Ekologisen datan integrointia erikseen käsitteleviä artikkeleita ei juurikaan ole, mutta teemana se näkyy

jossain määrin kaikissa ekoinformatiikkaa sivuavissa artikkeleissa ja vaikuttaa haasteellisimmalta ekologisen datan valmisteluun liittyvältä teemalta.

Vaikka yksittäisten tutkijoiden havainnoista ja kokeista saatu data on yhä ekologisen tutkimuksen keskiössä, kasvaa tällaisen datan arvo huomattavasti mikäli sitä yhdistellään paljastamaan tärkeitä malleja ja tuottamaan laajoja yleistyksiä. Datan yhdisteleminen mahdollistaa laajemman perspektiivin ajallisesti, alueellisesti ja tieteenalallisesti kuin yhden tai muutaman yksittäisen tutkimuksen avulla on mahdollista. (Jones et al. 2006, 521) Tarve erilaisen datan integroimiselle ekologiassa ei ole uusi, mutta mahdollisuus sen toteuttamiselle on (Porter & Ramsey Jr. 2002, 396). Monet tulevaisuuden edistysaskeleet ekologiassa tulevatkin todennäköisesti riippumaan kyvystä integroida erilaisia tietokoneistettuja datatiedostoja (Brunt 2000, 27).

Datan integrointiprosessiin kuuluu päättäminen siitä, voidaanko kaksi tai useampi dataresurssi tehokkaasti yhdistää ja kuinka. Integroinnin lopputuloksena pitäisi syntyä yhtenäinen datatuote, jossa on ratkaistu mukaan otettujen datalähteiden väliset eroavaisuudet. (Jones et al. 2006, 532) Datan integrointi voidaan suorittaa joko manuaalisesti tai tietokoneavusteisesti.

Porterin ja Ramseyn (2002, 396) mukaan projektikohtainen ekologisen datan integrointi on säilynyt ensisijaisesti manuaalisena prosessina. Prosessi alkaa potentiaalisten datalähteiden tunnistamisella ja hankkimisella. Tämän jälkeen päätetään datan liittämiseen käytetyt parametrit ja aggregaation laajuus ja aste sekä tehdään tarvittavat muutokset eri datalähteisiin. Aggregaatiota tarvitaan esimerkiksi

silloin kun osa datasta on kerätty tunnin välein ja muu data kerran vuodessa. Avainvaihe datan integroinnissa on liityntöjen tunnistaminen. Ekologiselle datalle tärkeimmät liitynnät ovat yleensä ajallisia ja maantieteellisiä, mutta myös taksonomiset liitynnät ovat mahdollisia. Järjestelmäperustainen integrointi on riippuvainen standardoitujen datalähteiden ja niiden integrointia tukevien ohjelmistojärjestelmien kehittämisestä. Laajanmittakaavan datan integroinnin haasteet ovatkin jatkuvan tutkimuksen kohteena niin tietojenkäsittelytieteessä kuin ekologisissa yhteisöissäkin. (Porter & Ramsey 2002, 397)

Integrointia tapahtuu käytännössä monella eri tasolla. Järjestelmätason integroinnissa pyritään ratkaisemaan eroavaisuudet liittyen verkostoprotokolliin (esim. http / ftp tiedoston siirrossa), käyttöjärjestelmiin ja datanhallintasovelluksiin. Järjestelmätason integrointia tarvitaan tukemaan datan saavutettavuutta ja siirtoa, mutta se ei yksinään riitä takaamaan, että muodostuva datatuote olisi tieteellisessä mielessä käyttökelpoinen tai tulkittavissa. Rakennetason integroinnissa käsitellään eroavaisuuksia datan esittämistavoissa, kuten sitä, onko dataobjekti kuvatiedosto vai taulukko. Datamalli-tason integroinnissa puolestaan lähdetään liikkeelle siitä, kuinka datalähteet on loogisesti strukturoitu ja keskitytään sovittamaan yhteen datatiedoston piirteitä, kuten samalla tavoin nimettyjä muuttujia. (Jones et al. 2006, 533) Viimeinen integroinnin taso eli semanttinen integrointi, on kaikkein haasteellisin ja siitä syystä sitä käsitellään seuraavassa muita tarkemmin.

Semanttisessa integroinnissa datan sisältöä selvitellään kontrolloituja sanastoja muistuttavalla tavalla, mutta käyttäen tehokkaampia formaaleja rakenteita eli ontologioita. Datan sisällön tulkinnassa voi esiintyä epävarmuutta koskien koko

datatiedostoa, sen yksittäisiä muuttujia (taulukkomuotoisissa tiedostoissa) tai jopa sitä laajempaa kontekstia, joka johti datan keräämiseen. Ontologiat luovat formaalin mallin ala- tai aihepiirikohtaiseen tietämykseen sisältyvistä tarkkaan määritellyistä käsitteistä tai termeistä sekä selvät täsmennykset näiden välisistä suhteista. Ekologian alalla on käynnissä kaksi projektia (SEEK ja SPIRE; ks. 4.2), joissa kehitetään ontologioita tutkimaan kuinka semanttiset lähestymistavat voivat auttaa datan integroinnissa. (Jones et al. 2006, 533-534) Ontologioita voidaan tulevaisuudessa käyttää esittämään tieteellistä ekologista dataa semanttisessa webissä, mutta ilmiön uutuudesta johtuen semanttista webiä ei vielä olla käytetty ekologisen datan julkaisukanavana (Williams et al. 2006, 241).

Datan tarkan merkityksen eli semantiikan selvittämiseen liittyvän epävarmuuden lisäksi ekologisen datan integroinnista tekee haasteellista näytteenoton ajallinen ja alueellinen vaihtelevuus, taksonomiset epäyhtenäisyydet näytteiden tunnistamisessa sekä muuttujien ja niiden mittayksiköiden omaperäinen merkitseminen (Jones et al. 2006, 521).

5.1.3 Datan analysointi

Datan analysoinnilla tarkoitetaan dataan kohdistettuja toimenpiteitä, joilla siitä saadaan rikastettua informaatiota. Ekologisen datan muuttamiseksi informaatioksi on käytettävissä monia erilaisia analyyttisiä välineitä. Käytännössä tutkijat ovat usein rajoitettuja suorittamaan analyysinsa tietyssä ohjelmistopaketissa saatavilla olevia menetelmiä ja välineitä käyttäen. Useissa analyyseissa kuitenkin haluttaisiin hyödyntää monien eri ohjelmistopakettien tarjontaa, jolloin dataa joudutaan siirtelemään paketista

toiseen. Tämänkaltainen ohjelmistopakettien yhdisteleminen on haastavaa ja se tekee datan kulun seurannasta ja uudelleenrakentamisesta sekä analyysin toistamisesta vaikeaa ennen kaikkea siksi, että prosessin yhteydessä yleensä häviää informaatiota dataan sovelletuista analyyttisista menettelytavoista. (Jones et al. 2006, 535) Ratkaisuksi näihin ongelmiin on alettu kehittää kohdassa 3.2.1 esiteltyjä analyysiprosesseja mallintavia työnkulkujärjestelmiä.

Yleisiä keinoja ekologisen datan analysoinnissa ovat visualisointi ja mallintaminen, jotka ovat näkyvä osa etenkin eurooppalaista ekoinformatiikkaa. Tämän tutkielman painotus informaationäkökulmaan ei kuitenkaan näe mallintamista niin keskeisenä osana ekoinformatiikkaa, että sitä, sen paremmin kuin muitakaan analysointi-menetelmiä käsiteltäisiin tässä sen tarkemmin. Mallintaminen ja muut analysointi-menetelmät sijoittuvat informaationäkökulman sijaan pikemminkin ekologian ja teknologian rajapintaan.

5.1.4. Datan laatu

Datanhallinnan I vaiheessa ekologisen datan laatu on ensinnäkin erottamattomasti sidoksissa dataa keräävän henkilöstön tietämys- ja taitotasoon, mikä tarkoittaa sitä, että laadukkaan datan takaamiseksi kannattaa henkilöstön koulutukseen panostaa. Toinen datan laatuun vaikuttava tekijä on käytettyjen välineiden luotettavuus ja tarkkuus. Kolmanneksi tapa, jolla data hankitaan (paperilomake, nauhuri, kenttäkäyttöinen tietokone, ym.) liittyy myös datan laatuun vaikuttamalla inhimillisten virheiden esiintymiseen mittaustuloksissa. (Michener et al. 2002, 2)

Ekologinen data on keräysvaiheessaan altis monenlaisille virheille, mukaan lukien lajitunnistusvirheet, laadullisen tai kategorisen datan koodausvirheet ja määrällisten muuttujien mittausvirheet. Esimerkiksi riski tunnistaa laji väärin vaihtelee luonnollisesti lajiryhmästä toiseen ja on vahvasti sidoksissa kenttätyöntekijöiden taitoihin ja kokemukseen sekä ammattiohjauksen saatavuuteen. Ohjausta voi saada joko erilaisista lajintunnistusoppaista tai kokeneemmilta tutkijoilta. (Wilson 2007a) Aina ei siltikään, muun muassa hankalien olosuhteiden vuoksi, voida olla varmoja onko luotettava havainto todella tehty (Roth & Bowen 1999, 719).

Tarkoituksenmukaisesti suunnitellut datakaavakkeet ovat helppokäyttöisiä ja tarjoavat pitkäaikaisen paperitulosteen datasta (Michener et al. 2002, 2), mutta käsin tehdyt merkinnät voivat toisinaan olla myöhemmin vaikeasti tulkittavia (Zimmerman 2003, 5). Datan kerääminen nauhurilla eliminoi datakaavakemenetelmään liittyvään puhtaaksikirjoittamiseen liittyvät ongelmat, mutta datan laatua voivat heikentää teknisten toimintahäiriöiden lisäksi erilaiset hälyäänet kuten tuulen humina (Michener et al. 2002, 2). Datan kerääminen suoraan kämmenmikroon tai automaattisilla datankeräysmenetelmillä on suositeltavaa aina kun niiden käyttö on mahdollista, sillä datan siirtely muodosta toiseen on aina laadun kannalta kriittinen vaihe ja sitä tulisi pyrkiä välttämään. Kämmenmikroihin kerätyn sekä muun tietokoneelle myöhemmin siirretyn datan osalta on tärkeää huolehtia riittävän usein tapahtuvasta tiedostojen varmuuskopioinnista. (Brunt 2000, 35-36)

Myös datan valmistelu- ja analysointivaiheet vaativat huolellista keskittymistä datan laadun ylläpitämiseen, sillä virheiden esiintyminen missä tahansa analyysiprosessin vaiheessa voi saada aikaan harhaanjohtavia tuloksia (Porter & Ramsey 2002). Näissä

vaiheissa datan laadusta huolehditaan erilaisilla laadunvarmistus- ja -valvontatekniikoilla, kuten datanhallinnan seuraavassakin vaiheessa, ja näitä tekniikoita käsitellään kohdassa 5.2.4.

5.2 Datanhallinnan II vaihe

Ekologisen datan hallinnan toisen vaiheen tarkoituksena on tavallaan valmistella dataa mahdollista uudelleenkäyttöä varten. Vaiheeseen voidaan katsoa sisältyvän sellaiset varsin laajat ja ekoinformatiikassa paljon huomiota saaneet teemat kuin jakaminen, arkistointi ja dokumentointi. Nämä teemat ovat laajemmassa mitassa varsin tuoreita ilmiöitä ekologisen datan hallinnassa, mutta uudenlaisen ekologisen tutkimuksen mahdollistajina erittäin tärkeitä tutkimusaiheita ekoinformatiikassa.

5.2.1 Jakaminen

Jakaminen on tutkimusaineiston perusteella yksi oleellisimmista ekoinformatiikan teemoista, sillä termi jakaminen esiintyy poikkeuksetta kaikissa ekoinformatiikkaan liittyvissä artikkeleissa. Jakamisella tarkoitetaan tässä yhteydessä lähinnä datan asettamista julkisesti muiden tutkijoiden saataville, sillä tämä on erilaisista jakamisen muodoista se, mitä ekoinformatiikassa ennen kaikkea halutaan edistää.

Olson ja McCord (2000) ovat erottaneet ekologisen datan jakamisessa kolme erilaista vaihetta. Ensimmäisessä vaiheessa datan tuottaneen tutkimuksen tuloksia ei ole vielä julkaistu eikä datan jakamista juuri tapahdu kuin korkeintaan tutkimuksen muille

tutkijoille tai muille lähikollegoille. Toisessa vaiheessa tutkimustulosten julkaisu herättää yleisempää kiinnostusta dataan, myös muiden alojen edustajissa, jotka yleensä hankkivat datan arkistoista. Toisen vaiheen jälkeen saattaa kulua pitkäkin aika, jolloin dataan kohdistuu hyvin vähän mielenkiintoa. Aikanaan kiinnostus dataan jälleen herää liitettäessä se osaksi laajempaa kokoelmaa. Dataa odottaa näin ollen tulevaisuudessa mahdollisuus tulla jälleen ekologisen tutkimuksen kohteeksi. Muuntuminen uudeksi tietämykseksi kuitenkin katoaa sellaiselta datalta, jota ei ole arkistoitu tai riittävästi dokumentoitu. (Olson & McCord 2000, 125-127)

Suurin osa ekologisesta tutkimuksesta on julkisrahoitteista, jolloin voidaan väittää, että myös tällaisesta tutkimuksesta saatu data on julkista ja tulisi olla julkisesti saatavilla. Datan jakaminen tuo lisäkatetta tutkimusinvestoinneille ja poistaa tarpeen kerätä samanlaista dataa uudelleen. Datan jakamisen laajemmat vaikutukset ovat myös tärkeitä. Voivatko tutkijat moraalisesti perustella datan yksityisenä pitämistä, jos tämän datan avulla voitaisiin nopeuttaa ratkaisujen löytämistä ympäristöllisiin ja ympäristönsuojelullisiin haasteisiin? (Parr & Cummings 2005, 362)

Perinteisesti ekologit eivät ole jakaneet dataa. Tavallisimmin ekologisen datan jakamista onkin tähän mennessä tapahtunut vain lähikollegojen kesken (Gross et al. 1995, 8). Syynä on riittävien yllykkeiden puuttumisen lisäksi ollut ennen kaikkea ekologisen datan kompleksinen luonne. (Zimmerman 2007, 7). Ekologisen datan jakamisen tarve on kuitenkin tutkimuskysymysten muuttumisen ja siitä seuranneen monitieteisten tutkimusten yleistymisen myötä lisääntynyt ja laajentunut. Ajatus ekologisen datan laajemmasta jakamisesta onkin periaatteessa jo hyväksytty, mutta käytännössä jakamista tapahtuu yhä vähän, johtuen esimerkiksi kysynnän vähyydestä,

standardien puuttumisesta sekä datan julkaisemiseen, omistajuuteen, laatuun ja etiikkaan kohdistuvista huolista (Borgman et al. 2007, 17).

Kysynnän puutteella viitataan todennäköisesti siihen, että useilla ekologisilla, pienistä tutkimuksista kertyneillä datatiedostoilla koetaan olevan vain pieni uudelleenkäyttö-potentiaali, johtuen niiden rajallisista keräyskonteksteista (Zimmerman 2003, 154). Datan omistajuuteen liittyvät seikat nousevat merkittäviksi erityisesti pienen tutkimusryhmän keräämän datan kohdalla (Zimmerman 2007, 7).

Datan jakamista estävät lisäksi palkitsemisen puute, huoli datan väärinkäytöstä ja prioriteettien luominen (Zimmerman 2003, 217). Myös jakamiseen läheisesti liittyvistä datan dokumentoinnista, siirrosta ja tallennuksesta aiheutuvat kustannukset haittaavat ekologisen datan jakamisen yleistymistä (Zimmerman 2003, 3).

Parr ja Cummings (2005) pitävät datan jakamisen yleistymättömyyden syynä kahta ilmeistä seikkaa: ensinnäkin tutkijat haluavat käyttää dataa seuraaviin töihinsä ilman kilpailua ja toiseksi he uskovat, että datan jakamiselle on olemassa logistisia esteitä. Datan salaaminen mahdollisen tulevaisuuden hyödyn vuoksi on kuitenkin lyhytnäköistä, sillä akateeminen palkkiojärjestelmä suosii datan jakamista: datan arvo nousee muiden käyttäessä sitä, millä on suoria seurauksia käsitykseen tutkimuksen tärkeydestä sekä objektiivisiin mittareihin (kuten viittausasteeseen). Näitä käsityksiä ja mittareita on käytetty sekä virallisesti että epävirallisesti kriteereinä julkaisemiselle, rahoituksen myöntämiselle ja uran edistämiselle. Myös datan jakamisen logistiset esteet ovat siinä mielessä harhaa, että datan jakamiselle on jo olemassa sopivia menetelmiä. Tutkija voi esimerkiksi liittää datatiedoston julkaistavaksi lähettämänsä

artikkelin mukaan alan lehtiin sekä lähettää datan instituutio- tai projektikohtaisille verkkosivuille. (Parr & Cummings 2005, 362)

Ekologia ja evoluutiobiologia ovat Porterin ja Calahanin (1994, 195) mukaan lähestulkoon ainoita aloja ympäristötieteiden ja luonnonympäristöön liittyvien tieteiden joukossa, joilta yleisesti ottaen puuttuvat jonkin laitoksen tai yhteisön vaatimat datan arkistointi- ja jakamisperiaatteet. Yleisten datan jakamiskäytäntöjen puuttuminen on myös yksi merkittävä ekologisen datan jakamista haittaava tekijä. Ekologialta puuttuu vielä myös formaali infrastruktuuri datan jakamiseen (Zimmerman 2003, 94).

Datan jakamisen yleistymiseksi ekologiassa tulee siitä koituvat edut ensin tunnistaa yksittäisten tutkijoiden tasolla ja heidät tulee saada näkemään olemassa olevat keinot voittaa aiemmin jakamisen tiellä havaitut esteet. Aktiivisempi datan jakaminen itsessään kannustaisi standardoinnin lisääntymiseen, sillä parhaiten selitettyä ja kerättyä dataa tullaan todennäköisimmin käyttämään uudelleen ja viittaamaan. (Parr & Cummings 2005, 362) Datan jakamisen kustannusten, hyötyjen ja seurausten arvioinnista tai datan jakamisen vaikutuksista tieteen sisältöön, tekemiseen ja kommunikointiin liittyvien olettamusten testaamisesta on kuitenkin tehty vasta hyvin niukasti empiiristä tutkimusta. Erityisesti yksittäisten tutkijoiden tai pienten tutkimusryhmien keräämien havaintodatatiedostojen jakamiseen liittyvistä haasteista, eduista ja tuloksista tiedetään hyvin vähän. (Zimmerman 2003, 40)

Tällä hetkellä ekologit tuntuisivat olevan halukkaampia jakamaan jo julkistamiinsa tutkimuksiin liittyvää dataa kuin dataa, josta saadut tulokset he ovat vasta aikeissa

julkistaa. Ekologit ovat myös valmiimpia jakamaan dataa, jonka keräämiseen on nähty vähiten vaivaa. Toisin sanoen, automaattisten datankeräysteknologioiden yleistyminen on yksi ekologisen datan jakamista edistävä tekijä. (Borgman et al. 2007, 25)

5.2.2 Arkistointi

Ekologisen tutkimusdatan pitkäaikaisarkistointi on keskeinen keino edellä käsitellyn datan jakamisen ja sitä kautta myös ekologisen tutkimuksen edistämiseksi. Arkistointi mainitaan pääsääntöisesti kaikissa ekoinformatiikkaa käsittelevissä artikkeleissa, mutta nimenomaan ekologisen datan arkistointiin keskittyviä artikkeleita ei juurikaan löytynyt.

Perinteisesti tieteellisen datan jakaminen on tapahtunut tieteellisten lehtien välityksellä, joissa julkaistu data on tarkistettu, analysoitu ja tulkittu kokemuksen ja muun datan valossa. Tässä prosessissa primaaridata on muokattu erilaisiksi johdetuiksi ja syntetisoiduiksi datatuotteiksi (Karasti et al. 2006, 330), mikä vähentää datan hyödyllisyyttä muulle tiedeyhteisölle. Tieteellisten julkaisujen pätevyyttä ekologisen tiedeyhteisön data-arkistoina heikentää myös se, että ne säilyttävät vain osan tutkimuksiin liittyvästä datasta (Baker et al. 2000, 966). Julkaisut eivät myöskään säilytä kaikkea datan tulkintaan tarvittavaa informaatiota, vain ainoastaan subjektiivisesti valitun määrän metadataa datasta tehtyjen päätelmien ymmärtämiseksi (Michener et al. 1997, 332).

Tässä tutkielmassa ja ekoinformatiikassa yleisemminkin datan arkistoinnilla tarkoitetaan yksittäisen tutkimuksen tarpeet ylittävää datan säilyttämistä, jonka

tavoitteena on taata datan laaja-alainen käytettävyys tulevaisuudessa. Data-arkistointi on ekologiassa laajemmassa mitassa vasta aluillaan, eikä ekologiselle datalle ole vielä olemassa kovin monia pysyviä data-arkistoja (Olson & McCord 2000, 117). Yksi keskeinen ekologisen datan arkistoinnin yleistymistä haittaava tekijä on vallitseva tieteellisen tutkimuksen rahoitusmalli, jossa ei huomioida tarvetta säilyttää dataa pitkään, vaan huomion keskipisteenä on tieteellisten tulosten tuottaminen ja julkaiseminen tieteellisessä kirjallisuudessa (Jones et al. 2006, 530). Tähän on toivottavasti tulossa muutos muun muassa sen myötä, kun käsitys myös tieteellisestä datasta ja dataa sisältävästä tietokannasta tieteenteon lopputuotteena yleistyy (Borgman et al. 2007, 18; Bowker 2000, 643).

Olsonin ja McCordin (2000, 117) määritelmän mukaisesti data-arkisto on kokoelma yleensä elektronisessa muodossa olevia datatiedostoja, jotka on talletettu siten, että erilaiset käyttäjät voivat paikantaa, hankkia, ymmärtää ja käyttää dataa. Lisäksi arkistoidun datan säilyvyydestä ja ylläpidosta tulee huolehtia niin, että se on suojassa luonnon ja ihmisen aiheuttamilta tuhoilta ja sellaisessa muodossa, jossa se on saavutettavissa teknologian muuttuessa.

Arkistoidun datan jatkuvasta käytettävyydestä huolehtiminen (curation, stewardship) on sekä teknisesti haastavaa että kallista (Jones et al. 2006, 530). Data-arkistoa täytyy esimerkiksi voida jatkuvasti täydentää uutta dataa lisäämällä. Lisäksi käyttäjille on annettava erilaista tukea ja tietokonejärjestelmää täytyy ylläpitää muun muassa uusimalla ohjelmistoja, laitteistoja, tallennusvälineitä ja verkkoyhteyttä (Olson & McCord 2000, 131). Tämän hetkisen datasta huolehtimisen ja arkiston käytön

tukemisen lisäksi data-arkistointiin liittyy retrospektiivistä vanhan datan pelastamista arkistoon sekä tulevan toiminnan suunnittelua (Karasti et al. 2006, 332).

Jatkuvatoiminen ja luotettava, vakiintuneille metadata- ja datan huolehtimis-standardeille perustuva data-arkisto on yhä suurelta osin kehittämättä (Jones et al. 2006, 530). Ekologisten data-arkistojen vähäisen kehityksen syynä voidaan pitää edellä mainittujen syiden lisäksi ainakin datan kompleksisuutta ja moninaisuutta sekä laaja-alaisten, pitkäkestoisten tutkimusohjelmien puuttumista. Täydellisimpiä data-arkistoja on tähän mennessä kehitetty sellaiselle ympäristöinformaatiolle, joka liittyy rajalliseen teemaan ja josta muodostuu laaja kokoelma dataa sekä ajallisesti että alueellisesti (Jones et al. 2006, 119).

Yksi merkittävä ekologisiin data-arkistoihin liittyvä ongelma on siinä, että eri kokoelmat sisältävät samoja tai päällekkäisiä datatiedostoja. Lisäksi mistä tahansa datatiedostosta on kokoelmissa todennäköisesti monia eri versioita, johtuen tiedostoihin tehdyistä lisäyksistä, virheiden korjaamisista ja muista muutoksista. Yksi mahdollinen ratkaisu tähän olisi standardoitujen datatunnistimien käyttö, jolloin datatiedostojen selkeä ja yksilöllinen tunnistaminen tulisi mahdolliseksi. Standardoitujen tunnistimien käyttö on vakiintunut kirjojen (ISBN) ja aikakauslehtien (ISSN) julkaisemiseen, mutta käytäntö ei ole yhtä yleistä elektronisen datan kohdalla. Kyky viitata tiettyyn datatiedostoon yksiselitteisesti on erityisen tärkeää tieteellisen toistettavuuden ihanteen vuoksi. (Jones et al. 2006, 529)

Vaikka ekologiselle datalle on vasta rajallinen määrä virallisia data-arkistoja, on ekologeilla tänä päivänä jo olemassa edellytykset halutessaan hallita dataansa siten, että se vastaa paikallisia turvallisuustarpeita, helpottaa datan jakamista ja valmistaa niitä lopulliseen data-arkistoon sisällyttämiseen (Olson & McCord 2000, 117). Tilanne on kuitenkin Olsonin ja McCordin (2000, 123) mukaan valitettavasti yhä sellainen, että vaikka useimmat ekologit kyllä tukevat ajatusta datan arkistoinnista ja jopa käyttävät arkistoitua dataa omissa tutkimuksissaan, he eivät tavallisesti kuitenkaan arkistoi omaa dataansa.

Virallisten arkistojen vähäisen määrän lisäksi oleellinen data-arkistoinnin suosion vähyyteen liittyvä seikka on se, että tietyssä mielessä arkistoinnista on enemmän hyötyä arkistoidun datan käyttäjälle kuin datan arkistoon luovuttajalle (Porter & Callahan 1994, 193). Olson ja McCord (2000, 123) vertaavat datan arkistointiprosessia julkaisun valmistamiseen, koska ne vaativat heidän käsityksensä mukaan yhtä paljon aikaa ja resursseja. Tiedeyhteisössä arkistointiprosessia ei kuitenkaan vielä arvosteta yhtä paljon kuin julkaisuprosessia, mikä johtaa siihen, että niukat resurssit käytetään mieluummin arvostetumpaan julkaisutoimintaan kuin datan pelastamiseen arkistoihin. Olson ja McCord (2000, 117) kuitenkin ennustavat, että sekä datan arkistoon luovuttaminen että arkistoidun datan käyttäminen tulee olemaan olennainen osa tulevaisuuden tieteellistä tutkimusprosessia.

Arkistoinnin lisäksi tutkimusaineistossa viitataan usein myös digitaalisiin kirjastoihin mahdollisena ekologisen datan jakamiskeinona. Digitaaliset kirjastot voidaan käsittää kontrolloiduiksi informaatio-objektien kokoelmiksi, joiden sisältö on digitaalisessa muodossa ja organisoitavissa, saavutettavissa, evaluoitavissa ja käytettävissä

hajautettujen, digitaalisella teknologialla tuettujen palveluiden kautta (Smith 1997, 105). Tieteellistä dataa sisältävien digitaalisten kirjastojen tavoitteena on helpottaa etsintää ja navigointia laajassa ja hajanaisessa tieteellisen datan universumissa ja datan yhdistämistä uudenlaisiin tieteellisiin kysymyksiin vastaamisessa. Tavoitteeseen liittyy haasteita koskien tarvetta organisoida, ylläpitää ja mahdollistaa saavutettavuus tieteelliseen dataan sekä datan uudelleenkäytön tukemista (Zimmerman 2007, 6).

Digitaaliset kirjastot eroavat data-arkistoista muun muassa siinä, että digitaalisilla kirjastoilla on taipumus painottaa metadatasisältöä, kun taas data-arkistot painottavat enemmin datasisältöä (Helly et al. 2002, 99). Digitaaliset kirjastot jakavat näin ollen ensisijaisesti informaatiota erilaisten dataresurssien olemassaolosta, kun taas arkistot jakavat itse dataa.

Esimerkiksi eräässä yhdysvaltalaisessa kansallisen tiedesäätiön (National Science Foundation, NSF) rahoittamassa projektissa pyritään rakentamaan alakohtaista digitaalikirjastoa (Metadata++) luonnonvarojen hallinnalle. Tämä digitaalinen kirjasto sisältää yleisluonteisen tesaurusmallin ohjelmistoineen ja yksi sen perustavoitteista on täsmällisesti esittää ja hyödyntää monia kontrolloituja sanastoja luonnonvarojen hallinnan alan eri diskursseista. Nämä ammattilaiset, ekologit mukaan lukien, käyttävät usein samoja termejä mutta monesti hieman ja joskus huomattavankin paljon eri merkityksissä. (Delcambre et al. 2005)

5.2.3 Dokumentointi

Ekologisen datan jakamisesta arkistoinnin keinoin ei ole datan uudelleenkäyttöä ajatellen minkäänlaista hyötyä, mikäli dataa ei ole huolellisesti ja yksiselitteisesti dokumentoitu ja tätä dokumentaatiota, eli metadataa, liitetty datan mukana arkistoon. Metadata on toisin sanoen elintärkeä elementti ekologisen datan uudelleenkäytön mahdollistamisessa. Metadataa voidaan pitää ratkaisevana keinona ekologisen datan paikantamiseen, saavutettavuuteen, tulkintaan ja analysointiin liittyvissä haasteissa (NCEAS 2008). Metadatan käsite kuuluukin olennaisena osana ekoinformatiikka-artikkeleihin, mutta myös täysin ekologiseen metadataan keskittyviä artikkeleita on laadittu huomattavasti runsaammin kuin esimerkiksi ekologisen datan arkistointia käsitteleviä.

Metadata on datan tulkintaan ja käyttöön tarvittavaa informaatiota ja kattavan metadatan liittäminen dataan data-arkistossa yleensä lisää datatiedoston arvoa ajan kuluessa (Jones et al. 2006, 524; Michener et al. 2002, 3). Metadatan tulisi sisältää informaatiota datatiedoston sisällöstä, laadusta, rakenteesta ja saavutettavuudesta (Michener et al. 1997, 331). Ekologisen metadatan tulee myös kertoa kuka, koska, missä, milloin, miksi ja miten ekologinen data on kerätty (Fegraus et al. 2005, 159).

Lähtökohta on se, että kaikki ekologinen data tarvitsee metadataa, sillä yksikään datakokoelma ei ole täydellinen eikä itse itsensä selittävä (Michener et al. 1997, 330). Se kuinka paljon ja minkälaista metadataa dataan liitetään vaikuttaa puolestaan datan uudelleenkäyttöön siten, että mitä kattavampi metadata on, sitä kauemmin ja laajemmin data on uudelleenkäytettävissä (Michener et al. 1997, 339). Keskeinen kysymys

ekoinformatiikassa metadataan liittyen onkin, että kuinka paljon metadataa vähintään tarvitaan, jotta alun perin tiettyä suhteellisen rajallista tarkoitusta varten kerättyä dataa on mahdollista ymmärtää ja käyttää tarkoituksenmukaisesti moniin muihin tarkoituksiin? (Andelman et al. 2004, 244). Kysymys on tärkeä siksi, että kaikkea ekologiseen dataan liittyvää informaatiota ei ole mahdollista taltioida (Bowker 2000, 675). Kysymykseen metadatan minimimäärästä ei kuitenkaan ole löydettävissä yleispätevää vastausta, sillä metadatan riittävyys on aina tapauskohtaista ja sidoksissa datan erilaisiin käyttäjiin ja käyttötarkoituksiin (Michener et al. 1997, 335).

Metadataa tarvitaan myös muun muassa hidastamaan niin sanottua informaatioentropiaa, eli dataan liittyvän informaatiosisällön turmeltumista ajan myötä. Informaatioentropiassa tyypillisesti yksityiskohtaisimmat erityistiedot menetetään ensin ja tämän jälkeen yleisempi informaatiosisältö. (Michener et al. 2002, 3) Informaatioentropiaan voivat johtaa muun muassa datan tallennusvälineen tuhoutuminen, datan keränneen henkilön työpaikanvaihdos, eläkkeellesiirtyminen tai kuolema sekä datan analysointi- ja tallennusteknologian vanheneminen (Michener et al. 1997, 331). Informaatioentropiaa voi periaatteessa tapahtua missä tahansa datanhallinnan vaiheessa, esimerkiksi jo datan keräämisen ja analysoinnin yhteydessä, mutta informaation häviämismahdollisuuden todennäköisyys kasvaa merkittävästi tutkimustulosten julkaisemisen ja / tai tutkimusprojektin päättymisen jälkeen (Michener 2000b, 94).

Tieteen näkökulmasta metadatan tehtävän voidaan tulkita kehittyneen pelkästä datan löytämisen tukemisesta ihmisvoimin suoritettavan datan hankinnan, tulkinnan ja käytön helpottamiseen ja tästä edelleen viime aikoina vähitellen myös automaattisen

datan löytämisen, sulattamisen, käsittelyn ja analysoinnin mahdollistamiseen metadata-avusteisissa tieteellisissä työnkulkujärjestelmissä (Michener 2006, 3).

Datan dokumentoinnin tärkeys ekologisen tutkimuksen helpottamisessa on tiedostettu 1980-luvulta lähtien (Gross et al. 1995, 6) ja ekologinen tiedeyhteisö on laatinut metadataa osana arkistointiprosessia yli vuosikymmenen ajan (McCartney & Jones 2002, 379). Nykyään ekologisen tutkimusprojektin rahoittaja- tai muu organisatorinen taho monesti vaatii tai ainakin suosittelee perusteellisen metadatan tuottamista tutkimusdatalle (Michener 2000b, 95).

5.2.3.1 Metadatastandardit

Metadatalla voidaan väljästi ajatellen viitata mihin tahansa informaatioon, joka tuottaa lisäarvoa datan tulkintaan, mutta käytännössä termi tavallisesti assosioituu systemaattiseen datatiedoston kriittisten aspektien dokumentointiin tarkoitettuihin rakenteellisiin ja hyvin määriteltyihin kategorioihin. Johdonmukaista ja tarkkaa määritelmäkokonaisuutta metadatakategorioille kutsutaan metadatan sisältö-spesifikaatioksi, joka puolestaan muuntuu metadatastandardiksi jonkin yhteisön laajan hyväksynnän ja omaksumisen myötä. (Jones et al. 2006, 525) Jokaisen kehitetyn metadatastandardin todellinen testi tulee olemaan siinä onko standardi yksinkertainen käyttää ja ymmärtää ja parantaako se tieteentekoa (Michener 2000b, 112).

Ekologisen datan dokumentointiin voidaan käyttää useita metadatan sisältö-spesifikaatioita tai standardeja. On sanottu, että ekologisen datan kompleksisesta luonteesta johtuen on epätodennäköistä, että yhden standardin avulla kyettäisiin

kuvailemaan kaikenlainen ekologinen data (Michener 1998, 48). Monien eri metadatastandardien käytöstä aiheutuu kuitenkin yhteentoimivuusongelmia datan tuottajien halutessa saada datansa saavutettavaksi eri data-arkistoista (Jones et al. 2006, 525-526).

Ekologian alalle ensimmäisenä sovelletut metadatastandardit on kehitetty spatiaaliselle datalle. Esimerkiksi FGDC (Federal Geographic Data Committee) kehitti sisältö-standardin digitaaliselle geospatiaaliselle metadatalle vuonna 1994. Standardia on sovellettu myös ekologiselle geospatiaaliselle datalle. Sittemmin standardia on muokattu ja sen rinnalle on kehitetty oma versio biologiselle ei-geospatiaaliselle datalle. Standardi on nimetty biologiseksi dataprofiiliksi (Biological Data Profile). (FGDC 1999, 1-2)

Eniten tutkimusaineistossa puhutaan ekologisesta metadatakielestä (Ecological Metadata Language, EML), joka on ekologian alalla kehitetty ja alun alkaen nimenomaan ekologian alalle tarkoitettu metadatastandardi. EML perustuu Amerikan ekologisen yhdistyksen (ESA) sekä Michener et al.:n (1997) tekemälle esityölle (KNB). Standardin kuvailuelementtien esikuvana on osittain ollut Dublin Core metadatastandardi (McCartney & Jones 2002, 381) ja osittain kehitystyön pohjana ovat olleet viisi dataa uudelleenkäyttävän ekologin mieleen oletettavasti nousevaa kysymystä. Kysymykset koskevat relevantin datan olemassaoloa, saavutettavuutta, keräämisen syitä ja soveltuvuutta omaan tarkoitukseen, strukturointitapaa sekä lisäinformaation saamista käytön ja tulkinnan helpottamiseen (Michener 200b, 101). Rakenteellisesti EML koostuu moduuleista, joista jokaisen on tarkoitus kuvailla yhtä

loogista osaa siitä metadatakokonaisuudesta, joka tulisi sisällyttää jokaiseen arkistoitavaan ekologiseen datasarjaan (Michener et al. 2002, 3).

Fegraus et al. (2005) painottavat, että ilman tällaista EML:n kaltaista yleisesti käytettävää metadata-ohjeistusta, ekologit liittävät dataansa juuri sen verran ja sen kaltaista informaatiota kuin itse kulloinkin päättävät. Tästä syystä metadatan muoto ja sisältö usein vaihtelee yksittäisen tutkijankin tiedostosta toiseen ja ongelma luonnollisesti paisuu haluttaessa yhdistää eri tutkijoiden dokumentoimia datatiedostoja. EML:n avulla yritetäänkin vähentää ekologisen datan moniselitteisyyttä ja tulkinnallista epävarmuutta vakiinnuttamalla tietyt metadatakäsitteet kattavaksi ja standardoiduksi erityisesti ekologiselle datalle tarpeellisten termien ja määritelmien joukoksi. (Fegraus et al. 2005, 159-160)

Vaikka EML tuottaa johdonmukaisen syntaksin datakokoelmien paikantamiselle, se ei McCartneyn ja Jonesin (2002, 383) mukaan vielä mahda juuri mitään integroinnillekin ongelmia aiheuttaville datakokoelmien välisille semanttisille eroille. He povaavatkin metadatatutkimuksen uusien tavoitteiden suuntautumista integrointitutkimuksen tapaan ontologiaperusteisiin ratkaisuihin.

5.2.3.2 Metadatan tuottaminen ja hallinta

Metadatan tallennusvälineen ja rakenteen määräävät Michenerin (2000b, 106) mukaan käytännössä usein metadatan tuottamisvälineiden ja koulutetun henkilökunnan saatavuus, aika- ja rahoitusrajoitukset sekä suunniteltu metadatan käyttöaste. Mikäli erityisesti metadataa varten tarkoitettuja välineitä ei ole saatavilla tai ne ovat

riittämättömiä, saatetaan metadata sisällyttää tekstinkäsittelytiedostoihin, analyyttisiin ohjelmiin tai rakenteellisempiin tietokantahallintajärjestelmän (DBMS) ohjelmiin. Michener nostaakin metadatan toteuttamista helpottavat joustavat metadatan tuottamis- ja hallintavälineet keskeiseksi tutkimusta ja kehitystyötä vaativaksi alueeksi.

Ekologisen metadatakielen käyttäminen ekologisen datan dokumentointiin on mahdollista ainakin kahdella tapaa. Ensimmäinen näistä on käyttää datan- ja metadatanhallintaohjelmistoa nimeltä Morpho. Sen lisäksi, että Morpho auttaa ekologeja luomaan, editoimaan ja hallitsemaan metadataa ja datataulukoitaan, sen avulla on myös mahdollista etsiä ja tehdä hakuja EML:ään pohjautuvista yleisistä ekologisista data-arkistoista. Toinen keino on käyttää jonkin instituution perustamaa verkossa toimivaa data-arkistoa, johon rekisteröidyttyään ekologit voivat luoda EML:n mukaista metadataa joutumatta asentamaan ja opettelemaan Morpho-ohjelmaa. Tällä hetkellä nämä verkkoarkistot mahdollistavat metadatan luomisen sekä metadata-hakujen tekemisen internetissä, mutta eivät mahdollista suoraa pääsyä dataan, ellei metadata sisällä online-sijaintitietoa (Fegraus et al. 2005).

Ekologit ja datanhallintohenkilöstö ovat yhtä mieltä siitä, että kukaan ei ymmärrä dataa paremmin kuin se, joka sen on kerännyt ja siten juuri tämän henkilön tulisi dokumentoida kyseinen datakokoelma (Zimmerman 2003, 224). Tämä ei kuitenkaan ole ekologian alalla kulttuurinen normi eikä vallitseva käytäntö (Karasti et al. 2006, 335). Paras tapa dokumentointitehtävän helpottamiseksi on kehittää välineitä, jotka auttavat ekologeja tarjoamaan juuri sen tietämyksen, jonka vain he kykenevät antamaan, eikä pyytää heitä dokumentoimaan dataansa kaikille mahdollisille käyttäjille ja kaikkia mahdollisia käyttötarkoituksia varten (Zimmerman 2003, 224).

Useimmilla ekologeilla on ollut joskus vaikeuksia muistaa tärkeitä yksityiskohtia omasta datastaan jopa vain muutaman kuukauden kuluttua datan keräämisestä (Fegraus et al. 2005, 159). Dokumentointi tulisikin mieluiten tehdä mahdollisimman pian keräämisen jälkeen, jolloin datan muodostumiseen vaikuttaneet tekijät ovat vielä tuoreessa muistissa, oikeastaan siis jo datanhallinnan I vaiheessa. Koska dokumentointi kuitenkin on nimenomaan datan uudelleenkäytön mahdollistamiseksi tehty toimi, on se tässä sijoitettu vasta datanhallinnan II vaiheeseen.

5.2.4 Laadunvarmistus ja -valvonta

Vaikka datan laadun tärkeyttä ekologisissa tutkimuksissa ei voi ylikorostaa, saa datan laadunvarmistus yllättävän vähän huomiota tutkimuskirjallisuudessa. Laadunvarmistus- ja -valvontatoimenpiteiden tarkoituksena on estää tai havaita datan kontaminoituminen. Datan kontaminoitumiseksi kutsutaan tilannetta, jossa jokin muu kuin kiinnostuksen kohteena oleva prosessi tai ilmiö vaikuttaa muuttujan arvoon. Kontaminoitumista voivat aiheuttaa esimerkiksi datan tietokonejärjestelmään syöttämisessä tapahtuneet virheet sekä mittalaitteiden väärinlukemisesta tai havaintojen väärinkirjaamisesta aiheutuvat virheet. Ekologisen datan laadunvarmistus- ja -valvontakeinoihin kuuluu standardien määrittely ja toteuttaminen käytettäville formaateille, koodeille, mittayksiköille ja metadatalle; epätavallisten ja mahdottomien datamallien tarkkailu; arvojen vertailu datatiedostojen välillä sekä yleinen laadunseuranta. (Edwards 2000, 70; Brunt 2000, 36-37) Datatiedostoihin datanhallinnan eri vaiheissa tehdyt laadunvarmistus- ja –valvontatoimenpiteet tulisi aina raportoida metadatassa (Brown 1994, 24).

Eräs suuremmissa projekteissa toimiva hallinnallinen keino tieteellisen datan laadunvarmistuksen helpottamiseksi on Edwardsin (2000, 71) mukaan niin sanotut laatupiirit. Nämä ovat lyhyitä, säännöllisiä tapaamisia, joissa tutkijat, teknikot, järjestelmäasiantuntijat ja dataa syöttävä henkilöstö keskustelevat datan laatuun liittyvistä asioista. Näin kaikki ammattiryhmät olisivat tietoisia datan laatutekijöistä ja osaisivat ennakoida mahdollisia ongelmia sekä arvostaa paremmin omaa rooliaan datan laadun ja koko tieteellisen hankkeen suhteen.

Yksi mahdollinen ratkaisu datan laadun varmistamiseen nimenomaan datanhallinnan tässä vaiheessa on omaksua jaettavaksi annettavan tieteellisen datan suhteen vastaavanlainen vertaisarviointikäytäntö kuin tieteellisten artikkelien julkaisemisessa (Helly et al. 2002, 100). Tällainen käytäntö on toiminnassa jo amerikkalaisen ekologisen yhdistyksen (ESA) lehdissä julkaistuihin artikkeleihin liittyvälle tutkimusdatalle. Datan vertaisarviointiprosessiin sisältyy tekninen arviointi, jossa varmistetaan, että data on järjestetty loogisesti ja yhdenmukaisesti, että metadata on riittävän kattava ja että datan laadun ja eheyden ylläpitämiseksi on tehty asianmukaiset toimenpiteet (Helly et al. 2002, 100).

5.3 Datanhallinnan III vaihe

Ekologisen datan hallinnan kolmanteen vaiheeseen kuuluu datan uudelleenkäyttö ja sen tukeminen. Datan jakamiseen, arkistointiin ja dokumentointiin nähty vaiva muuttuu kannattavaksi vasta mikäli data tulee näiden toimien ansioista myöhemmin uudelleenkäytetyksi. Termillä uudelleenkäyttö tai sekundaarikäyttö tarkoitetaan tässä

Zimmermanin (2003, 7) tavoin yhtä tarkoitusta varten kerätyn datan käyttöä uuden ekologisen ongelman tutkimisessa. Ekologista dataa voivat toki haluta käyttää alkuperäisten tutkijoiden jälkeen myös muut ryhmät kuin tutkijat, esimerkiksi erilaiset yhteiskunnalliset päätöksentekijät tai tiedotusvälineiden edustajat. Suurin osa tähänastisesta ekoinformatiikkatutkimuksesta on kuitenkin tähdännyt nimenomaan ekologisissa tutkimuksissa tarvittavan datan uudelleenkäyttömahdollisuuksien parantamiseen (Cushing & Wilson 2005, 4). Tarve monien eri käyttäjäryhmien huomioonottamiselle on silti viime kädessä yksi keskeinen tulevaisuuden haaste ekologisen datan uudelleenkäytön kehittämisessä.

Uudelleenkäyttö-termi mainitaan kaikissa yhteyksissä, joissa ekoinformatiikkakin. Siihenhän ekoinformatiikkatutkimus suurelta osin tähtää. Ekologisen datan uudelleenkäytöstä on kuitenkin vasta vähän empiiristä tutkimustietoa. Zimmerman (2003, 73) huomauttaakin varsin ainutlaatuisessa ekologien datan uudelleenkäyttöä valottaneessa tutkimuksessaan, että tutkijoiden motivaatioita datan etsintään, olemassa olevien formaalien ja epäformaalien datan löytämistä tukevien mekanismien arviointiin, sekä sekundaarikäyttäjien kokemuksiin, kohdistetuista tutkimuksista on pulaa.

Mikäli dataa aiotaan käyttää uudelleen, sen tulee olla tarjolla muodossa, joka on saavutettavissa, vaihdettavissa ja manipuloitavissa, ja datan tulee olla kuvailtu ja dokumentoitu siten, että sekundaarikäyttäjä pystyy ymmärtämään sitä (Zimmerman 2003, 31). Ekologisen datan uudelleenkäyttöä edeltää uudelleenkäyttöön tarvittavan datan löytäminen, hankinta, tulkinta ja laadun arviointi. Näitä käsitellään seuraavassa

enemmän, aloittaen datan hankintaan liittyvistä teemoista, siirtyen tämän jälkeen hankitun datan käyttöä edeltäviin toimenpiteisiin.

5.3.1 Datan hankinta

Ekologit hankkivat uudelleenkäytettäväksi monenlaista dataa monista eri lähteistä, sekä elektronisessa että painetussa muodossa (Zimmerman 2003, 141-142). Koska ekologeilla ei juurikaan ole käytettävissä kattavia tietokantoja ekologisen datan hankintaan uudelleenkäyttöä varten, täytyy heidän rajata datanhankintansa muilla tavoin (Zimmerman 2003, 185). Muita relevantin data-aineiston keräämiseen käytettyjä keinoja ovat ainakin kirjallisiin lähteisiin tai maantieteelliseen sijaintiin perustuvat rajaukset. (Zimmerman 2003, 213)

Datanhallinnan III vaiheeseen kuuluva datan saavutettavuus (access) on nostettu erittäin keskeisenä ongelmana usein esille tutkimusaineistossa. Ekologisen datan saavutettavuuden suhteen tilanne on se, että suurin osa olemassa olevasta ekologisesta datasta on käytännöllisesti katsoen yksittäisten tutkijoiden tai tutkimusprojektien tiedossa ja muille saavuttamattomissa (Jones et al. 2006, 538).

Tutkijoiden kiinnostuksen kohdistuessa paikallisen datakeräämishankkeen ulkopuolelle, vaikeutuu datan saavutettavuus esimerkiksi erilaisista yhteiskunnallisista ja oikeudellisista syistä, kuten luvan puuttumisesta datan käyttöön, kuin myös teknologisista syistä, kuten integrointivaikeuksista tai siitä, että data sisältyy oman järjestelmän kanssa yhtyeensopimattomaan hallintajärjestelmään. Haastavaksi ekologisen datan saavutettavuuden tekee lisäksi ekologisen datan luontainen

monimuotoisuus sekä se, että suuri osa ekologisesta datasta on edelleen digitoimatta (Jones et al. 2006, 521). Ekologinen data on siis tyypillisesti talletettu paikalliseen arkistojärjestelmään ja sen hallintaan ja dokumentointiin käytetyt menetelmät ovat usein varsin omalaatuisia, mikä tekee sen paikantamisesta ja hankinnasta haastavaa. Sekundaarikäyttöön tarjolla olevan datan määrä tulee Zimmermanin (2007, 6-7) mukaan kuitenkin jatkuvasti kasvamaan datan keruun teknologistumisen, jakamiseen ja uudelleenkäyttöön kannustavien periaatteiden sekä laajamittaisten yhteistyö-projektien lisääntymisen myötä.

Ekologian alalla vallitsevasta linjasta poiketen LTER-verkostolla on laadittuna viralliset datan saavutettavuusperiaatteet, joiden pääasiallisena tarkoituksena on edistää ekologisen datan maksimaalista saatavuutta. Periaatteiden mukaisesti datan tulee pääsääntöisesti olla julkisesti saatavilla viimeistään kahden vuoden kuluttua sen keräämisestä (LTER 2005b). Käytäntö on ollut olemassa LTER-verkostossa jo 1990-luvun puolivälistä lähtien (Karasti & Baker 2004, 2).

Datan hankintaan uudelleenkäyttöä varten kuuluu olennaisena teemana myös löytäminen. Zimmermanin (2003, 143) tutkimuksen mukaan ekologit käyttävät sekundaarikäyttöön soveltuvan datan löytämiseen henkilökohtaisia verkostojaan, bibliografisia tietokantoja, julkaistua kirjallisuutta sekä artikkeleiden lähdeviitteitä. Lisäksi he ottavat myös yhteyttä tuntemattomiin tutkijoihin ja tutkimuslaitoksiin, jotka he ovat tunnistaneet mahdollisiksi datalähteiksi sekä käyttävät internetiä.

Ekologisen datan löytämiseen elektronisista lähteistä liittyvistä ongelmista ja yrityksestä ratkaista niitä voisi mainita LTER-verkostoon liittyvän esimerkin.

LTER-verkostoon kuuluvilla tutkimusalueilla käytetään yleisesti datatiedoston kuvailuun asiasanoja, joita ei Porterin (2006) mukaan kuitenkaan ole useimmilla alueilla vielä mitenkään kontrolloitu, vaan ne ovat datan tuottajan itsensä valitsemia. Asiasanoina käytettyjen sanojen ja monisanaisten termien joukko onkin erittäin kirjava ja yli puolta niistä on käytetty kuvaamaan yhtä ainoaa tapausta. Tämä asiasanojen vaihteleva käyttö on yksi LTER:n omien sekä muiden tutkijoiden kohtaamista haasteista datan löytämisessä. Esimerkiksi haluttaessa löytää dataa hiilidioksidi-mittauksista, täytyy etsiä sekä sanalla ´hiilidioksidi´ että kemiallisella lyhenteellä ´CO2´. Toisaalta haulla ´kaasut´ ei löydä kummallakaan edellä mainitulla asiasanalla kuvailtuja datatiedostoja. Smith et al. (2002) puolestaan ovat kehitelleet erityisesti ekologiselle datalle soveltuvaa hakujärjestelmää, joka luo haettavissa oleville datatiedostoille dynaamisesti omat ennalta määrätyt käyttöliittymänsä.

5.3.2 Datan tulkinta ja laadunarviointi

Mikäli ekologit eivät pysty selvittämään sitä prosessia, jonka tuloksena tarjolla oleva data on syntynyt, on Zimmermanin (2003, 220) mukaan epätodennäköistä, että he käyttävät kyseistä dataa omissa tutkimuksissaan. Ekologisten kokeiden ja havaintojen tekotapojen ymmärtäminen on tärkeää ennen kaikkea siksi, että tiedemaailmassa yleinen kokeiden toistaminen datan laadunvarmistukseksi ei ole ekologisen datan kohdalla monestikaan mahdollista eikä näin ollen yleinen käytäntö ekologiassa (Zimmerman 2003, 136-137). Esimerkiksi ekologisista tutkimuksista kertovat julkaisut eivät kuitenkaan yleensä sisällä riittävästi informaatiota siitä, kuinka data on kerätty (Gross et al. 1995, 7).

Datan tulkinta sisältää arvioinnin siitä, vastaako data omia tarpeita ja onko se käyttökelpoista omassa tutkimuksessa. Datan kyllin hyvään ymmärtämiseen ennen kuin sitä on mahdollista käyttää kuluu paljon aikaa. (Zimmerman 2003. 183) Erityisesti toisten alojen edustajien keräämän datan tulkintaa ja arviointia soveltuvuudesta omaan tutkimukseen vaikeuttaa muun muassa se, että tämä data on usein kerätty hyvin erilaiset tavoitteet omaavien projektien yhteydessä (Michener 2000c, 3). Ekologisten tutkimusten tulkintaa helpottaisi osaltaan standardoitujen tutkimusmenetelmien käyttäminen (Michener 2000c, 16).

Menetelmällisen kirjavuuden lisäksi ekologiassa ei ole myöskään aina käytettävissä samalla lailla selkeästi määriteltyjä käsitteitä kuin esimerkiksi elotonta luontoa tutkivissa tieteissä (kuten fysiikka ja kemia), joten yleisesti ymmärrettäväksi tarkoitettua dataa saatetaan joutua luokittelemaan ja kuvailemaan termeillä, joiden määritelmistä ei alalla ole täyttä yhteisymmärrystä (NRC 1997). Tämä on standardoimattomien menetelmien käytön ohella toinen uudelleenkäytettävän datan tulkintaa hankaloittava tekijä. Andelman et al. (2004, 245) ovat esittäneet asiasta oivallisen esimerkin. Heidän mukaansa termiä tutkimuspaikka (site) saatetaan toisessa tutkimuksessa käyttää viittaamaan koko tutkimusalueeseen ja toisessa tutkimuksessa taasen yksittäiseen näytteenottokohtaan tutkimusalueella. Termin tarkoittaman alueen laajuuskin vaihtelee siten suuresti.

Suurimpana esteenä ekologisen datan uudelleenkäytölle pidetään riittämätöntä dokumentointia. Metadatan kattavuutta on kyllä pyritty parantamaan standardeja luomalla, mutta metadatan käytännön tehokkuudesta elektronisten resurssien

paikantamisen, arvioinnin ja käytön avustamisessa tiedetään vasta vähän (Zimmerman 2003, 90 ja 92).

Löydetyn datan sisällöllisen tulkinnan lisäksi sekundaarikäyttäjän tulisi myös kyetä arvioimaan tarjotun datan laatu. Julkaisemattoman datan osalta datan laadusta ei ole mitään varmuutta. Julkaistuksi hyväksytty data puolestaan sisältää jonkinlaisen todistuksen siitä, ettei siinä ole ainakaan kovin räikeitä virheitä (NRC 1997). Ekologit ymmärtävät, että tutkimusartikkeleiden julkaisemiseen liittyvä vertaisarviointi ei vielä riitä takaamaan datan laatua. Julkaisuja käytetäänkin pikemminkin keinona löytää dataa ja rajaamaan datan keräämistä kuin datan laadunvarmistuskeinona. Ekologit ovat näin ollen valmiita käyttämään myös julkaisematonta dataa, mikäli he löytävät muita tapoja keräämisensä rajaamiseen ja mikäli he voivat saada datan ymmärtämiseen tarvittavan informaation muualta. Kerättäessä pieniä määriä dataa useasta eri lähteestä, vähenee huoli yksittäisen datatiedoston laadusta ja myös ero julkaistun ja julkaisemattoman datan välillä muuttuu merkityksettömämmäksi. (Zimmerman 2003, 225)

Myöskään standardoitujen menetelmien käyttö sinänsä ei ole riittävä tae datan laadusta. Standardoidut menetelmät antavat kyllä aavistuksen siitä, kuinka data on kerätty, mutta eivät kerro, onko kyseiset mittaukset tai havainnot tehty taidokkaasti (Zimmerman 2003, 212).

Ekologien päätökseen siitä, mikä data uudelleenkäyttöön lopulta tulee valituksi, vaikuttaa Zimmermanin (2003, 160/210/221) tutkimustulosten mukaan formaalin alakohtaisen tietämyksen, henkilökohtaisen epävarmuuden sietokyvyn ja informaalin

henkilökohtaisen tietämyksen yhdistelmä. Formaalia tietämystä käytetään datan uudelleenkäyttövaiheessa sekä datan ymmärtämiseen että laadunarviointiin. Informaali henkilökohtainen tietämys puolestaan vaikuttaa esimerkiksi siten, että luottamus datan kerääjän taitoihin tai arvostus häntä kohtaan voi vähentää huolta datan laadusta ja toisaalta henkilökohtaiset kontaktit voivat lisätä datan ja sen tulkintaan tarvittavan informaation saavutettavuutta. Henkilökohtainen epävarmuuden sietokyky taasen on tarpeen, jotta toisten keräämää dataa ylipäätään pystytään käyttämään, koska tällaiseen dataan voi olla hyvin vaikea luottaa.

Yleinen trendi näyttäisi olevan, että uudelleenkäytettäväksi valitaan useimmiten itselle aiemmista tutkimuksista tuttua dataa (Zimmerman 2003, 221), koska sen ymmärtäminen ja laadunarviointi on omien käytännön kokemusten pohjalta helpointa. Ekologien omat kokemukset datan kerääjinä nimittäin antavat heille asiantuntemuksen ymmärtää ratkaiseva yhteys tutkimustarkoituksen, menetelmien ja datan välillä, tunnistaa erityisiin datatyyppeihin liittyvät rajoitukset sekä pärjätä datan kompleksisuuden kanssa (Zimmerman 2003, 147).

Kuten aiemmin on tullut ilmi, on ekologisen datan uudelleenkäytön mahdollistaminen yksi ekoinformatiikan keskeisistä pyrkimyksistä. Sim et al. (2004) ovat tunnistaneet ekologisen datan uudelleenkäytön edistämiseen kaksi lähestymistapaa. Ensimmäinen tapa edellyttää sosiaalista kanssakäymistä. Siinä käytetään henkilökohtaisia suhteita ja tietämyksen siirtämistä epävirallisissa ympäristöissä auttamaan datan tulkinnassa ja laadunvarmistuksessa. Toinen tapa tukeutuu tekniseen infrastruktuuriin datan paikantamisen, hakemisen ja yhdistelemisen helpottamiseksi. Nämä ratkaisu-ehdotukset osoittavat osaltaan kuinka monialaista yhteistyötä ekoinformatiikassa

tarvitaan. Teknologisten apukeinojen soveltaminene ekologiaan ei yksin riitä vaan myös esimerkiksi informaatiotutkimukselle ja sosiaalitieteille ominaisia lähetysmis-tapoja kaivataan datanhallinnallisten päämäärien saavuttamiseksi.

6 JOHTOPÄÄTÖKSET

Tutkielman tarkoituksena oli tutkimuskirjallisuuden sisällönerittelyn avulla selvittää mitä ekoinformatiikka on. Tutkielmassa kartoitettiin ekoinformatiikasta esitettyjä määritelmiä, alaan kuuluvia osa-alueita, tavoitteita ja haasteita. Samalla pyrittiin selvittämään ekoinformatiikan tarjoamia tutkimusaiheita nimenomaan informaatio-tutkijoille. Sisällönerittely osoittautui hyväksi keinoksi näiden keskeisten piirteiden kartoitustyössä.

Ennakkokäsityksen mukaisesti ekoinformatiikasta saatavilla oleva informaatio on erittäin hajanaista, yksittäisiin tutkimusprojekteihin keskittyvää, eikä suomenkielistä aineistoa ollut löydettävissä. Termiä ekoinformatiikka ei suomenkielessä vielä yleisesti käytetä. Ekoinformatiikka on tutkimusalanakin nuori ja hakee vielä rajojaan. Tämä näkyy muun muassa alasta esitettyjen määritelmien kirjavuutena ja englanninkielistenkin perusartikkeleiden puuttumisena. Tässä tutkielmassa ekoinformatiikka määriteltiin tieteidenväliseksi tutkimusalaksi, joka tähtää ekologisen datan ja informaation hallinnan edistämiseen.

Ekoinformatiikassa on tutkimusaineiston perusteella nähtävissä kolme perusulottuvuutta: ekologia, teknologia ja informaatio. Ekoinformatiikan syntyyn ovat vaikuttaneet toisaalta ekologisessa tutkimuksessa tapahtuvat muutokset ja toisaalta teknologian huima kehittyminen, jotka ovat sekä lisänneet että monimuotoistaneet ekologisissa tutkimuksissa hallittavan datan määrää. Tähän mennessä ekoinformatiikkatutkimus on suurelta osin keskittynyt ekologisen ja teknologisen näkökulman risteyskohtaan. Tässä tutkielmassa esitetty näkemys ekoinformatiikasta

puolestaan painottaa informaationäkökulmaa ja datan koko elinkaaren kattavaa hallintaa.

Informaationäkökulmasta ekologia näyttäytyy hyvin dataintensiivisenä alana ja myös ekoinformatiikka on tähän mennessä suurelta osin keskittynyt nimenomaan datanhallinnallisiin haasteisiin. Yhtenä ekoinformatiikan keskeisistä pyrkimyksistä voidaan nähdä datan elinkaaren pidentäminen niin, että alun perin tiettyä tarkoitusta varten kerättyä dataa olisi mahdollista käyttää myös moniin muihin tarkoituksiin pitkienkin aikojen kuluttua alkuperäisen tutkimusprojektin päättymisen jälkeen. Tämän näkemyksen puitteissa ekologiselle datalle esitettiin tutkielmassa kolmivaiheinen datanhallinnan malli, jossa dataa hallitaan ensin (vaihe I) osana sen synnyttänyttä tutkimusta, josta se siirtyy (vaihe II) erilaisten jakamismekanismien avulla uudelleenkäytettäväksi (vaihe III) toisessa tutkimuksessa tai muutoin.

Kaikkiin vaiheisiin liittyy omanlaisiaan haasteita, mutta suuri osa varsinkin Euroopassa tehdystä ekoinformatiikkatutkimuksesta on kohdistunut datanhallinnan ensimmäiseen vaiheeseen, eli ekologisen datan analysoinnin tukemiseen. Merkittävässä roolissa muiden vaiheiden datanhallinnallisten haasteiden ratkaisemisessa on ollut alun perin yhdysvaltalaisen, sittemmin kansainvälistyneen pitkäkestoista ekologista tutkimusta harjoittavan LTER-ohjelman puitteissa tehty tutkimus- ja kehitystyö.

Suurimmat ongelmat ekoinformatiikassa juontavat juurensa ekologian tutkimuskentän laajuudesta ja tutkimuksessa tarvittavan ja siinä syntyvän datan erityisen heterogeenisestä luonteesta. Myös ekologisten tutkimusten pitkäkestoisuus aiheuttaa

merkittäviä haasteita toimivan datanhallinnan tehostamiseksi. Pitkäkestoinen tutkimus kun tarvitsee pitkäkestoista dataa, mutta teknologiassa tapahtuu nopeita muutoksia.

Ekoinformatiikkaa ei olla informaatiotutkijoiden taholta tähän mennessä tutkittu juuri lainkaan. Mahdollisia tutkimusaiheita kuitenkin löytyisi paljon ekoinformatiikan joka saralta. Esimerkiksi ekologisen datan uudelleenkäytöstä tiedetään vasta vähän ja se vaatisi nimenomaan tiedonhallinnallista tutkimusta, jotta esimerkiksi sopivan datan löydettävyyttä saataisiin parannettua. Informaatiotutkijoilla olisi varmasti paljon annettavaa erityisesti myös datan jakamismekanismien kuten arkistoinnin ja digitaalisten kirjastojen kehittämisessä sekä dokumentoinnin parantamisessa. Esimerkiksi informaatiojärjestelmien kehittämisessä ei olla totuttu ajattelemaan pitkäkestoisuuden tarvetta, joka on ollut keskeinen huolenaihe arkistoammattilaisten keskuudessa jo pitkään.

Ylipäätään kuvaa ekoinformatiikasta eri osa-alueiden kohdalta informaatio-näkökulmasta tulisi tarkentaa, sillä tässä tutkielmassa ei ollut mahdollista yleiskuvan tavoittelemisen vuoksi mennä kovin syvälle olemassa oleviin haasteisiin ja niiden ratkaisuyrityksiin. Mielenkiintoista olisi myös selvittää ekologisen datanhallinnan käytäntöjä Suomessa.

LÄHTEET

Altintas, I.; Berkley, C.; Jaeger, E.; Jones, M.; Ludäscher, B. & Mock, S. (2004). Kepler: An extensible system for design and execution of scientific workflows. Teoksessa: Proceedings of the 16th International Conference on Scientific and Statistical Database Management, 423-424. Saatavilla sähköisesti: <http://users.sdsc.edu/~ludaesch/Paper/ssdbm04-kepler.pdf> Viitattu 28.3.2008.

Andelman, S. J.; Bowles, C. M.; Willig, M. R. & Waide, R. B. (2004). Understanding environmental complexity through a distributed knowledge network. BioScience 54 (3) 240-246.

Baker, K. S. & Bowker, G. C. (2007). Information ecology: open system environment for data, memories, and knowing. Journal on Intelligent Information Systems 29 (1) 127-144.

Baker, K. S.; Bowker, G. C. & Karasti, H. (2002). Designing an infrastructure for heterogeneity in ecosystem data, collaborators and organizations. Teoksessa: Proceedings of the 2002 Annual National Conference on Digital Government Research, 1-4. ACM International Conference Proceedings Series 129. Saatavilla sähköisesti: <http://diggov.org/library/library/pdf/baker.pdf> Viitattu 28.3.2008.

Baker, K. S.; Benson, B. J.; Henshaw, D. L.; Blodgett, D.; Porter, J. H. & Stafford, S. G. (2000). Evolution of a multisite network information system: the LTER information management paradigm. BioScience 50 (11) 963-978.

Biodiversity Informatics -lehden esittely.
<https://journals.ku.edu/index.php/jbi/index> Viitattu 2.4.2008

Blair, D. C. (2002). Knowledge management: hype, hope, or help? Journal of the American Society for Information Science and Technology 53 (12) 1019-1028.

Borgman, C. L.; Wallis, J & Enyedy, N. (2007). Little science contronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries 7 (½) 17-30.

Bowker, G. C. (2000). Biodiversity Datadiversity. Social Studies of Science 30/5, 643-83.

Brown, J. H. (1994). Grand challenges in scaling up environmental research. Teoksessa: Michener, Brunt & Stafford (toim.) Environmental Information Management and Analysis: Ecosystem to Global Scales, 21-26.

Brunt, J. W. (2000). Data management principles, implementation and administration. Teoksessa: W. K. Michener & J. Brunt (toim.) Ecological Data: Design, Management and Processing, 25-47.

Brunt, J. W.; McCartney, P.; Baker, K. & Stafford. S. G. (2002). The future of ecoinformatics in long term ecological informatics. Teoksessa: Proceedings of the 6th World Multi-Conference on Systematics, Cybernetics and Informatics (sci2002).

Saatavilla sähköisesti: <http://www.icess.ucsb.edu/lter/biblio/2002/sci2002brunt.pdf> Viitattu 28.3.2008.

Costello, M. J. & Berghe, E. V. (2006). ´Ocean biodiversity informatics´: a new era in marine biology research and management. Marine Ecology Progress Series 316: 203-214.

Cushing, J. B; Nadkarni, N.; Finch, M.; Fiala, A.; Murphy-Hill, E.; Delcambre, L. & Maier, D. (2007). Component-based end-user database design for ecologists. Journal of Intelligent Information Systems 29 (1) 7-24.

Cushing, J. & Wilson, T. (toim.) (2005). Eco-informatics for decision makers: advancing a research agenda. Report of an Workshop on ecoinfromatics for resource management decision makers. Dec 13-15, 2004. Saatavilla sähköisesti: <http://academic.evergreen.edu/projects/bdei/documents/finalReport.pdf> Viitattu 28.3.2008.

Delcambre, L.; Nielsen, M. L.; Tolle, T.; Weaver, M.; Maier, D. & Price, S. (2005). NSF-Project: Harvesting information to sustain our forests. Teoksessa: Proceedings of the 2005 National Conference on Digital Government Research. SESSION: Data integration and eco-informatics, 213-214. ACM International Conference Proceeding Series; vol 89.

Ecoinformatics. Elektroninen lehti ekoinformatiikasta. <http://www.ecoinformatics.net/index.php/ecoinfo/about/editorialPolicies#focusAndScope> Viitattu 28.3.2008.

Edwards, D. (2000). Data quality assurance. Teoksessa: W. K. Michener & J. Brunt (toim.) Ecological Data: Design, Management and Processing, 70-91.

ESA. Ecological Society of America -yhdistyksen www-sivut. <http://www.esa.org> Viitattu 2.4.2008.

FGDC (1999). Federal Geographic Data Committee. Content Standard for Digital Geospatial Metadata. Part 1: Biological Data Profile. FGDC-STD-001.1-1999. Saatavilla sähköisesti: <http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/biometadata/biodatap.pdf> Viitattu 28.3.2008.

Fegraus, E.; Andelman, S. J.; Jones, M. B. & Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. Bulletin of the Ecological Society of America 86 (3), 158-168.

FinLTSER (2007). The Finnish Long-Term Socio-Ecological Research Network -veroston esite. Saatavilla sähköisesti: <http://www.ymparisto.fi/download.asp?contentid=66827&lan=fi> Viitattu 28.3.2008.

Gross, K. L. et al. (1995). Final report of the Ecological Society of America Committee on the Future of Long-Term Ecological Data (FLED). Volume I: Text of the report.

saatavilla sähköisesti:
<http://intranet.lternet.edu/archives/documents/other/fledvol1.pdf> Viitattu 28.3.2008.

Hanski, I.; Lindström, J.; Niemelä, J.; Pietiläinen, H. & Ranta, E. (1998). Ekologia. Juva: WSOY.

He, Shaoyi (2003). Informatics: a brief survey. The Electronic Library 21 (2) 117-122.

Helly, J.; Case, T.; Davis, F.; Levin, S. & Michener, W. (1999). The State of Computational Ecology. San Diego Supercomputer Center. Technical Report. SDSC TR-1999-5.

Helly, J. J.; Elvins, T. T.; Sutton, D.; Martinez, D.; Miller, S. E.; Pickett, S. & Ellison, A. M. (2002). Controlled publication of Digital scientific data. Communications of the ACM 45 (5) 97-101.

Helsingin yliopisto (2005). Bioinformatiikka 2005-2006 -esite. Saatavilla sähköisesti: <http://www.helsinki.fi/bioinfo/Bio-esite_0506.pdf> Viitattu 28.3.2008.

Hobbie, J. E.; Carpenter, S. R.; Grimm, N. B.; Gosz, J. R. & Seastedt, T. R (2003). The US Long Term Ecological Research Program. BioScience 53 (1) 21-32.

Hook, L. A.; Beaty, T. W.; Santhana-Vannen, S.; Baskaran, L. & Cook, R. B. (2007). Best practices for preparing environmental data sets to share and archive. Päivitetty versio ohjeesta: Cook et al. (2001) Best practices for preparing ecological and ground-based data sets to share and archive. Oak Ridge National Laboratory. Saatavilla sähköisesti: <http://daac.ornl.gov/cgi-bin/MDEDIT/bestprac.html>. Viitattu 28.3.2008.

ISEI. International Society of Ecological Informatics -yhdistyksen www-sivut. <http://www.waite.edelaide.edu.au/ISEI/> Viitattu 2.4.2008.

Jones, M. B. (2007). Meta-information systems and ontologies. A special feature from ISEI'06. Ecological Informatics 2 (3) 193-194.

Jones, M. B.; Schildhauer, M. P.; Reichman, O. J. & Bowers, S. (2006). The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere. Annual Review of Ecology, Evolution, and Systematics 37, 519-544.

Jørgensen (2002). How many Eco-sub-disciplines do we need? Teoksessa: 3rd Conference of the International Society for Ecological Informatics - Abstract book, s. 40. Saatavilla sähköisesti: <http://www.isei3.org/ISEI3_abstract_book.pdf> Viitattu 28.3.2008.

Kalra, H. P. S. (2005). Bioinformatics and the library and information science. IFLA Journal 31 (4) 333-341.

Karasti, H. (2007). Steps towards understanding long-term problematics in information infrastructure work – case of long-term ecological research. Teoksessa: Proceedings of the 30th Information Systems Research Seminar in Scandinavia IRIS 2007.

Karasti, H. & Baker, K. (2004). Infrastructuring for the long-term: ecological information management. Teoksessa: Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS). Saatavilla sähköisesti: <http://csdl2.computer.org/comp/proceedings/hicss/2004/2056/01/205610020c.pdf> Viitattu 28.3.2008.

Karasti, H.; Baker, K. & Halkola (2006). Enriching the notion of data curation in e-science: Data management and information infrastructuring in the Long Term Ecological Research (LTER) Network. Computer Supported Cooperative Work 15 (4) 321-358.

Kineman, J. J. & Kumar, K A. (2006). Information as communication: the new ecoinformatics. Teoksessa: Proceedings of the 50th Annual Meeting of the ISSS, s.367. Saatavilla sähköisesti: <http://journals.isss.org/index.php/proceedings50th/article/view/367> Viitattu 28.3.2008.

KNB. The Knowledge Network for Biocomplexity -verkoston www-sivut. <http://knb.ecoinformatics.org> ja <http://knb.ecoinformatics.org/software/eml/> Viitattu 2.4.2008.

Kratz et al. (2006). Toward a global lake ecological observatory network. Teoksessa: Proceedings of the Karelian Institute, University of Joensuu, Finland, Fall, 2006. Saatavilla sähköisesti: <http://grid.cs.binghamton.edu/projects/publications/global-KI06/global-KI06.pdf> Viitattu 28.3.2008.

Kuopion yliopisto. Ympäristötieteen laitos. www-sivut. <http://envi.uku.fi> Viitattu 2.4.2008.

Lin, C.-C.; Porter, J. H. & Lu, S.-S. (2006). A metadata-based framework for multi-lingual ecological information management. Taiwan Journal of Forest Science 21 (3) 377-82.

LTER (2005a). LTER-esite. Celebrating 25 years of excellence in Long-Term Ecological Research. Saatavilla sähköisesti: <http://intranet.lternet.edu/archives/documents/Publications/brochures/LTERNetwork Brochure1.pdf> Viitattu 28.3.2008.

LTER (2005b). LTER network data access policy, data access requirements, and general data use agreement. Saatavilla sähköisesti: <http://www.lternet.edu/data/netpolicy.html> Viitattu 28.3.2008.

McCartney, P. H. & Jones, M. B. (2002). Using XML-encoded metadata as a basis for advanced information systems for ecological research. Teoksessa: Proceedings from the 6th World Multiconference on Systematics, Cybernetics and Informatics, 379-384. Saatavilla sähköisesti: <http://intranet.lternet.edu/archives/documents/presentations/sci2002/sci2002_xml_encodeded_metadata.pdf> Viitattu 28.3.2008.

Mélendez-Colom, E. C. & Baker, K. S. (2002). Common information framework: in practice. Teoksessa: Proceedings from the 6th World Multiconference on Systematics, Cybernetics and Informatics. Saatavilla sähköisesti: <http://oceaninformatics.ucsd.edu/kbaker/docs/02iiis_CIMF_inpractice.pdf> Viitattu 28.3.2008.

Michener, W. K. (1997). Advances in infromation management and metadata development for ecological data. Kalvo-esitys istunnossa nimeltä Long Term Research in Ecology. Cross-Site Collaborations for the Future, Albuquerque, New Mexico. <http://www.lternet.edu/asm/1997/michener/sld005.htm> Viitattu 28.3.2008.

Michener, W. K. (1998). Ecological metadata. Teoksessa: W.K. Michener, J. H. Porter & S. G. Stafford (Eds.) Data and Information Management in the Ecological Sciences: A Resource Guide. LTER Network Office, University of New Mexico, Albuquerque, NM, 47-51.

Michener, W. K. (2000a). Ecological knowledge and future data challenges. Teoksessa: W. K. Michener & J. Brunt (toim.) Ecological Data: Design, Management and Processing, 162-173.

Michener, W. K. (2000b). Metadata. Teoksessa: W. K. Michener & J. Brunt (toim.) Ecological Data: Design, Management and Processing, 92-116.

Michener, W. K. (2000c). Research design: Translating ideas to data. Teoksessa: W. K. Michener & J. Brunt (toim.) Ecological Data: Design, Management and Processing, 1-24.

Michener, W. K. (2000d). Transforming data into information and knowledge. Teoksessa: W. K. Michener & J. Brunt (toim.) Ecological Data: Design, Management and Processing, 142-161.

Michener, W. K. (2006). Meta-information concepts for ecological data management. Ecological Informatics 1:3-7.

Michener, W. K. et al. (2007). A knowledge environment for the biodiversity and ecological sciences. Journal of Intelligent Information Systems 29, 111-126

Michener, W. K. & Brunt, J. W. (toim.) (2000). Ecological data, design, management and processing. Oxford: Blackwell Science.

Michener, W. K.; Brunt, J. W.; Helly, J. J.; Kirchner, T. B. & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. Ecological Applications 7 (1) 330-342.

Michener, W. K.; Brunt, J. W. & Vanderbilt, K. L. (2002). Ecological Informatics: a Long-Term Ecological Research Perspective. Teoksessa: Proceedings from the 6th World Multiconference on Systematics, Cybernetics and Informatics. Saatavilla sähköisesti: <http://intranet.lternet.edu/archives/documents/presentations/sci2002/sci2002_ecological_informatics_lter_perspective.pdf> Viitattu 28.3.2008.

NCEAS (2008). National Center for Ecological Analysis and Synthesis -tutkimuslaitoksen www-sivut. <http://www.nceas.ucsb.edu/ecoinfo> Viitattu 28.3.2008.

NEON. The National Ecological Observatory Network -verkoston www-sivut. <http://www.neoninc.org> Viitattu 2.4.2008.

Neumann, M.; Baumeister, J. & Puppe, F. (2003). ILMAX: A system for managing experience knowledge in a long-term study of stream ecosystem regeneration – an application of ecological informatics. Teoksessa: Proceedings of the First International NAISO Symposium on Information Technologies in Environmental Engineering (ITEE), 2003.

NRC (1997). National Research Council. Bits of power: Issues in global Access to scientific data. Washington, DC: National Academy Press. Saatavilla sähköisesti: <http://www.nap.edu/readingroom/books/BitsOfPower/> Viitattu 28.3.2008.

Olson, Richard J. & McCord, Raymond A. (2000). Archiving Ecological Data and Information. Teoksessa: W. K. Michener & J. Brunt (toim.) Ecological Data: Design, Management and Processing, 118-141.

Parr, C. S. & Cummings, M. P. (2005). Data sharing in ecology and evolution. Trends in Ecology and Evolution 20 (7) 362-363.

PEaCE Lab. Pacific Ecoinformatics and Computational Lab -tutkimuslaitoksen www-sivut. <http://wow.sfsu.edu/index.html> Viitattu 28.3.2008.

Pietilä, V. (1973). Sisällön erittely. Helsinki: Gaudeamus.

Porter, J. H. (2000). Scientific Databases. Teoksessa: W. K. Michener & J. Brunt (toim.) Ecological Data: Design, Management and Processing, 48-69.

Porter, J. (2006). Improving data queries through use of controlled vocabulary. LTER DataBits – Information Management Newsletter of the Long Term Ecological Research Network. Spring 2006.

Porter, J. et al. (2005). Wireless sensor networks for ecology. BioScience 55 (7) 561-572.

Porter, J. H. & Callahan, J. T. (1994). Circumventing a dilemma: Historical approaches to data sharing in ecological research. Teoksessa: Michener, Brunt & Stafford (toim.) Environmental Information Management and Analysis: Ecosystem to Global Scales, 193-202.

Porter, J. H. & Ramsey, K. W. Jr. (2002). Integrating ecological data: tools and techniques. Teoksessa: Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics: Volume VII Information Systems Development II, 396-401. Saatavilla sähköisesti: <http://intranet.lternet.edu/archives/documents/presentations/sci2002/sci2002_integrating_ecological_data.pdf> Viitattu 28.3.2008.

Recknagel, F. (toim.) (2002). Ecological informatics: understanding ecology by biologically-inspired computation. Berlin: Springer.

Recknagel, F. (2006) Editorial. Ecological Informatics 1 (2006) 1.

Roth, W-M & Bowen, G. M. (1999). Digitizing Lizards: The topology of 'vision' in ecological fieldwork. Social Studies of Science 29 (5) 719-64.

Sachs, J. et al. (2006). Using the semantic Web to support ecoinformatics. Teoksessa: Proceedings of the AAAI Fall Symposium on the Semantic Web for Collaborative Knowledge Acquisition. American Association for Artificial Intelligence. Saatavilla sähköisesti: <http://ebiquity.umbc.edu/_file_directory_/papers/297.pdf> Viitattu 28.3.2008.

Salo, M. & Sääksjärvi, I. E. (2007). Tuntematon maa: luonnon monimuotoisuuden käsikirja. Helsinki: Otava.

Schaminée, J. H. J.; Hennekens, S. M. & Ozinga, W. A. (2007). Use of the ecological information system SynBioSys. Journal of Vegetation Science 18: 463-470.

SEEK (2008 / 2004). Science Environment for Ecological Knowledge -projektin www-sivut. <http://seek.ecoinformatics.org/> Päivitetty 11.2.2008 ja <http://seek.ecoinformatics.org/Wiki.jsp?page=Glossary> Päivitetty 21.5.2004. Viitattu 28.3.2008.

Sim, Zimmerman & Nardi (2004). Converging conversations and repositories: bringing archival data to life. The Center for Organizational Research. Working Papers. Saatavilla sähköisesti: <www.cor.web.uci.edu/ufiles/working_papers/ sim-zimmerman-nardi.doc> Viitattu 28.3.2008.

Smith, T. R. (1997). Meta-information in digital libraries. International Journal on Digital Libraries 1 (2) 105-107.

Smith, D. J; Benson, B. J. & Balsinger, D. F. (2002). Designing Web database applications for ecological research. Teoksessa: Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics. Saatavilla sähköisesti: <http://intranet.lternet.edu/archives/documents/presentations/sci2002/sci2002_designing_web_database_applications.pdf> Viitattu 28.3.2008.

SPiRE. Semantic Prototypes in Research Ecoinformatics -projektin www-sivut. <http://spire.umbc.edu/> Viitattu 2.4.2008.

Strebel, D. E.; Meeson, B. W. & Nelson, A. K. (1994). Scientific information systems: a conceptual framework. Teoksessa: Michener, Brunt & Stafford (ed.) Environmental information management and analysis: ecosystem to global scales, 59-85.

Tuomi, I. (1999). Data is more than knowledge - Implications of the reversed knowledge hierarchy for knowledge management and organizational memory. Journal of Management Information Systems Winter 16 (3) 103-117.

Tuomi, J. & Sarajärvi, A. (2002). Laadullinen tutkimus ja sisällönanalyysi. Helsinki: Tammi.

Turun yliopisto. Biologian laitoksen tutkimusprojektit. <http://www.sci.utu.fi/biologia/tutkimus/projektit/> Viitattu 2.4.2008.

Williams, R. J.; Martinez, N. D. & Golbeck, J. (2006). Ontologies for ecoinformatics. Journal of Web Semantics 4: 237-242.

Wilson, P. D. (2007a). Principles of data management. Version 1.2. Saatavilla sähköisesti: <http://www.ecoinformatics.com.au/Features/DataMgt1.2.pdf> Viitattu 28.3.2008.

Wilson, P. D. (2007b). What is ”ecoinformatics”?. Saatavilla sähköisesti: <http://www.ecoinformatics.com.au/Features/WhatIsEcoinfo.pdf> Viitattu 28.3.2008.

Wyoming (2008). Wyoming Geographic Information Science Center. University of Wyoming. What is Ecoinformatics. <http://www.wygisc.uwyo.edu/ecoinformatics/whatis.htm> Viitattu 28.3.2008.

Yu, Y. Y.; Stamberger, J. A.; Manoharan, A. & Paepcke, A. (2006). EcoPod: A mobile tool for community based biodiversity collection building. JCDL ´06. Teoksessa: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries. Saatavilla sähköisesti: <http://portal.acm.org/citation.cfm?id=1141753.1141807> Viitattu 7.4.2008.

Zimmerman, A. S. (2003). Data sharing and secondary use of scientific data: experiences of ecologists. Väitöskirja. Michiganin yliopisto. Saatavilla sähköisesti: <http://deepblue.lib.umich.edu/bitstream/2027.42/39373/2/ann_zimmerman_dissertation_2003.pdf> Viitattu 28.3.2008.

Zimmerman, A. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. International Journal on Digital Libraries 7 (½) 5-16.

Zimmerman, A. & Nardi, B. A. (2006). Whither or whether HCI: requirements analysis for multi-sited, multi-user cybeinfrastructures. Conference on Human Factors in Computing Sytems. CHI ‘06 Extended Abstracts on Human Factors in Computing Systems. Session: Work-in-progress, s.1601-1606. Saatavilla sähköisesti: <http://www-personal.si.umich.edu/~asz/zimmerman_nardi_chi06.pdf> Viitattu 28.3.2008.

Zuur, A. F.; Ieno, E. N. & Smith, G. (2007). Analyzing ecological data. New York: Springer.