Linked Data -palvelu luontohavaintoaineistoille

Mikko Koho

Pro gradu -tutkielma
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 1. helmikuuta 2015

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Mikko Koho

Työn nimi — Arbetets titel — Title

Linked Data -palvelu luontohavaintoaineistoille

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	1. helmikuuta 2015	61 sivua + 7 liitesivua

Tiivistelmä — Referat — Abstract

Biologisten havaintoaineistojen julkaiseminen linkitettynä datana mahdollistaa useiden aineistojen yhdistämisen toisiinsa. Yhdistämällä toisiinsa useita samaan asiaan liittyviä aineistoja, voidaan saavuttaa parempi ymmärrys kiinnostuksen kohteena olevasta ilmiöstä kuin tutkimalla aineistoja erikseen. Näin voidaan mahdollistaa tarkempien päätelmien tekeminen aineistojen pohjalta sekä etsiä odotettuja tai odottamattomia yhteyksiä aineistojen välillä. Linkitetyssä datassa käytetty RDF-tietomalli tuo aineistoihin koneluettavuuden ja helpon tavan viitata kaikkiin aineistojen osiin. Linkitettynä datana julkaistuja aineistoja voidaan helposti rikastaa yhä uusilla aineistoilla.

Tässä tutkielmassa käsitellään Hangon lintuaseman havaintoaineiston sekä Ilmatieteenlaitoksen Hangon Russarön säähavaintoaineiston mallinnusta, käsittelyä ja hyödyntämistä linkitettynä datana. Aineistot on mallinnettu käyttäen RDF Data Cube -sanastoa, joka parantaa aineistojen yhteentoimivuutta. Lintuhavaintoaineistoon on annotoitu lajitietoa käyttäen ontologiaa Suomen linnuista, jota on rikastettu mm. lajien tuntomerkkiontologialla sekä uhanalaisuustiedoilla.

Aineistot on julkaistu Linked Data Finland -alustalla, ja aineistojen välisten yhteyksien hahmottamiseksi on kehitetty visualisointipalvelun prototyyppi. Säätilan tiedetään olevan tärkeimpiä päivittäisen lintumuuton voimakkuuteen vaikuttavia tekijöitä. Visualisointipalvelulla pyritään näyttämään käyttäjälle, miten säätila vaikuttaa lintuhavaintomääriin ja erityisesti havaittuun lintumuuttoon. Aineistojen välisten suhteiden parempi tuntemus mahdollistaa tarkempien päätelmien tekemisen lintuhavaintoaineiston perusteella.

Tutkielmassa esitetyt menetelmät ovat yleistettävissä lintu- ja säähavaintoaineistojen lisäksi muihin rakenteeltaan samankaltaisiin aineistoihin.

ACM Computing Classification System (CCS):
Information systems → Resource Description Framework (RDF)
Human-centered computing → Visualization
Applied computing → Life and medical sciences

Avainsanat — Nyckelord — Keywords
linkitetty data, ontologiat, RDF, RDF Data Cube, SPARQL, visualisointi

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Sisältö

1 Johdanto

Luontohavainnoista koostuvia aineistoja on olemassa pitkältä ajalta. Digitaalisessa muodossa saatavilla olevien luontohavaintoaineistojen määrä kasvaa jatkuvasti [GFH+04]. Datamäärien kasvaessa niiden hyödyntäminen tulee haastavammaksi. Aineistot ovat hajallaan erilaisissa järjestelmissä ilman yhteismitallisuutta. Luontohavaintoaineistojen ytimessä oleva taksonomia eli eliöiden luokittelu on jatkuvassa muutoksessa tuoden oman hankaluutensa aineistojen yhdistämiseen.

Ihmisen toiminnalla on suuri vaikutus luonnonympäristöihin kaikkialla maailmassa. Tämän seurauksena lajien levinneisyydet muuttuvat ja lajeja häviää. Muutosten seuraamisessa olennaisessa osassa ovat laadukkaat pitkäaikaisseurannat. Esimerkiksi kvantitatiivisten lintumuuttoa käsittelevien seuranta-aineistojen perusteella voidaan tehdä päätelmiä lajien kannankehityksestä [Sve78, L+08] ja vuosittaisista pesintöjen onnistumisista [Kje98]. Seurantatutkimusten perusteella tiedetään ilmastonmuutoksen aikaistaneen monien lintulajien kevätmuuttoa [LSB+10].

Hangon lintuasemalta (Halias) on olemassa vuodesta 1979 lähtien kerättyä lintumuuton seuranta-aineistoa, jota on käytetty aktiivisesti tutkimuksissa [Han13a]. Tutkielmassa selvitetään, miten Hangon lintuaseman havaintoaineisto ja yleisemmin biologisista havainnoista koostuva tutkimusaineisto voidaan julkaista linkitettynä datana ja tutkitaan hyötyjä julkaistavan tutkimusaineiston rikastamisesta muilla aiheeseen liittyvillä aineistoilla kuten säädatalla sekä laji- ja tuntomerkkiontologialla ja uhanalaisuustiedoilla. Säätila tiedetään merkittävimmäksi yksittäisten päivien lintumuuttoa selittäväksi tekijäksi [Abl73, ND68, Ale11]. Tutkielmassa käsitellään tilastollisen linkitetyn datan visualisointia ja tutkitaan, voidaanko visualisoinneilla löytää biologian tutkimuksen kannalta kiinnostavia yhteyksiä lintuhavaintoaineiston ja säähavaintoaineiston välillä.

Luontohavaintoaineistoja voidaan esittää digitaalisesti erilaisissa tiedostomuodoissa ja tietojärjestelmissä lukemattomilla eri tavoilla. Yhteistä useimmille datan esitysmuodoille on se, että tarvitaan erikseen dataa kuvailevaa tietoa eli metatietoa, joka kertoo miten dataa tulee tulkita [BHL01]. Tämä rajoittaa datan ymmärtämisen ainoastaan ihmisille sekä järjestelmille, joille

ihminen on eksplisiittisesti määritellyt, miten dataa tulee tulkita. Aineistojen yhdistäminen toisiinsa vaatii yleensä käsityötä. Eräs ratkaisu koneluettavien ja helposti yhdisteltävien aineistojen esittämiseksi on julkaista niitä linkitettynä datana. [BHL01, VAR09].

Linkitettynä datana julkaistuja aineistoja voidaan rikastaa yhdistelemällä niitä toisiinsa. Edellytyksenä on, että aineistoissa viitataan joihinkin samoihin käsitteisiin kuten paikkoihin, henkilöihin tai ajankohtiin tai niiden kuvailuun on käytetty samoja ontologioita. Linkitetty data mahdollistaa uudenlaisia tapoja hyödyntää aineistoja, kuten uuden tiedon päättelemisen annetun tiedon pohjalta ja aineistojen helpon rikastamisen niitä yhdistelemällä [BHL01]. Aineistojen rikastaminen muilla aineistoilla luo mahdollisuuden etsiä odotettuja tai odottamattomia yhteyksiä erilaisten aineistojen ja tietojen välillä. Näin voi aueta mahdollisuuksia uuden tiedon löytämiseen (knowledge discovery) [MR05]. Semanttinen web on visio linkitetyn datan muodostamasta globaalista verkosta [B$^+$06].

Linkitetyn datan kuvailemiseen käytetään ontologioita. Ne kuvaavat jotain osaa todellisuudesta ja koostuvat formaalista tietyn aihealueen yhteisestä käsitteistöstä ja formaaleista suhteista käsitteiden välillä [GOS09]. Biologisia sukulaisuussuhteita kuvaavien taksonomioiden esittäminen ja suhteuttaminen toisiinsa on mahdollista esimerkiksi ontologioiden avulla [TLH11].

Tutkielman tavoitteena on löytää ratkaisut seuraaviin tutkimuskysymyksiin:

- Millaista lisäarvoa lintuhavaintoihin on mahdollista saada yhdistämällä niihin säädataa ja lajitietoa?
- Miten biologisista havainnoista koostuvaa tutkimusdataa kannattaisi julkaista linkitettynä datana ja rikastaa muilla aineistoilla, jotta lisäarvo on parhaiten saavutettavissa?
- Onko lintumuuttoaineiston päivittäisten havaintomäärien ja säähavaintoaineiston väliltä mahdollista hahmottaa yhteyksiä visualisoimalla dataa?
- Voidaanko lintuhavaintoaineiston laatua parantaa julkaistaessa sitä linkitettynä datana?

Tämän tutkielman alustavia tuloksia on julkaistu aiemmin havaintoaineistojen yhdistämistä käsittelevässä artikkelissa [KHL14].

Tässä työssä toteutettiin Hangon lintuaseman havaintoaineiston ja Ilmatieteenlaitoksen Hangon Russarön säähavaintoaineiston muuntaminen RDF-muotoon, linkittäminen toisiinsa ja lintuaineiston kuvailuun käytetyn taksoniontologian rikastaminen. Aineistot on julkaisu Linked Data Finland (LDF) -alustalla [HTAM14]. Lisäksi on toteutettu visualisointipalvelu aineistojen ja niiden välisten suhteiden visualisoimiseksi.

Biologisia havaintoaineistoja ja taksonomiaan liittyviä tietojenkäsittelyllisiä haasteita käsitellään luvussa 2. Luvussa 3 tarkastellaan lintu- ja säähavaintoaineistojen alkuperäismuotoja. Luku 4 on yleiskatsaus linkitettyyn dataan ja sen käyttämiin tekniikoihin ja standardeihin. Luku 5 käsittelee lintu- ja säähavaintoaineistojen mallintamista RDF-muodossa. Luvussa 6 käydään läpi aineistojen muunnosprosessi alkuperäismuodoista RDF-muotoon. Linkitettyjen RDF-muotoisten aineistojen visualisointia ja kehitettyä visualisointipalvelua tarkastellaan luvussa 7. Tutkimuksen tuloksia tarkastellaan luvussa 8. Yhteenveto tutkielmasta on luvussa 9.

2 Biologiset havaintoaineistot

Tässä tutkielmassa tarkasteltavat biologiset havaintoaineistot koostuvat lajien (tai muiden *taksonien* tai lajiryhmien) esiintymisestä ajassa ja paikassa, eli ovat ns. *biodiversiteettidataa*.

2.1 Yleisiä piirteitä

Biodiversiteettidata koostuu tyypillisesti havainnoista, joista tiedetään jollain tarkkuudella ainakin paikka, aika sekä havaitun eliön laji tai lajia yleisempi tai suppeampi määritys, kuten suku tai alalaji. Aineistot voidaan jakaa kahteen päätyyppiin:

- **Systemaattisesti kerätyt tutkimusaineistot** ovat esimerkiksi tutkimusprojekteissa syntyneitä havaintoaineistoja, joissa voidaan mahdollisesti keskittyä joihinkin lajeihin tai lajiryhmiin. Monesti myös havaintopaikat on rajattu johonkin maantieteelliseen alueeseen ja myös

aika on jotenkin rajattu. Tutkimusprojektien havaintoaineistoissa tyypillisesti merkitään ylös kaikki havainnot tutkimuksen kohteena olevista lajeista ja havainnointitapa on vakioitu ja dokumentoitu. Ajan, paikan ja lajin lisäksi tutkimuksissa voidaan kerätä monenlaisia muita tietoja havaituista eliöistä.

- **Epäsystemaattiset aineistot** ovat esimerkiksi avointen havaintopalvelujen aineistoja tai muut aineistot, joissa havaintojen tekemisen ja kirjaamisen käytännöt eivät ole vakioituja tai dokumentoituja. Lajeja havainnoidaan vaihtelevin menetelmin ja havainnoista tyypillisesti kirjataan muistiin vain jollain tapaa mielenkiintoisena pidetyt.

Systemaattisesti kerätyistä aineistoista on mahdollista päätellä esimerkiksi populaatiotrendejä kohtalaisella varmuudella [L+08, Sve78, Kje98]. Päättely tämän tyyppisistä aineistoista on suoraviivaista. Epäsystemaattisista aineistoista on vaikeampi tehdä minkäänlaista päättelyä, kun muuttujina ovat kiinnostuksen kohteena olevien muuttujien lisäksi havainnointiteho, havainnointimenetelmät sekä vaihteleva havaintojen ilmoitusaktiivisuus.

Erilaisia ohjelmallisen päättelyn menetelmiä on onnistuneesti sovellettu luontohavaintoaineistoihin [FCH04, HCF+07, YWH10, SEK08]. Laajoista havaintoaineistoista voi löytyä tämän kaltaisilla menetelmillä mielenkiintoisia tuloksia, joita ei ole vielä osattu etsiä.

Lintuasemilla havainnointi on hyvin standardoitua ja niiden aineistojen perustella voidaan tehdä luotettavia päätelmiä lajien kannankehityksestä. Tyypillisesti jokainen havaittu lintu merkitään muistiin ja pyritään määrittämään lajilleen [Han13b]. Jos havaintoa ei saada määritettyä lajilleen, havainto merkitään jollekin laajemmalle taksonille.

2.2 Taksonomia

Havaitut lajit yleensä merkitään muistiin lajin tieteellisellä nimellä tai sen lyhenteellä. Tieteelliset nimet ovat kuitenkin puutteellinen tapa esittää nimien taustalla olevaa taksonomiaa [KHKP06, LVG14, TLH11].

Tämän kappaleen katsaus luonnon monimuotoisuuden tutkimiseen ja esittämiseen noudattelee pääpiirteissään J. Muonan selvitystä [Muo04] aiheesta.

Luonnon monimuotoisuuden tutkimisessa olennaista on näkemys eliöiden sukulaisuussuhteista. *Systematiikka* on tutkimusta eliöiden sukulaisuussuhteiden ja luokittelun teoreettisista ongelmista eli siitä, miten maailman eliöitä ja niiden monimuotoisuutta ylipäänsä voidaan luokitella. *Taksonomia* on systematiikkaa soveltavaa tutkimusta siitä, miten eliöt käytännössä jakautuvat eri luokkiin. Yleisnimitys näille käytännön jaoille on *taksoni*. Esimerkkejä taksoneista ovat liito-orava, valkovuokko, koiraeläimet, pääjalkaiset, peippo, linnut ja eläinkunta. Jokainen taksoni sijoitetaan johonkin *taksonomiseen tasoon*, joka kuvaa sitä, mille tasolle eliöiden välisissä suhteissa kyseinen taksoni sijoittuu. Taksonomisia tasoja ovat esimerkiksi laji, heimo, pääjakso ja alalaji. *Taksonin rajaus* (circumscription) tarkoittaa sitä tuntomerkkien joukkoa, joka erottaa kyseiseen taksoniin kuuluvat eliöt muihin taksoneihin kuuluvista.

Lajilistat (checklist) ovat listoja jollain alueella esiintyvästä lajistosta, jotka ovat rajattu vielä johonkin taksonomiseen ryhmään, kuten lintuihin. Tyypillisesti lajilistoja muodostetaan eri maiden lajistosta, mutta myös suppeampien alueiden ja koko maapallon kattavia lajilistoja on myös olemassa. Lajilistoissa usein on myös lajia ylemmät taksonit ja mahdollisesti myös lajia suppeampia taksoneita kuten alalajeja, rotuja tai muotoja. Lajilista kertoo aina jostain taksonomisesta tulkinnasta eli siitä, miten eliöt jaetaan lajeihin ja muihin taksoneihin. Uuden biologisen tutkimuksen myötä taksonomiset käsitykset tarkentuvat, joten lajilistat kehittyvät jatkuvasti.

Lajilistojen välillä voidaan huomata esimerkiksi jonkin lajin siirtyneen suvusta toiseen ja jonkin lajin tieteellisen nimen muuttuneen. Lajin rajaus voi olla pysynyt samana, vaikka lajin tieteellinen nimi olisi muuttunut. Toisaalta kyse voi olla myös jonkun lajin jakamisesta useaksi tai usean lajin yhdistämisestä samaksi lajiksi. Lajin taksonominen rajaus ei välttämättä ole pysynyt samana, vaikka nimi olisi kahdessa listassa sama. Esimerkiksi alueellisesti rajatusta lajilistasta ei nähdä, jos laji on jaettu kahdeksi lajiksi, joista toista ei esiinny alueella. Myös uuden lajin ilmaantuminen listalle vaatii tietoa siitä, onko kyseinen laji jaettu jostain toisesta lajista vai onko kyseessä laji, joka on levittäytynyt alueelle. Maantieteellisesti rajatut lajilistat viestivät taksonomiasta ainoastaan paikallisesta näkökulmasta. Ontologisesti tämä on ongelmallista, koska taksoneiden rajauksia olisi mielekkäämpää esittää

globaalissa laajuudessa.

Lajilistasta ei siis ilmene taksoneiden rajaukset, mutta lajilistojen välisten muutosten perusteella on mahdollista päätellä muutoksia tapahtuneen, jos esimerkiksi jonkin suvun sisältämät lajit eivät ole enää täysin samat. Kysymykseen siitä, mitä kyseiset muutokset ovat, tarvitaan lajilistan ulkopuolista tietoa tapahtuneista taksonomisista muutoksista. Taksonien nimien tulkinta taksonien rajauksiksi vaatii asiantuntijan työpanosta [Fra11, SSB08], eikä oikeaa rajausta välttämättä ole jälkikäteen edes mahdollista päätellä.

3 Käytetyt havaintoaineistot

Tässä luvussa kerrotaan Hangon lintuaseman ja Russarön sääaseman havaintoaineistoista.

3.1 Hangon lintuaseman havaintoaineisto

Hangon Lintuasema sijaitsee Hangon Tulliniemen kärjessä, lähellä Hangon kaupunkia. Hankoniemi toimii syksyisin merkittävänä lintumuuton johtolinjana.

Lehikoinen ja Vähätalo [LV00] sekä aseman miehitysohjeet [Ohj14] selvittävät aineiston keräämisen käytäntöjä. Aseman miehittäjien velvollisuuksiin kuuluu suorittaa havainnointia vakioidulla menetelmällä jokaisena aamuna auringon noususta alkaen. Vakioidun aamuisen havainnoinnin eli ”aamuvakion” pituus on normaalisti neljä tuntia. Marraskuulta maaliskuulle aika on typistetty kahteen tuntiin muuton vähyyden ja inhimillisten syiden takia. Aamuvakio suoritetaan aina samasta toisen maailmansodan aikaisesta tulenjohtobunkkerista, josta on hyvä näkyvyys jokaiseen suuntaan. Vakiohavainnointiin kuuluu merkitä ylös jokainen muuttava lintu. Koko päivän ajalta lasketaan yhteen jokaisen lajin muuttajamäärät, paikalliset linnut aseman alueelta sekä paikalliset linnut asema-alueen ulkopuolelta.

Miehitys asemalla perustuu vapaaehtoisuuteen ja eri aikojen erilainen miehitys tuokin kenties suurinta vaihtelevuutta aineistoon. Kuitenkin vuodesta 2000 alkaen havainnointiaktiivisuus on pysynyt lähes ympärivuotisena ja tasaisempana yhden henkilön tuoman pysyvän havainnointipanoksen ansiosta [L$^+$08].

Havaintoaineiston kerääminen on aloitettu vuonna 1979 ja sitä on vuosien saatossa käytetty useissa tutkimuksissa ja muissa julkaisuissa [Han13a]. Asemalla kerätään myös muitakin aineistoja [LV00], kuten lintujen rengastustoiminnasta syntyviä laji- ja päiväkohtaisia lintumääriä, mutta nämä eivät kuulu tässä tutkimuksessa käytettyihin aineistoihin. Havaintodataa on kerätty järjestelmällisesti riittävän kauan, jotta sen pohjalta pystytään päättelemään populaatiotrendejä niin yleisten kuin harvalukuistenkin lajien osalta [L$^+$08].

Taulukkomuotoinen havaintoaineisto on kahdessa tiedostossa. Näistä yhdessä on laji- ja päiväkohtaisesti paikallisten aseman alueella olevien lintujen, aseman ulkopuolella havaittujen lintujen ja muuttavien lintujen summat vuodesta 1979 vuoteen 2009. Toisessa tiedostossa ovat päivittäiset aamuvakion havainnot vuodesta 1979 vuoteen 2008. Esimerkki ensimmäisen havaintotiedoston muodosta on taulukossa 1. Aineistojen tulkintaan on saatu ohjeistusta aineiston omistavan Tringa ry:n edustajalta. Koska lintuhavaintodatan käyttöehdot eivät salli aineiston osien julkaisua, tutkielmassa esitettyjen esimerkkien lukuarvoja on muutettu alkuperäisestä.

Laji	**Päivämäärä**	**Paikalliset**	**Muutto**	**Lisäalue**
TADTAD	5/15/1991	2	1	
TADTAD	5/17/1991	3	0	

Taulukko 1: Kaksi riviä alkuperäisdatan muotoista havaintodataa.

3.2 Russarön sääaseman havaintoaineisto

Ilmatieteenlaitoksen Russarön säähavaintoasema sijaitsee Hangossa alle 6 kilometrin etäisyydellä Hangon lintuasemasta.

Säähavaintoaineisto on alkuperäismuodossaan kahtena taulukkomuotoisena tiedostona. Päätiedosto koostuu 3 tunnin välein tehdyistä säähavainnoista, joissa havainnoituja muuttujia ovat lämpötila, suhteellinen ilmankosteus, tuulen suunta, tuulen nopeus, tuulen puuskanopeus, ilmanpaine ja kokonaispilvisyys. Tuulen suunnat ovat alkuperäisessä datassa ilmaistuina 10 asteen tarkkuudella. Esimerkki datan muodosta on taulukossa 2. Kaikkia näistä muuttujista ei ole saatavissa koko ajalta 1979–2011 ja lisäksi datassa on

satunnaisesti aukkoja.

Toisessa tiedostossa on päivittäinen sademäärä koko päivän summana. Sademääriä ei ole käytössä koko lintuhavaintoaineiston aikajänteeltä, vaan ne rajoittuvat vuodesta 1979 vuoteen 2005.

Vuosi	1987	1987
Kuukausi	2	2
Päivä	3	3
Tunti	18	21
Lämpötila	-12.2	-13.1
Suhteellinen kosteus	74	
Tuulen suunnan 10 min keskiarvo	60	60
Tuulen nopeuden 10 min keskiarvo	9	9
Tuulen puuskanopeus		
Ilman paine merenpinnan tasolla	1012.1	1012.4
Kokonaispilvisyys	8	8

Taulukko 2: Kaksi esimerkkiriviä säähavaintodatan alkuperäisestä muodosta esitettynä sarakkeina.

4 Linkitetty data

Tässä luvussa käsitellään linkitetyn datan periaatteita, tekniikoita ja standardeja.

Käsite *linkitetty data* (Linked Data) viittaa Tim Berners-Leen esittelemiin [Ber06] periaatteisiin rakenteellisen tiedon ja metatiedon julkaisemiseksi ja yhteenliittämiseksi verkossa [BHB09, HB11]. Nämä periaatteet ovat:

1. käytä *URI*-tunnisteita (uniform resource identifier) asioiden nimeämiseen,

2. käytä URI-tunnisteina HTTP-osoitteita, jotta ihmiset voivat löytää niistä lisätietoa,

3. kun joku etsii lisätietoa URI:sta, tarjoa hyödyllistä tietoa käyttäen standardeja kuten RDF, RDFS, OWL ja SPARQL sekä

4. tarjoa tunnisteeseen liittyviä linkkejä muihin URI-tunnisteisiin, jotta ihmiset voivat löytää uusia asioita.

Dokumenteista koostuva *World Wide Web* (WWW) on rakennettu pienelle joukolle standardeja, joista tärkeimmät ovat [JW14]: URI, HTTP ja HTML. Ideana linkitetyssä datassa ja laajemmin semanttisessa webissä on käyttää WWW:n arkkitehtuuria pohjana rakenteellisen datan globaaliin jakamiseen.

Linkitetty data esitetään käyttäen RDF-tietomallia, jossa käsitteistä voidaan luoda linkkejä toisiin käsitteisiin. Näin voidaan rikastaa aineistoja tekemällä linkityksiä aineistoista toisiin. Aineistot voivat käsitellä samoja aiheita tai olla keskenään täysin erilaisia, kuitenkin viitaten samoihin käsitteisiin, kuten esimerkiksi paikkoihin tai henkilöihin. Käsitteisiin viitataan niiden URI- tai *IRI*-tunnisteen (internationalized resource identifier) perusteella, jotka ovat tietynlaisia merkkijonoja. Nämä merkkijonot tulisi olla linkitetyn datan periaatteiden mukaisesti HTTP-osoitteita, joista löytyy kyseisen käsitteen koneluettava kuvailu [CWL+14]. URIn ja IRIn erona on se, että uudempi IRI on suunniteltu tukemaan kansainvälisiä merkistöjä ja vanhempi URI ei [DS05, CWL+14].

Toisiinsa linkittyneiden käsitteiden muodostamasta kokonaisuudesta käytetään nimitystä *graafi* eli verkko. Toisiinsa linkitettyjen ja avoimesti Internetissä saatavilla olevien aineistojen muodostama verkko yhdessä käytettyjen standardien ja teknologioiden kanssa tunnetaan nimellä *Semanttinen Web* (*Semantic Web* tai *Web of Data*) [HB11].

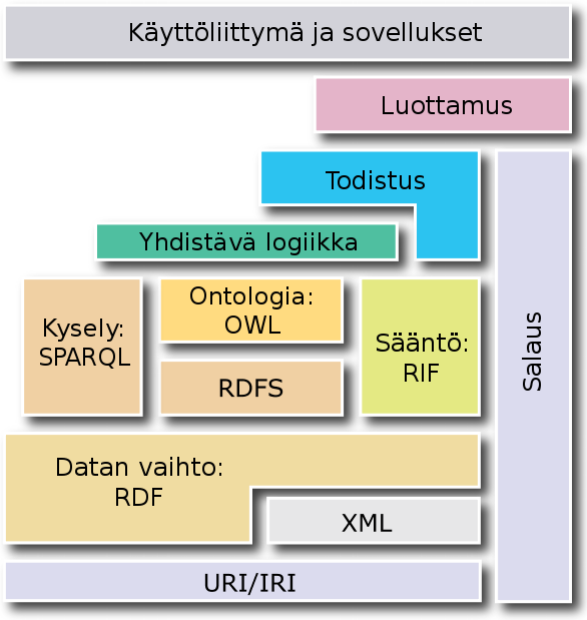
4.1 Semanttinen Web

World Wide Web esittää tietoa muodossa, jossa ihmisen on sitä helppo omaksua: etupäässä tekstinä ja kuvina. Tietokoneen on mahdollista tutkia tällaisia sivuja niiden ulkoasun perusteella ja löytää otsikoita ja linkkejä toisiin sivuihin, mutta tietokoneella ei ole juurikaan mahdollisuuksia ymmärtää merkityksiä näiltä sivuilta [BHL01].

Semanttinen Web on visio koneluettavan ja koneymmärrettävän datan muodostamasta verkosta [B+06]. Se tuo rakenteen verkossa esitettävän sisäl-

lön merkitysten esittämiseen, mikä mahdollistaa älykkäiden järjestelmien toiminnan verkosta löytyvien tietojen pohjalta [BHL01]. Älykkäät järjestelmät voisivat suorittaa monimutkaisia tehtäviä käyttäjille, kuten selvittää mikä on lähin avoinna oleva ruokakauppa tai varata liput seuraavalle lennolle Berliiniin. Datan esittäminen laajana koneluettavana verkkona tarjoaa uudenlaisia mahdollisuuksia datan hyödyntämiseksi kuten uuden tiedon päätteleminen annetun tiedon pohjalta ja aineistojen helpompi yhdistäminen.

Semanttisen webin keskeiset komponentit on esitetty kuvassa 1. *World Wide Web Consortium* (W3C) on vastuutaho Semanttisen Webin kehityksessä ja standardien luomisessa [AH11]. Tämän tutkielman kannalta olennaisia komponentteja käsitellään tarkemmin seuraavissa aliluvuissa.



Kuva 1: Semanttisen webin kerrosrakenne [H+13].

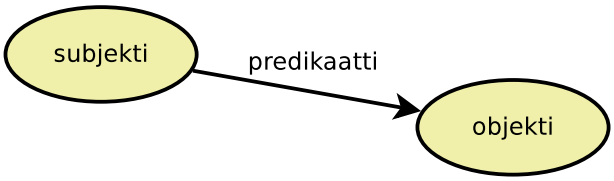
4.2 Resource Description Framework (RDF)

Semanttisen webin keskeisin standardi on *Resource Description Framework* (RDF), jonka avulla voidaan esittää formaalisti semanttista kuvailua eli annotointia [Pan09]. Semanttinen kuvailu pyrkii esittämään asioiden merkityksiä

formaalissa muodossa. Annotoinnit eivät itsessään määrittele käsiteltävän asian semantiikkaa, vaan merkitykset annotointeihin luodaan ontologioiden avulla [Pan09].

RDF on tietomalli ontologiaperustaisten annotointien esittämiseen [Pan09, Hog14]. RDF-lause (statement) on subjektin, predikaatin ja objektin muodostama kolmikko kuvan 2 mukaisesti. Predikaatilla ilmaistaan minkälainen suhde subjektilla on objektiin. Verkkojen osien nimiä *kaari* (arc) ja *solmu* (node) käytetään myös RDF-tietomallin yhteydessä.

RDF-lauseiden osat voivat olla joko *resursseja* (resource), *tyhjiä solmuja* (blank node) tai *literaaleja*. Resursseihin viitataan niiden URI-tunnisteella. Literaaleilla tarkoitetaan joko yksinkertaista merkkijonoa tai jotain tietotyypitettyä merkkijonona ilmaistua arvoa. RDF tukee osaa XML Scheman[1] tietotyypeistä, joiden lisäksi voidaan käyttää itse määriteltyjä tietotyyppejä [CWL+14].



Kuva 2: Subjekti-predikaatti-objekti-kolmikko.

RDF-muotoiset aineistot voidaan esittää sarjallistettuina erilaisilla syntakseilla, joista keskeisimpiä ovat RDF/XML, RDFa, Turtle, N-Triples, N-Quads ja Notation3.

4.3 Ontologia

Ontologia-käsitteen (ontology) historia juontaa juurensa antiikin Kreikkaan, jossa ontologia oli oppi olevaisesta. Matematiikan näkökulmasta ontologia on suunnattu verkko, joka esittää jotakin tietoa todellisuudesta. Ontologian matemaattinen malli koostuu loogisista aksioomista, jotka mahdollistavat päättelyn verkon tietojen perusteella [Ehr07].

[1] `http://www.w3.org/TR/xmlschema11-2/`

Nykyisin tekoäly- ja web-tutkijat ovat ottaneet ontologia-käsitteen käyttöönsä ja sanan määritelmät vaihtelevat kirjallisuudessa [Ehr07]. Alunperin tietojenkäsittelyn näkökulmasta vuonna 1993 luodun määritelmän mukaan ontologia on ”eksplisiittinen määritelmä käsitteellistämisestä” (”explicit specification of a conceptualization”) [G+93]. Tätä määritelmää on myöhemmin hieman tarkennettu ja esitetty formaalimmin. Keskeisimpinä tarkennuksina määritelmään on tullut vaatimus jaetusta käsitteistöstä ja koneluettavuudesta, jolloin ontologian määritelmä on ”formaali eksplisiittinen määritelmä jaetusta käsitteellistämisestä” (”formal, explicit specification of a shared conceptualization”) [GOS09]. Ontologia on siis tapa mallintaa formaalisti jonkin järjestelmän rakennetta.

Linkitetyn datan kuvailuun käytetyt ontologiat kuvaavat jotain osaa todellisuudesta ja koostuvat formaalista tietyn aihealueen yhteisestä käsitteistöstä ja formaaleista suhteista käsitteiden välillä [GOS09]. Ontologian kehittäjä organisoi aiheeseen liittyvät entiteetit *käsitteisiin* (concepts) ja *suhteisiin* (relations) [GOS09].

4.4 Ontologia- ja sanastokielet

RDF Schema (RDFS) on RDF-sanaston laajennos, jolla voidaan ilmaista yksinkertaisia ontologioita RDF-tietomallilla [Pan09, BGM14]. Keskeinen komponentti RDFS:ssä on *luokka* (class), joita ilmaistaan resurssilla `rdfs:Class`. Jokin RDF-resurssi on luokan ilmentymä, jos sillä on predikaatilla `rdf:type` määritelty suhde luokkaan. RDFS tukee luokkahierarkioita yläluokkaan viittaavan ominaisuuden `rdfs:subclassOf` avulla. Luokan ilmentymä on aina myös kyseisen luokan yläluokkien ilmentymä.

RDFS tarkentaa *ominaisuuksien* (property) ilmaisemista, joita käytetään RDF-lauseiden predikaatteina. Ominaisuudet ovat luokan `rdf:Property` ilmentymiä. RDFS mahdollistaa esimerkiksi ominaisuuden määrittely- ja arvojoukkojen ilmaisemisen. Ominaisuuksista voidaan luoda hierarkioita viitaten yläominaisuuteen käyttäen ominaisuutta `rdfs:subPropertyOf`. RDFS-ontologioita kutsutaan usein *sanastoiksi* [HB11].

Simple Knowledge Organization System (SKOS) on W3C:n standardoima yksinkertainen sanasto käsitehierarkioiden esittämiseen, joista käytetään

myös nimitystä *taksonomia*. Tätä ei kuitenkaan tule sekoittaa biologiseen taksonomiaan. *Web Ontology Language* (OWL) on RDFS:n laajennos semanttisesti ilmaisuvoimaisempien ontologioiden esittämiseen [Mv04].

Formaalista metatietomallista tai tietomallista käytetään usein nimeä *skeema* (schema).

4.5 SPARQL-kyselykieli

SPARQL on laajasti käytetty W3C:n standardoima kyselykieli RDF:lle [HSP14, Dod05]. SPARQL on tuettuna kaikissa yleisimmissä RDF-tietovarastoissa [HB11]. Kyselyt muotoillaan osittain saman kaltaisten rakenteiden ja termien avulla kuin SQL-kielillä. Kyselyn vastauksena saadaan kyselystä riippuen joko vastausjoukko tai RDF-graafi [HSP14].

4.6 Biologiset aineistot ja RDF

Biologisten havaintoaineistojen kuvaamiseen ja jakamiseen on kehitetty Darwin Core [WBG$^+$12, W$^+$13] metatietomalli. Darwin Core:n käsitteiden semantiikka on määritelty RDF-muodossa.

Luontohavaintoaineistoissa usein käytetyt tieteelliset nimet ovat riittämättömiä taksonomisten käsitteiden esittämiseen. RDF:ää ja ontologioita voidaan käyttää taksonomian ja taksonomisten muutosten mallintamiseen [Rod06, Fra11, TLH11]. Yksi malli tällaisten ontologioiden kuvaamiseen on TaxMeOn-metaontologia [TLH11].

On todettu, että biologisten aineistojen parempi yhteentoimivuus voidaan saavuttaa julkaisemalla aineistot linkitettynä datana [RJS11].

4.7 Linkitetty tilastollinen data

Kiinnostus julkaista tilastollista dataa linkitettynä datana on ollut kasvussa viime vuosien aikana [KH11, HHR$^+$09]. *Statistical Core Vocabulary* (SCOVO) on ensimmäisiä yrityksiä luoda ontologia tilastollisten aineistojen mallintamiseksi RDF:nä [HHR$^+$09]. SCOVO on kuitenkin hyvin rajoittunut [Dat11], eikä enää ylläpidetty.

RDF Data Cube Vocabulary on sanasto, joka on kehitetty moniulotteisen datan, kuten tilastotietojen, esittämiseen RDF-muodossa [CR14, SAB$^+$12].

Sanaston pohjana on käytetty *Statistical Data and Metadata eXchange* (SDMX) -standardin tietomallia, ja RDF Data Cube -sanasto on yhteensopiva tämän kanssa [CR14]. SDMX on 2001 käynnistynyt hanke tilastollisten aineistojen yhteentoimivuuden parantamiseksi [Sta09]. SDMX tarjoaa standardoituja datan ja metadatan esitysmuotoja sekä sisältöohjeita ja IT-arkkitehtuurin datan ja metadatan välittämiseen [Sta09].

RDF Data Cube -sanasto on saavuttanut statuksen virallisena W3C:n suosituksena tammikuussa 2014 [CR14]. Sanasto on jo ennen tätä ollut aktiivisessa käytössä. Esimerkiksi linkitetyn avoimen datan aineistojen tilastotietoja julkaiseva LODStats on julkaissut tilastojaan käyttäen RDF Data Cube -sanastoa [ADML12, EMLA13].

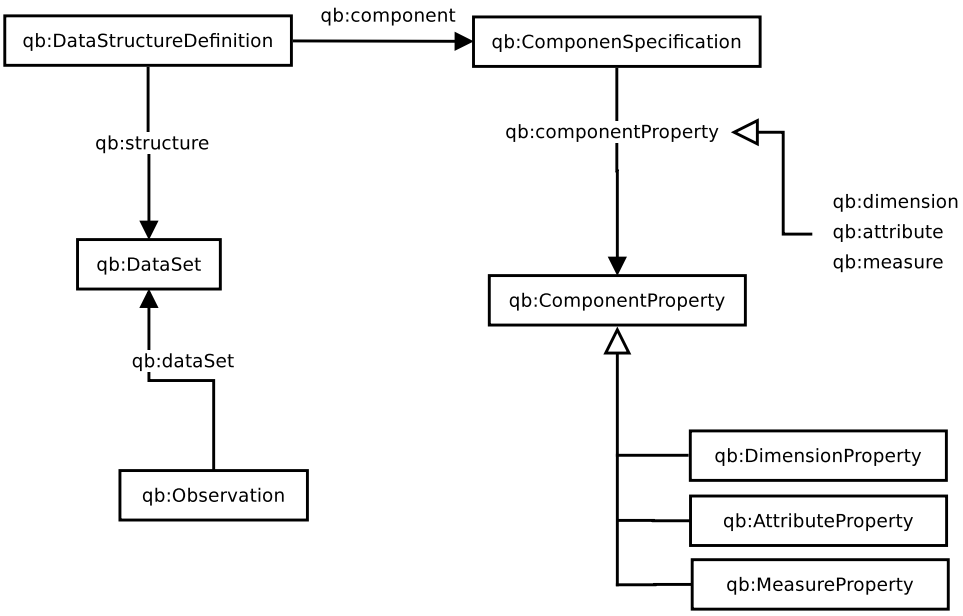
Sanastoa käyttävien aineistojen julkaisun helpottamiseksi Salas ja kumppanit [SAB+12] ovat kehittäneet työkalut *OLAP2DataCube* ja *CSV2DataCube*, joilla tilastollista dataa voidaan muuntaa RDF Data Cube -muotoon. Muunnoksissa on mahdollista käyttää myös geneerisiä työkaluja linkkien etsintään muunnettavien aineistojen ja muiden aineistojen välillä.

Cyganiak ja Reynolds [CR14] mainitsevat seuraavat hyödyt moniulotteisen datan julkaisemisesta RDF-muodossa käyttäen linkitetyn datan periaatteita:

- Yksittäisiin havaintoihin ja havaintoryhmiin voidaan viitata yksilöllisellä tunnisteella. Tällöin esimerkiksi jokin raportti voi viitata suoraan lukuihin, joihin se perustuu.
- Dataa voidaan joustavasti yhdistellä aineistojen välillä. Tällöin tilastotiedosta tulee keskeinen osa laajempaa linkitetyn datan verkkoa.
- Julkaisu linkitettynä datana tarjoaa joustavan, avoimeen formaattiin perustuvan ja koneluettavan julkaisutavan, jolle löytyy valmis Web-rajapinta.
- Linkitetty data mahdollistaa työkalujen ja komponenttien uudelleenkäytön.

Kuvassa 3 on esitettynä tärkeimpiä RDF Data Cube -sanaston käsitteitä ja niiden suhteita toisiinsa. Käsitteiden nimiavaruus `qb` viittaa RDF Data Cube:n nimiavaruuteen `http://purl.org/linked-data/cube#`. Kuvan

mustat nuolet ilmaisevat predikaattia, jonka objektiin nuoli osoittaa. Valkoiset nuolet tarkoittavat aliluokkaa tai -ominaisuutta. Yleisen käytännön mukaisesti isolla alkukirjaimella kirjoitetut resurssien nimet ovat luokkia ja pienellä kirjoitetut ovat ominaisuuksia.



Kuva 3: Otos RDF Data Cube -sanaston tärkeimpiä käsitteitä ja niiden suhteet toisiinsa. [CR14].

Aineistot ovat luokan `qb:DataSet` ilmentymiä ja aineiston rakenne määritellään luokan `qb:DataSetDefinition` ilmentymällä. Aineiston rakenne koostuu dimensioista (dimension), ominaisuuksista (attribute) ja mittauksista (measure). Dimensioiden avulla yksilöidään havainto, eikä samoilla dimensioiden arvoilla saa olla useita havaintoja. Dimensioita ovat esimerkiksi havainnon ajankohta ja paikka. Mittaukset esittävät havainnoitavana olevaa ilmiötä. Ominaisuudet mahdollistavat mittausyksiköiden ja muun havaintoon liittyvän metatiedon esittämisen. Itse aineistot koostuvat havainnoista (`qb:Observation`) ja niiden `qb:DataSetDefinition`:ssa määritellyistä komponenteista.

Data Cube -sanaston yhteydessä suositellaan käytettävän DCMI Meta-

data Terms[2] -sanaston mukaisia metatietoja aineistoista [CR14]. Toimijoille, kuten julkaisija, suositellaan käytettävän sanastoja FOAF[3] tai Core Organization Ontology[4].

4.8 Linkitetyn datan mallinnus

Koska aineistoja voidaan mallintaa linkitettynä datana myös eri tavoin, on käytetyillä ontologioilla ja aineiston muodolla merkitystä aineiston käytettävyyteen. Dataa mallinnettaessa ei kaikkia käyttötapauksia kuitenkaan voi ottaa huomioon. Tapana onkin linkitetyn datan piirissä ollut mallintaa aineisto juuri käsillä olevaan tarpeeseen soveltuvana ottaen kuitenkin hieman huomioon myös muiden aiheesta kiinnostuneiden tahojen mahdolliset käyttökohteet [AH11].

Datan mallintamiseen linkitettynä datana kuuluu kaikkien datassa olevien käsitteiden tunnistaminen ja aihetta käsittelevien ontologioiden luominen niiltä osin, kuin ei ole olemassa käyttökelpoisia valmiita ontologioita [PMP13].

T. R. Gruber [Gru95] listaa viisi ontologioiden suunnittelun periaatetta:

1. **Selkeys**: Määritellyille käsitteille tarkoitettu merkitys pitäisi välittyä tehokkaasti ontologiasta. Määrittelyt tulisi dokumentoida luonnollisella kielellä.

2. **Johdonmukaisuus**: Ontologian olisi hyväksyttävä päättelyt, jotka ovat yhdenmukaisia määritelmien kanssa. Jos ontologiaa ja ontologiakielen aksioomia käyttäen päätelty lause on ristiriidassa jonkin määritelmän kanssa, ontologia on epäjohdonmukainen.

3. **Laajennettavuus**: Ontologia tulisi suunnitella ennakoiden mahdollisia tulevia käyttötapauksia ja ontologian laajentamista.

4. **Pienin koodauksen puolueellisuus** (minimal encoding bias): Käsitteet tulisi mallintaa tietämyksen tasolla oikein, eikä räätälöidä esitysmuotoa tiettyä käyttötapausta tai helppoa esitysmuotoa varten. Tämä edistää ontologian toimivuutta toisistaan poikkeavissa järjestelmissä.

[2] `http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms`
[3] `http://xmlns.com/foaf/spec/`
[4] `http://www.w3.org/ns/org#`

5. **Pienin ontologinen sitoumus**: Ontologian pitäisi tehdä vain sen verran väitteitä mallinnettavasta maailmasta kuin on tarpeen ontologian käsillä olevan käytön kannalta. Tämä edesauttaa ontologian jatkokäyttöä.

Linkitetyn datan aineistojen metatietoja ilmaistaan tyypillisesti *Vocabulary of Interlinked Datasets (VoID)* [C+11] -sanastolla [HB11].

4.9 Linkitetyn datan visualisointi

Linkitetyn datan visualisointiin on erilaisia tapoja, jotka voidaan jakaa seuraavaan kolmeen luokkaan [KHL+07, DR11]. Voidaan 1) visualisoida datan rakennetta, kuten ontologioita, 2) esittää datan analyysien tuloksia, kuten tilastoja tai 3) esittää ilmiöitä käyttäen erilaisia graafisia menetelmiä, kuten datan esittäminen kartalla, aikajanalla tai muulla aiheeseen sopivalla tavalla.

Mutlu ja kumppanit [MHS+13] ovat kehittäneet visualisointityökalun, joka automaattisesti ehdottaa soveltuvia visualisointeja datan ja semanttisen määrittelyn perusteella. Kämpgen ja Harth ovat käsitelleet [KH11] avoimen lähdekoodin OLAP-visualisointityökalujen käyttöä tilastollisen linkitetyn datan visualisointiin.

Salas ja kumppanit [SAB+12] ovat kehittäneet *CubeViz*-visualisointityökalun RDF Data Cube -muotoiselle datalle.

4.10 Linkitetyn datan julkaiseminen

W3C on esittänyt hyviä tapoja linkitetyn datan julkaisemiseksi [HAV14]. Näiden perusteella jonkin aineiston julkaisu koostuu pääpiirteissään seuraavanlaisesta työnkulusta:

1. **Datan mallinnus**: Datassa esiintyvien tietojen ja niiden suhteiden esittäminen sovellusriippumattomalla tavalla.

2. **Lisenssin valinta**: Tulisi valita sopiva avoin lisenssi. Selvät käyttöehdot edistävät datan uudelleenkäyttöä.

3. **Luo asioille hyvät URI-tunnisteet**: HTTP-URI-tunnisteiden luominen huolellisesti suunnitellun strategian mukaisesti.

4. **Yleisten sanastojen käyttäminen**: Aina kun mahdollista, tulisi käyttää jo olemassa olevia sanastoja. Tarvittaessa tulisi laajentaa sanastoja ja jos välttämätöntä, luoda uusia sanastoja.

5. **Muunna data linkitetyksi dataksi**: Datamuunnos tehdään yleensä ohjelmallisesti.

6. **Tarjoa ohjelmallinen pääsy dataan**: Hakukoneita ja muita ohjelmia varten tulisi tarjota pääsy dataan standardeilla tavoilla.

7. **Ilmoitus aineiston julkaisusta**: Mahdollisia kiinnostuneita tahoja tulisi informoida avoimen aineiston julkaisusta.

8. **Ylläpito ja saatavuus**: Aineiston julkaisijalla on velvollisuus ylläpitää aineistoa ja taata sen saatavuus.

Petrou ja kumppanit [PMP13] ovat esitelleet menetelmän taulukkomuotoisten tilastollisten aineistojen julkaisemiseksi linkitettynä datana käyttäen RDF Data Cube -sanastoa. Menetelmä noudattaa W3C:n linkitetyn datan julkaisun periaatteita.

5 Havaintoaineistojen mallinnus

Säätila tiedetään merkittäväksi yksittäisten päivien lintumuuttoa selittäväksi tekijäksi. Aihetta on tutkittu paljon [Abl73] ja jo 1800-luvulla on hahmoteltu säätilan ja muuttavien lintujen määrien suhdetta [Coo88]. Yhdistämällä Hangon Lintuaseman päivittäiset havainnot läheltä mitattuihin säähavaintoihin pyritään tekemään mahdolliseksi tarkempien päätelmien tekeminen lintuhavaintoaineiston pohjalta.

Tapoja aineistojen esittämiseen ja yhdistämiseen RDF-muodossa on useita. Yksittäisistä biologisista havainnoista koostuva havaintodata voidaan esittää tilastollisena datana, jonka ytimessä olevat ulottuvuudet ovat paikka, laji ja ajankohta. RDF Data Cube -sanasto soveltuu hyvin yksittäisten havaintojen tai koko päivän havaittujen summien esittämiseen ja aineistot on mallinnettu tätä sanastoa käyttäen.

Havaintojen esittämisessä olisi mahdollista käyttää myös esimerkiksi *Open Geospatial Consortium*:in ja ISO:n standardia *Observations and Measurements* [Cox13a], joka on saatavilla myös OWL-versiona [Cox13b]. Standardi on etupäässä automaattisten sensorien havaintodatan esittämiseen kehitetty tietomalli, joka soveltuu myös muuhun havainto- ja mittausdataan. Mallin käyttämässä rakenteessa jokainen mittaustulos on erillinen havainto. Observations and Measurements ei salli moniulotteisen havaintodatan esittämistä kuten RDF Data Cube eivätkä nämä mallit ole keskenään yhteentoimivia.

Tässä tutkimuksessa käytetyt aineistot on pyritty mallintamaan niin, että aineistojen visualisointi käyttäjälle onnistuu hyvin. Koska aineistot ovat suuria, mallinnuksessa on pyritty pitämään tarvittavien RDF-lauseiden määrä pienenä. Suunnittelemalla ja kokeilemalla erilaisia mallinnuksia aineistoista RDF-muodossa on lopulta päädytty malliin, joka koostuu kolmesta eri RDF Data Cube -mallin datakuutiosta, joista yhdessä on lintuhavainnot, toisessa päiväkohtaiset keskiarvoistetut ja aggregoidut säämuuttujat ja kolmannessa alkuperäiset säähavainnot. Ontologioita suunniteltaessa on pyritty ottamaan huomioon luvussa 4.8 esitetyt ontologioiden suunnittelun periaatteet. Havaintojen rakenne on täytynyt pitää yksinkertaisena, jotta visualisointeja varten aineistoon tehtävät kyselyt toimivat mahdollisimman nopeasti.

RDF Data Cube -standardi ei salli puuttuvia arvoja havaintojen `measure`- tai `dimension`-ominaisuuksille. Itse standardi ei ota kantaa siihen, miten alkuperäisestä aineistosta puuttuvat arvot tulisi mallintaa, mutta tämä voidaan standardin mukaisesti toteuttaa ainakin seuraavilla tavoilla. Voidaan käyttää jonkin XML-tietotyypin sopivaa arvoa kuten 0 tai esim. `xsd:double` -tietotyypin arvoa NaN (Not a Number). Fernández ja kumppanit [FZ13] ovat käyttäneet `measure`-ominaisuuksille itse määriteltyä resurssia merkitsemään puuttuvaa tai tuntematonta arvoa, jolloin arvon esittämiseen käytetyn ominaisuuden arvojoukko on määriteltävä itse sen sijaan, että voisi käyttää valmiita XML Scheman tietotyyppejä. Lefort ja kumppanit [LBH+12] ovat käyttäneet erillisiä `attribute`-ominaisuuksia ilmaisemaan puuttuuko datasta measure-ominaisuuksien arvoja.

Linkitettyjä aineistoja varten tehtyjen ontologioiden luomiseen ja muok-

kaamiseen on käytetty TopBraid Composer -ohjelman[5] ilmaisversiota sekä tekstieditoria.

5.1 Käytetyt nimiavaruudet

Käytettyihin nimiavaruuksiin viitataan tekstistä, joten ne esitellään tässä ennen varsinaisten ontologioiden esittelyä. Käyttöön otetut uudet nimiavaruudet on listattu taulukossa 3. Näiden lisäksi muita käytössä olevia ja tässä tutkielmassa esiintyviä nimiavaruuksia lyhenteineen on listattu taulukossa 4.

Etuliite	**URI**	**Selite**
hs	`http://ldf.fi/schema/halias/`	Havaintoaineistojen skeema
halias	`http://ldf.fi/halias/observations/birds/`	Lintuhavainnot
hw	`http://ldf.fi/halias/observations/weather/`	Vuorokausittaiset aggregoidut säähavainnot
r	`http://ldf.fi/halias/observations/russaro/`	Alkuperäiset sääha-vainnot
winds	`http://ldf.fi/halias/observations/winds/`	Instanssit havaituista tuulista
bc	`http://ldf.fi/halias/bird-characteristics/`	Tuntomerkkiontologia

Taulukko 3: Linkitetyn datan julkaisua varten käyttöön otetut nimiavaruudet.

5.2 Lintuhavaintoaineiston skeema

Datan mallintamisessa on käytetty etenkin RDF Data Cube -sanastoa sekä muita aiheeseen liittyviä sanastoja. Päivittäiset lintuhavainnot on esitetty yhtenä ”kuutiomallisena” aineistona ja päivittäiset säähavainnot toisena kuutiomallisena aineistona. Linkitys näiden aineistojen välillä tapahtuu päi-

[5] `http://www.topquadrant.com/tools/modeling-topbraid-composer-standard-edition/`

Etuliite	**URI**	**Selite**
bio	`http://www.yso.fi/onto/bio/`	AVIO-ontologia
dc	`http://purl.org/dc/elements/1.1/`	Dublin Core
dct	`http://purl.org/dc/terms/`	DCMI Metadata Terms
dgu-intervals	`http://reference.data.gov.uk/def/intervals/`	data.gov.uk Time Intervals
dwc	`http://rs.tdwg.org/dwc/terms/`	Darwin Core
foaf	`http://xmlns.com/foaf/0.1/`	FOAF Vocabulary
owl	`http://www.w3.org/2002/07/owl#`	OWL
qb	`http://purl.org/linked-data/cube#`	RDF Data Cube Vocabulary
rdf	`http://www.w3.org/1999/02/22-rdf-syntax-ns#`	RDF
rdfs	`http://www.w3.org/2000/01/rdf-schema#`	RDF Schema
sapo	`http://www.yso.fi/onto/sapo/`	Suomen ajallinen paikkaontologia (SAPO)
sdmx-attribute	`http://purl.org/linked-data/sdmx/2009/attribute#`	SDMX-attribute
sdmx-dimension	`http://purl.org/linked-data/sdmx/2009/dimension#`	SDMX-dimension
sdmx-subject	`http://purl.org/linked-data/sdmx/2009/subject#`	SDMX-subject
taxmeon	`http://www.yso.fi/onto/taxmeon/`	Taxon Meta-Ontology TaxMeOn
taxonomic-ranks	`http://www.yso.fi/onto/taxonomic-ranks/`	Taxonomic Ranks
void	`http://rdfs.org/ns/void#`	Vocabulary of Interlinked Datasets (VoID)
xsd	`http://www.w3.org/2001/XMLSchema#`	XML Schema
ysa	`http://www.yso.fi/onto/ysa/`	Yleinen suomalainen asiasanasto (YSA)

Taulukko 4: Linkitetyissä aineistoissa käytetyt ulkopuolisten ontologioiden ja sanastojen nimiavaruudet.

vämäärän avulla. Osa tarpeellisista tiedoista linkittyy Data Cube -muotoisen datan ulkopuolelle esim. taksoniontologiaan.

Taulukkomuotoinen esitys lintuhavaintoaineiston kuutiomuotoisen mallin `hs:haliasDSD` komponenteista on taulukossa 5. Data Cube:ssa `qb:DimensionProperty` -ominaisuudet yksilöivät jokaisen havainnon, joten kaikkien näiden arvot eivät saa olla samoja usealla havainnolla. Aineiston `hs:haliasDataSet` -kuutiossa on käytetty Halias-ontologiassa määriteltyjä komponentteja, joista osa käyttää `rdfs:subPropertyOf` -suhdetta joidenkin muualla kehitettyjen ontologioiden käsitteisiin. Viikkojen ja kuukausien numerojen ilmaisemiseen käytetyt käsitteet linkittyvät ontologiaan data.gov.uk Time Intervals[6]. Havaintojen päivämääriä esittävät käsitteet linkittyvät SDMX-dimension-ontologiaan ja Darwin Core -ontologiaan. Linkittämisellä muihin ontologioihin pyritään parantamaan datan jatkokäyttöä.

nimi	**arvojoukko**	**kardinaliteetti**	**luokka**
hs:countAdditionalArea	xsd:int	1	qb:MeasureProperty
hs:countLocal	xsd:int	1	qb:MeasureProperty
hs:countMigration	xsd:int	1	qb:MeasureProperty
hs:countStandardizedMigration	xsd:int	1	qb:MeasureProperty
hs:countTotal	xsd:int	1	qb:MeasureProperty
hs:monthOfYear	xsd:int	1	qb:DimensionProperty
hs:observedSpecies	taxmeon:TaxonInChecklist	1	qb:DimensionProperty
hs:refTime	xsd:date	1	qb:DimensionProperty
hs:season	hs:SeasonOfYear	1	qb:DimensionProperty
hs:weekOfYear	xsd:int	1	qb:DimensionProperty
qb:dataSet	qb:DataSet	1	qb:DimensionProperty
sdmx-attribute:nonsamplingErr	-	0..1	qb:AttributeProperty

Taulukko 5: Linkitetyn lintuhavaintoaineiston RDF Data Cube -mallin (`hs:haliasDSD`) komponentit.

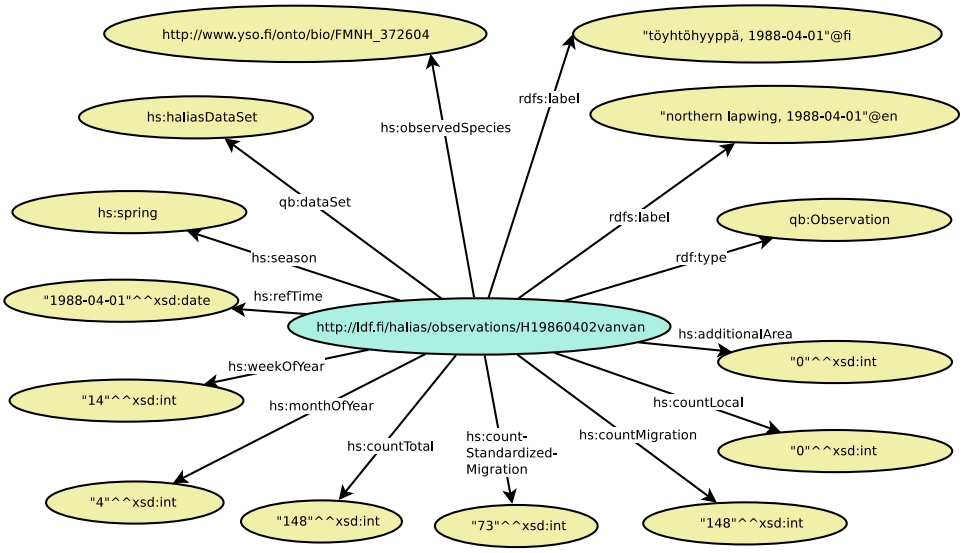
[6] `http://datahub.io/dataset/data-gov-uk-time-intervals`

Monia käytettyjä vastaavia termejä löytyy myös muista sanastoista, mutta näille on Data Cubea varten täytynyt tehdä Data Cuben mukaiset määritykset. Data Cubea käytettäessä määrittelyjoukko (domain) poikkeaa yleensä niiden ontologioiden määrittelyjoukosta, joita ei ole tehty erityisesti Data Cubea varten. Arvojoukkoon (range) Data Cube ei aseta mitään vaatimuksia. On mahdollista käyttää ulkopuolisten sanastojen käsitteitä, joille ei ole määritelty määrittelyjoukkoa. Itse määritellyille käsitteille on luotu englanninkieliset URI-tunnisteet, koska se edistää aineistojen jatkokäyttöä.

Pakollisia tietoja jokaiselle havainnolle on havaittu laji (`hs:observedSpecies`), päivämäärä (`hs:refTime`) sekä erilaiset vuorokauden lintumäärät.

Päivittäiset lintumäärät on esitettynä kokonaislukuina. Päivittäisten paikallisten lintujen (`hs:countLocal`) ja muuttavien lintujen (`hs:countMigration`) yhteenlaskettu päivän lintujen kokonaismäärä on ilmoitettu ominaisuudella `hs:countTotal`. Epäsäännöllisesti käytetty lisäalueen paikallisten lintujen määrä on ilmaistu ominaisuudella `hs:additionalArea`, jota on suositeltu käytettäväksi vain harvinaisille lajeille. Lisäalueen määrä puuttuu suuresta osasta havaintoja. Puuttuvat arvot ilmaistaan arvolla 0, koska tämä on merkitykseltään ja aineiston käytön kannalta sama kuin puuttuva arvo ja näin skeema pysyy Data Cube -mallin mukaisena. Esimerkki yhdestä havaintorivistä RDF-muodossa on esitetty verkkona kuvassa 4.

Kokonaislukujen tietotyyppinä on käytetty `xsd:int` -tietotyyppiä, joka on 32-bittinen etumerkillinen kokonaisluku. Tämä arvojoukko on riittävä tarvittaville muuttujille. Toinen varteenotettava vaihtoehto olisi `xsd:Integer` -tietotyyppi. Tällöin arvojoukkona olisi kaikki kokonaisluvut, mikä on ohjelmallisen käsittelyn kannalta hankaloittava tekijä, koska arvot voivat olla mielivaltaisen isoja ja vaatia mielivaltaisen määrän muistia. Ainoana etuna tästä tietotyypistä olisi se, että Turtle-kielelle serialisoidussa datassa ei tarvitsisi olla tietotyyppimäärittelyä kyseisillä arvoilla, mikä helpottaisi datan luettavuutta ihmisen näkökulmasta. Ihmisluettavuuden paraneminen on kuitenkin hyvin pieni hyöty, koska RDF-muotoinen data on tarkoitettu ensisijaisesti koneluettavaksi.



Kuva 4: Yhden lajin RDF-muotoinen havaintotieto yhdeltä havaintopäivältä.

5.3 Taksonomia ontologiana

Lintujen taksonomian pohjana on käytetty maailman lintulajit käsittävää AVIO-ontologiaa [TLKH13], jota on aiemmin käytetty mm. BirdWatch--mobiilipalvelussa [HAKT13]. Vaihtoehtona olisi ollut käyttää esim. Avibasen [LVG14] lajilistoja ja tunnisteita, mutta näitä ei ole saatavissa RDF-muodossa.

AVIO-ontologiasta tehtiin päivitetty versio nimeltä AVIO+, johon on lisätty Halias-aineistossa käytettyjä laajempia taksoneita, kuten lajipareja ja lajiryhmiä sekä generoitu taksonien lyhenteet. AVIO+ on rajattu Suomessa tavattuihin lajeihin.

AVIO ja AVIO+ käyttävät TaxMeOn-metaontologiaa, joka on muuttunut alkuperäisen AVIO-ontologian käyttämästä versiosta. AVIO+ päivitettiin vastamaan nykyistä versiota TaxMeOn:sta. Päivityksessä on poistettu joitain ominaisuuksia ja joidenkin nimi on muuttunut. Koska lajilistan taustalla olevista taksonomisista konsepteista ei ole tietoa helposti saatavilla, käytetään epämääräisemmin taksoniin viittaavaa `taxmeon:TaxonGeneral` kuvaamaan nimen takana olevaa merkitystä. Tieteelliset nimet ilmaistaan Darwin Core:n ominaisuudella `dwc:scientificName`.

AVIO+:n ja Halias-aineiston taksonomian välillä on joitain eroja. Halias-

aineiston taksonomia perustuu BirdLife Suomen ylläpitämään listaan Suomen lintulajeista ja AVIO+ perustuu *Association of European Records and Rarities Committees*:in ylläpitämään Euroopan lintulajien listaan. Listojen väliset erot rajoittuvat Halias-aineiston lajien osalta siihen, että jotkut lajit ovat eri suvuissa. Tämä vaikuttaa joidenkin sukutasolla tehtävien tarkastelujen tulokseen, mutta muuta haittaa tästä ei ole. Esimerkki tällaisesta lajista on jalohaikara, jonka tieteellinen nimi Halias-aineistossa on *Egretta alba* ja AVIO+:ssa *Ardea alba*.

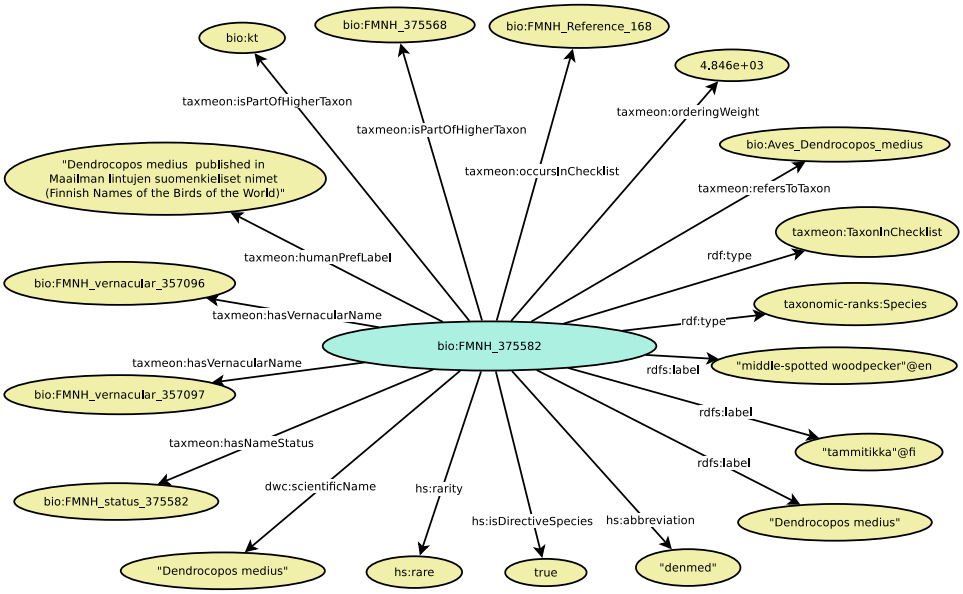
Halias-aineistossa käytetyt lajiparit ja lajiryhmät on lisätty AVIO+:aan niiltä osin kuin niille on olemassa hyvin dokumentoitu tulkinta [Uus06]. Lajiparit ja lajiryhmät on ripustettu osaksi AVIO+:n taksonihierarkiaa käyttämällä `taxmeon:isPartOfHigherTaxon` ominaisuutta sellaiseen taksoniin, joka käsittää kaikki lajiryhmän lajit, mutta mahdollisimman vähän muita taksoneita. Lisäksi lajit ja muut taksonit linkitetään samalla ominaisuudella näihin lajipareihin ja lajiryhmiin.

Halias-aineistossa havaitut taksonit on ilmaistu lyhenteinä, jotka ovat yleisesti käytettyjä ja niiden muodostamiseen on omat sääntönsä [Uus06]. Lajien lyhenteet ovat 6 merkkiä pitkiä koostuen ensin kolmesta suvun ilmaisevasta merkistä, joita seuraa 3 merkkiä, jotka ilmaisevat lajin. Sekä suvusta että lajiosasta otetaan ensisijaisesti kolme ensimmäistä kirjainta. Esimerkiksi peippo *(Fringilla coelebs)*, on lajilyhenteenä ilmaistuna Fri coe. Joillain lajeilla lyhenteiksi tulisi näin muodostettuna samat lyhenteet, joten näillä lajeilla käytetään lajiosan lyhenteenä kolmea viimeistä kirjainta. Esimerkiksi *Phylloscopus trochilus*, ja *Phylloscopus trochiloides*, ovat lyhenteinä ilmaistuna vastaavasti *Phy lus* ja *Phy des*. Suomessa käytettäviä lyhenteitä muodostettaessa tarvitsee ottaa huomioon vain Euroopassa ja sen lähialueilla tavatut lajit. Laajemmalla maantieteellisellä alueella tarkasteltuna päällekkäisten lajinimien vuoksi lyhenteet eivät enää vastaisi Suomessa käytettyjä lyhenteitä.

Käytännössä lyhenteet on generoitu AVIO-ontologiasta, mutta pienelle osalle on myös lisätty käsin lyhenne AVIO:n ja alkuperäisen havaintoaineiston välillä olevien taksonomisten eroavaisuuksien takia. Myös suvuille on generoitu lyhenteet lyhennesääntöjen mukaisesti, sekä on käsin lisätty lajiryhmien ja lajiparien lyhenteet.

Tieto lajien yleisyydestä Hangon Lintuasemalla on lisätty jokaiselle lajille. Yleisyys on karkeistettu luokkiin yleinen ja harvinainen. Tätä tietoa tarvitaan, jos halutaan käyttää lisäalueelta merkattuja summia, koska niitä tulisi käyttää vain harvinaisten lajien osalta. Yleisten lajien osalta niiden käyttäminen ei ole ollut säännönmukaista. Tämä ohjeistus saatiin aineiston haltijalta. Yleisyysluokkiin lajit on jaettu sen mukaan onko lajia havaittu Haliaksella yli 300 päivänä (yleinen), tai alle sen tai ei ollenkaan (harvinainen). Jakopiste on valittu melko mielivaltaisesti niin, että lajit jakautuvat oikean oloisesti. Yleisyystieto on lisätty taksoniontologiaan ominaisuudella `hs:rarity`

Kuva yhdestä AVIO-ontologian taksonista ja sen suhteista muihin käsitteisiin on kuvassa 5. Kyseiselle lajille (tammitikka) ei ole tuntomerkkiontologiassa tuntomerkkejä, mutta jos niitä olisi, ne näkyisivät kaarina taksonista yksittäisiin tuntomerkkeihin ominaisuudella `hs:hasCharacteristic`.



Kuva 5: Yksi AVIO+ -ontologian kuvaamista lintulajeista ja sen kaikki ominaisuudet.

Kyseistä lajia ei ole havaittu Hangon Lintuasemalla, mutta on huomionarvoista, että laji on osa lajiryhmää `bio:kt` eli ”keskikokoinen tikka”, jota

on käytetty lintuasema-aineistossa. Monimutkaisemmaksi asian tekee se, että havainnointiohjeiden [Han13b, Uus06] mukaan lintua määritettäessä otetaan huomioon vain Suomessa tavatut lintulajit, ellei ole erillistä syytä ottaa muitakin huomioon. Tammitikka on havaittu Suomessa ensimmäistä kertaa vuonna 2010, jolloin lajiryhmän `bio:kt` sekä suvun *Dendrocopos* (`bio:FMNH_375568`) merkitys voidaan tulkita muuttuneeksi kun laji on lisätty Suomessa tavattujen lajien listalle. AVIO+ -ontologiassa ei ole kuitenkaan otettu huomioon taksonien ja lajiryhmien ajallista muuttumista, vaan se pyrkii kuvaamaan nykyistä tilannetta mahdollisimman hyvin.

5.4 Tuntomerkkiontologia

Lajien tuntomerkkiontologia kattaa 245 Suomessa esiintyvää lajia ja se on yhteensopiva Luontoportti-palvelun[7] tuntomerkki-järjestelmän kanssa. Tuntomerkkiontologia on luotu aiemmassa tutkimuksessa [HAKT13] ja se koostuu tuntomerkeistä ryhmiteltynä OWL-hierarkiaksi. Tuntomerkkiontologiassa on 116 erilaista tuntomerkkiä, jotka on annotoitu AVIO-ontologian lajeille. Käsitteiden URI:t sisältävät muista käytetyistä sanastoista poiketen suomenkielisiä sanoja.

Tuntomerkkiontologia on muunnettu alkuperäisestä OWL-luokkahierarkiasta SKOS-sanastoksi, joka soveltuu paremmin tämän tyyppisille käsitteistöille. Ontologiassa käsitteet muodostavat hierarkian niin, että suppeammista käsitteistä on `skos:broader` -suhde laajempiin käsitteisiin. Esimerkiksi linnun kokoa ilmaisevat käsitteet `bc:suuri` ja `bc:hyvinpieni` omaavat `skos:broader` -suhteen käsitteeseen `bc:koko`, josta puolestaan on `skos:broader` -suhde käsitteeseen `bc:muotojakoko`, joka puolestaan on hierarkian ylimpiä käsitteitä.

5.5 Säähavaintoaineiston skeema

Säämuuttujien esittäminen on toteutettu ennakkotietojen ja kirjallisuuden perusteella olennaisiksi oletettujen lintujen muuttoon vaikuttavien säämuuttujien suhteesta. Tärkeimpänä vuorokaudenaikana on päivän valoisa aika, jolloin suurin osa havaittavasta muutosta tapahtuu painottuen aikaiseen

[7] `http://www.luontoportti.com/`

aamuun. Merkittäviä päivittäistä lintumuuttoa selittäviä säämuuttujia ovat tuulen nopeus ja suunta [Ale11]. Tuuliolot maanpinnan lähellä eivät välttämättä vastaa lainkaan tuulioloja ylempänä ilmakehässä, missä osa lintumuutosta tapahtuu [Ale11], mutta korkealla menevää lintumuuttoa ei voi juurikaan havainnoida maanpinnalta.

Tuulen suunnat ovat alkuperäisessä aineistossa ilmaistuina 10 asteen tarkkuudella ja nämä karkeistetaan 8-suuntaiseen asteikkoon, joka koostuu pääilmansuunnista ja väli-ilmansuunnista. Tällöin pääilmansuuntiin osuu 10 asteen tarkkuudella ilmaistuja ilmansuuntia 5 kappaletta, mutta väli-ilmansuuntiin niitä osuu vain 4. Tämä aiheuttaa vääristymää tuulien painottuessa jonkin verran pääilmansuuntiin. Tämä vääristymä kuitenkin on korjattavissa aineistoon tehtävissä hauissa normalisoimalla havaitut lintumäärät eri tuulen suuntien esiintyvyyden mukaan.

Tuuliolosuhteet ilmaistaan tuuli-instansseina, joista jokaisessa on kiinteä tuulen suunta ja nopeus pyöristettynä metreihin sekunnissa. Vuorokauden tuulet ilmaistaan neljällä eri ominaisuudella. Vakiohavainnointiajan tuulet on ilmaistu ominaisuudella `hs:standardWind`, auringonnousun ja -laskun väliset tuulet on ilmaistu ominaisuudella `hs:windDay` ja auringonnousua edeltävät sekä auringonlaskun jälkeiset tuulet vastaavasti ominaisuuksilla `hs:windPreSunrise` ja `hs:windPostSunset`.

Yhtä tuuliolosuhdetta voi alkuperäisessä aineistossa olla yhden päivän aika useita kertoja. Ne mallinnetaan RDF-muodossa kuitenkin niin, että jokainen tuuli-instanssi voi esiintyä ainoastaan kerran jokaisella neljästä eri tuuliominaisuudesta yhden vuorokauden aikana. Tämä johtuu siitä, että eri tuuliolosuhteisiin viitataan suoraan päiväkohtaisista aggregoiduista säähavainnoista ja jokainen subjekti-predikaatti-objekti -kolmikko voi esiintyä RDF-tietomallissa vain kerran. Jos aineisto mallinnettaisiin niin, että eri tuulioloista saadaan tietoon myös tieto siitä, kuinka monessa mittauksessa kyseistä tuulta on esiintynyt, saataisiin hieman tarkemmin vertailtua havaittuja lintumääriä suhteessa tuuliolosuhteisiin. On kuitenkin havaittu, että nykyisen mallinnustavan aiheuttama vääristymä on pieni, koska täsmälleen samoja tuuliolosuhteita esiintyy hyvin harvoin useita kertoja saman vuorokaudenosan aikana.

Visualisointeja varten aineistoa on aggregoitu ja tarkkuutta karkeistettu

`hs:weatherDataset` -datakuutiossa ja sitä kuvaavassa `hs:weatherDSD` kuutiomäärittelyssä, josta on taulukkomuotoinen esitys taulukossa 6. Jos datan pohjalta tehtäisiin päättelyä ohjelmallisesti, voisi olla perusteltua käyttää tarkempia arvoja. Osa lintuhavaintoaineiston kannalta olennaisista suureista ei ole kovinkaan yleiskäyttöisiä, kuten erilliset arvot sääoloista aamuvakiolta. RDF-muotoinen alkuperäisdata säämuuttujista on `hs:russaroDSD` kuutiomäärittelyssä ja sitä vastaavassa `hs:russaroDataset` -datakuutiossa. Taulukkomuotoinen esitys kuutiomäärittelystä löytyy taulukosta 7.

nimi	**arvojoukko**	**luokka**
hs:airPressure	xsd:double	qb:MeasureProperty
hs:cloudCover	xsd:double	qb:MeasureProperty
hs:haliasObservationDay	xsd:boolean	qb:AttributeProperty
hs:humidity	xsd:double	qb:MeasureProperty
hs:monthOfYear	xsd:int	qb:DimensionProperty
hs:rainfall	xsd:double	qb:MeasureProperty
hs:refTime	xsd:date	qb:DimensionProperty
hs:season	hs:SeasonOfYear	qb:DimensionProperty
hs:standardCloudCover	xsd:double	qb:MeasureProperty
hs:standardTemperature	xsd:double	qb:MeasureProperty
hs:standardWind	hs:WindObservation	qb:MeasureProperty
hs:sunriseTime	xsd:time	qb:AttributeProperty
hs:sunsetTime	xsd:time	qb:AttributeProperty
hs:temperatureDay	xsd:double	qb:MeasureProperty
hs:weekOfYear	xsd:int	qb:DimensionProperty
hs:windDay	hs:WindObservation	qb:MeasureProperty
hs:windPostSunset	hs:WindObservation	qb:MeasureProperty
hs:windPreSunrise	hs:WindObservation	qb:MeasureProperty
qb:dataSet	qb:DataSet	qb:DimensionProperty

Taulukko 6: Aggregoidun säähavaintoaineiston `qb:DataStructureDefinition` -määrittelyn komponentit. Jokaisen komponentin tulee esiintyä mallin mukaisessa havainnossa tasan kerran.

Alkuperäisdatasta puuttuvat arvot on ilmaistu RDF-muodossa `xsd:double` -tyypin arvolla `NaN`, joka tarkoittaa määrittelemätöntä numeroarvoa.

nimi	**arvojoukko**	**luokka**
hs:airPressure	xsd:double	qb:MeasureProperty
hs:cloudCover	xsd:double	qb:MeasureProperty
hs:humidity	xsd:double	qb:MeasureProperty
hs:observationTime	xsd:dateTime	qb:DimensionProperty
hs:temperature	xsd:double	qb:MeasureProperty
hs:wind	hs:WindObservation	qb:MeasureProperty
qb:dataSet	qb:DataSet	qb:DimensionProperty

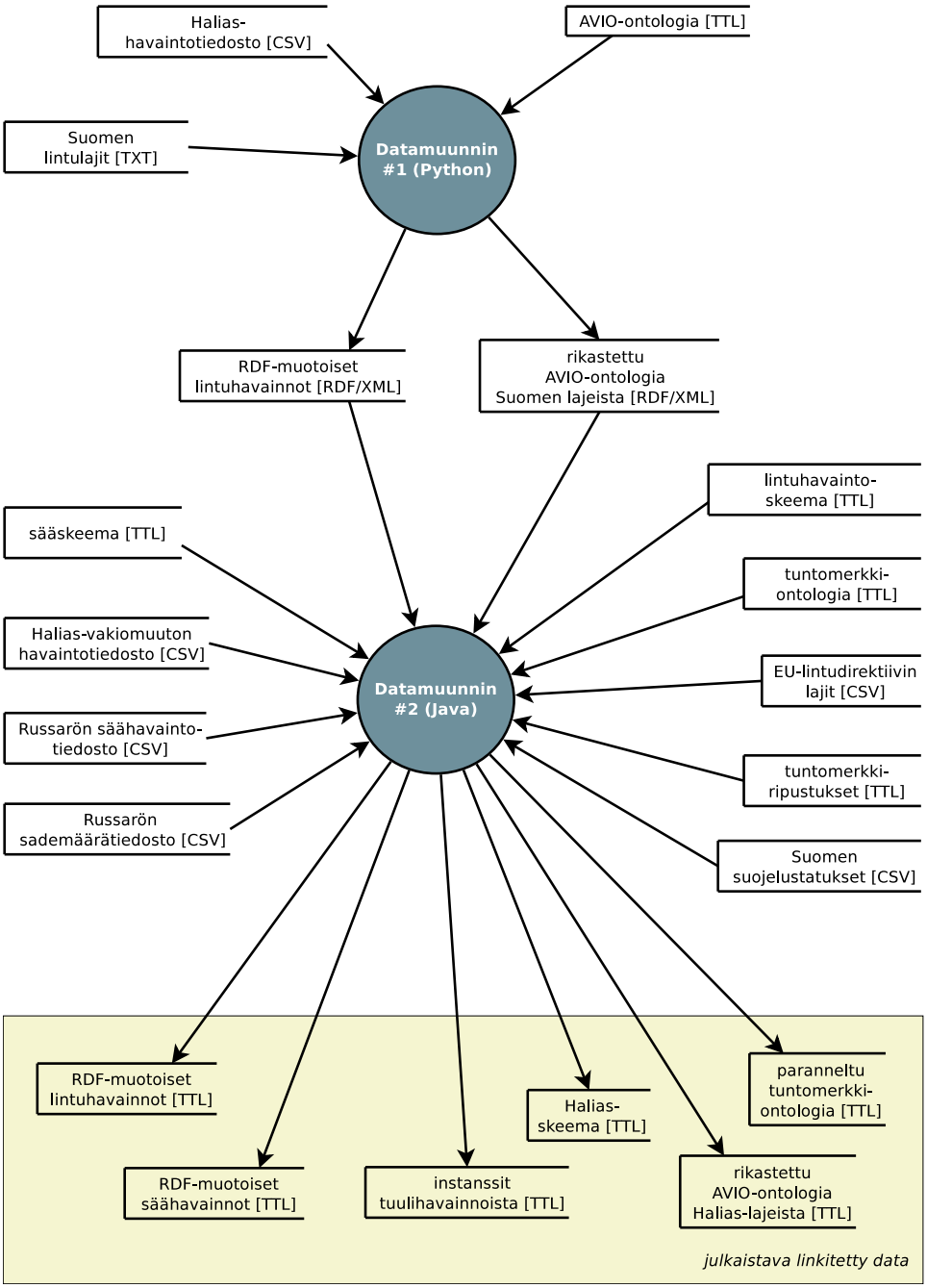
Taulukko 7: Alkuperäisten säähavaintojen `qb:DataStructureDefinition` -määrittelyn komponentit. Jokainen komponentti tulee esiintyä mallin mukaisessa havainnossa tasan kerran.

Puuttuviin arvoihin ei käy 0, koska sen käyttäminen vääristäisi aineistoa. Mittaustulosten lukuarvoja on `hs:weatherDataset` -datakuutiossa karkeistettu kokonaislukujen tarkkuuteen, mutta ne on silti ilmaistu `xsd:double` -arvojoukolla, jotta puuttuvat arvot saadaan näin helposti ilmaistua.

Linkitys eri datakuutioiden välillä tehdään päivämäärän perusteella. Lintuhavaintoaineiston ja sääaineiston Data Cube -malleissa käytetään ominaisuutta `hs:refTime` kuvaamaan havainnon päivämäärää. Alkuperäishavainnoista koostuvassa Data Cube -mallissa päivämäärä on esitetty havainnon kellonajan kanssa `hs:observationTime` -ominaisuudella.

6 Datamuunnos linkitetyksi dataksi

Datan muuntaminen alkuperäistiedostoista RDF-muotoon tehdään kahdessa vaiheessa. Muunnosprosessin kulkua on havainnollistettu kuvassa 6. Ensimmäisen vaiheen tekee Python-ohjelma, jonka lopputuotteena on RDF-muotoista dataa. Tätä dataa jatkokäsitellään vielä Java-ohjelmalla. Syynä tähän kaksivaiheiseen muunnokseen on se, että ohjelma oli tarkoitus toteuttaa alun perin Pythonilla, mutta tämä osoittautui mahdottomaksi ohjelman käyttämän RDFLib-kirjaston huonon skaalautuvuuden takia. Helpointa oli sitten pitää toteutetut osat muunnoksesta Python-koodina ja tehdä jatkokäsittelyä Javalla ja Jena-kirjastolla, jonka skaalautuvuus osoittautui erinomaiseksi.



Kuva 6: Kaksivaiheinen datan muunnosprosessi sekä muunnosohjelmissa käytetyt syötteet ja tulosteet, joiden tiedostomuodot ovat hakasulkeissa.

Ohjelmien rakenne on elänyt projektin kuluessa. Koodin määrän kasvaessa ja ohjelmien monimutkaistuessa on myös kirjoitettu yksikkö- ja integraatiotestejä sekä Python- että Java-komponenteille. Integraatiotestit lukevat sisään pieniä esitysmuodoltaan alkuperäisdataa vastaavia tiedostoja ja varmistuvat, että muunnosprosessi tuottaa näiden pohjalta oikeanlaiset lopputulokset. Ohjelmat myös suorittavat muunnosvaiheessa lintuhavaintoaineistolle joitain validointeja eli oikeellisuustarkistuksia.

Kaikki aineistojen muuntamisessa käytetty ohjelmakoodi on julkaistu avoimesti Github-palvelussa ja ne löytyvät osoitteesta `https://github.com/razz0/Halias-data-conversion`. Ohjelmien ajaminen kuitenkin vaatisi myös alkuperäiset havaintoaineistot, joiden käyttöehdot eivät tällä hetkellä mahdollista julkista levitystä.

6.1 Python-muunnosohjelma

Python-koodi käsittelee lintuhavaintoaineistoa ja tuottaa RDF-muotoisia havaintoinstansseja sekä rikastetun AVIO+ -ontologian Suomen lintulajeista. Ohjelman rakenne on suoraviivainen ja sen toiminnallisuudet on toteutettu käyttäen funktioita.

Ohjelma lukee ensin AVIO+ -taksoniontologian tiedostosta, muokkaa sen rakennetta vastaamaan nykyistä versiota TaxMeOn:sta, lisää taksoneille generoidut nimilyhenteet ja kirjoittaa taksoniontologiasta kaksi erilaista versiota tiedostoihin. Toinen tiedosto on rajattu Suomessa havaittuihin lajeihin ja toisessa on kaikki maailman lajit. Sen jälkeen ohjelma lukee sisään lintuhavaintodatan CSV-tiedostosta ja prosessoi havainnot RDF-muotoisiksi havaintoinstansseiksi, jotka sitten tallennetaan viiteen RDF-tiedostoon. Syynä tähän pilkkomiseen on etupäässä vähentää ohjelman muistin käyttöä. Muistin loppuminen muunnosvaiheessa on johtanut ohjelman ajon keskeytymiseen. Tästä syystä kaikkea aineiston käsittelyä ei tehdä Python-ohjelmalla.

Python-ohjelmaa on ajettu Python 2.7:llä ja se käyttää Pythonin standardikirjastojen lisäksi RDFLib-kirjastoa (versio 4.1.2) ja iso8601-kirjastoa (versio 0.1.10). RDFLib on Python-kirjasto RDF:n lukemiseksi, käsittelemiseksi ja sarjallistamiseksi. Iso8601 on kirjasto ISO 8601 -muotoisten päivämäärien tulkitsemiseen.

Itse Python-muunnosohjelma koostuu seuraavista tiedostoista:

- **observation_generator.py** on ohjelman päätiedosto,
- **halias_helpers.py** sisältää kokoelman apufunktioita muunnosprosessia varten sekä Validator-luokan validointiluokan datamuunnoksessa esiin tulevien validointivirheiden käsittelyyn ja
- **test_observation_generator.py** sisältää yksikkötestejä ja funktionaalisia testejä muunnosprosessin eri vaiheiden oikeellisuuden tarkastamiseksi.

Muunnosprosessi ajetaan käynnistämällä observation_generator.py:n ajo Python-tulkilla. Ohjelma hyväksyy parametrit:

-h antaa tietoa ohjelman hyväksymistä parametreista ja

-d ajaa ohjelman normaalisti, mutta mitään ei kirjoiteta levylle.

Ohjelma lukee ensin taulukkomuotoisesta lintuhavaintoaineiston tiedostosta kaikkien havaittujen taksoneiden havaintomäärät. Tämän jälkeen luetaan sisään AVIO-ontologia ja jotain taksonomisista muutoksista johtuvia käsin tehtyjä lisäyksiä siihen. Lajeille generoidaan lintulajien tieteellisten nimien lyhennyssäännön perusteella lyhenteet. AVIO-ontologian taksoneita muokataan poistamalla niiltä joitain tarpeettomia ominaisuuksia, lisäämällä lajinimet kokonaisuudessaan `RDFS:label` ja `dwc:scientificName` -ominaisuuksina ja lisäämällä lajinimien lyhenteet ominaisuudella `hs:abbreviation`. AVIO:n OWL-luokkahierarkia korvataan hierarkialla, joka käyttää `TaxMeOn:isPartOfHigherTaxon` -ominaisuutta. Ne taksonit poistetaan, joita ei tavata Suomessa ja joiden alempia taksoneita ei tavata Suomessa. Näiden muutosten jälkeen kirjoitetaan Suomen lajistoon rajattu AVIO+ -ontologia tiedostoon.

RDF-graafit sarjallistetaan tiedostoihin RDF/XML-kielellä. Tämän jälkeen käydään läpi Halias-aineiston alkuperäistiedostoa rivi kerrallaan ja kirjoitetaan havainnot RDF-graafiin. Samalla havaintoja validoidaan ja kaikista validointivirheistä tulostetaan ilmoitus. Lopuksi tulostetaan kooste validointivirheiden lukumääristä. Havainnot jaetaan 100 000 havaintorivin välein eri tiedostoihin. Vuoden 2009 havainnot jätetään pois, koska tälle

vuodelle ei ole vielä saatavissa aamuvakion havaintoja. Tämä ero aineistojen aikajänteissä tulisi muuten ottaa huomioon joissain aineistoon tehtävissä kyselyissä. Yksittäisen havaintorivin käsittely koostuu pääpiirteissään seuraavista vaiheista:

1. luetaan havaintorivi,
2. jos havaintorivin taksonilyhennettä ei löydy taksoniontologiasta, näytetään validointivirhe ja hypätään käsittelemään seuraavaa havaintoriviä,
3. luodaan havainnolle URI päivämäärän ja taksonilyhenteen perusteella, mutta jos URI on jo käytetty, näytetään validointivirhe ja hypätään käsittelemään seuraavaa riviä,
4. lisätään RDF-graafiin havaintoinstanssi luokasta qb:Observation ja
5. lisätään `qb:dataSet` sekä muut tiedot havaintoinstanssille nimiavaruuden `hs` ominaisuuksina.

Muunnosohjelman oikeellisuutta varmistavat automaattiset testit käyttävät Pythonin standardikirjastoon kuuluvaa unittest-kirjastoa. Testit voidaan ajaa komentoriviltä, joko komennolla:

```
python -m unittest test_observation_generator
```

tai selkeämmällä tulosteella käyttäen nose-kirjastoa, jolla testit ovat tavallisesti ajettavissa komennolla:

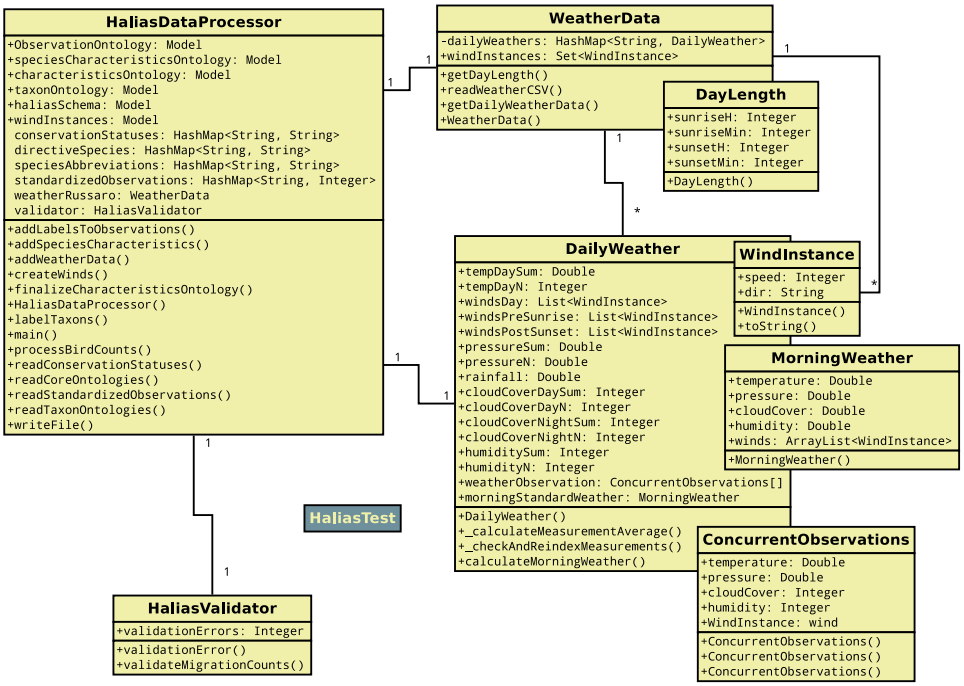
```
nosetests test_observation_generator.py
```

6.2 Java-muunnosohjelma

Java-ohjelma lukee sisään Python-ohjelman tuottamat RDF-tiedostot ja jatkokäsittelee näitä. Tässä vaiheessa luetaan ja käsitellään säähavaintodata ja luodaan näistä kaksi RDF-graafia.

Java-muunnosohjelma on kehitetty Java 6.0:lla. Ohjelma käyttää Javan standardikirjastojen lisäksi Jena 2.11.0 ja Sunrise/SunsetLib -kirjastoja. Jena on kehittynyt kirjasto RDF-datan käsittelyyn. Sunrise/SunsetLib-kirjasto laskee minkä tahansa päivän auringonnousun ja -laskun ajankohdat mielivaltaiselle maapallon pisteelle.

Java-ohjelman luokkarakenne on kuvattu karkealla tasolla kuvan 7 UML-kaaviossa. Kaaviossa näkyy luokat, niiden väliset suhteet ja niiden attribuutit sekä metodit. Metodien parametrejä ja paluuarvoja ei ole kuvattu. Automaattisia testejä sisältävän HaliasTest-luokan riippuvuudet ja rakenne ei ole UML-kaaviossa mukana.



Kuva 7: UML-luokkakaavio Java-muunnosohjelmasta

Java-muunnosohjelma ajetaan esim. komentoriviltä ajamalla Java-tulkilla HaliasDataProcessor-luokan main-metodi. Java-koodin yksikkö- ja integraatiotestit ovat luokassa HaliasTest. Testit on toteutettu käyttäen junit-kirjastoa.

Ohjelman alkuvaiheessa luetaan sisään tarvittavat ontologiat. Lisäksi luetaan muistiin aamuvakioiden lintuhavainnot sekä luetaan ja käsitellään säähavainnot kahdesta taulukkomuotoisesta tiedostosta. Säämuuttujista prosessoidaan muistiin kolmen tunnin välein tehdyt säähavainnot riittävällä tarkkuudella sekä aggregoidaan vuorokauden eri osien kuten aamuvakion ajan ja päivän valoisan ajan keskiarvoja. Säähavaintoaineistojen pohjalta luodaan seuraavat RDF-graafit omiin tiedostoihinsa:

1. ontologia havaituista tuulista,

2. aggregoidut säähavaintoaineistot RDF Data Cube -muodossa ja

3. alkuperäiset säähavainnot RDF Data Cube -muodossa.

Ohjelma tekee Python-muunnosohjelman luomaan AVIO+ -ontologiaan seuraavia muutoksia. Taksoneille lisätään Suomen suojelustatus ja esiintyminen EU-lintudirektiivissä, jos ne löytyvät. Nämä luetaan sisään tiedostoista, joihin ne on poimittu internetissä saatavilla olevista listauksista. Taksoneille lisätään kansankieliset englannin- ja suomenkieliset nimet ominaisuudella `rdfs:label`. Lisätään lajeille tuntomerkkien annotoinnit.

Python-muunnosohjelman luomat viisi RDF-havaintotiedostoa käydään läpi. Näistä jokaiselle tehdään seuraavat toimenpiteet:

1. havaintoihin lisätään `rdfs:label` -ominaisuudet suomeksi ja englanniksi, jotka sisältävät taksonin kansankielisen nimen ja päivämäärän,

2. havaintoinstansseihin lisätään vakiomuuttohavainnot ja tyypitetään aiemmin luodut havaintomäärät,

3. lisätään viikkojen ja kuukausien numerot havaintoinstansseihin ja

4. lisätään aggregoitujen säähavaintojen RDF-graafiin tieto siitä, onko kyseisenä päivänä havainnoitu lintuja. Tämä tieto lisätään, jotta saadaan nopeutettua SPARQL-kyselyjä.

Lopuksi luodaan Halias-skeema yhdistämällä aiemmin luodut sääskeema ja lintuhavaintoskeema. Tuntomerkkiontologiaa myös muokataan.

Java-ohjelma sarjallistaa RDF-graafit tiedostoihin käyttäen Turtle-kieltä. Tässä vaiheessa graafit ovat lopullisessa muodossaan ja ne voidaan laittaa saataville johonkin palveluun.

Lopullisessa muodossaan lintuhavaintodata koostuu 438 471 havainnosta ja säähavaintodata 12 051 havainnosta. Luodut linkitetyt aineistot koostuvat yhteensä 6 793 043 subjekti-predikaatti-objekti -kolmikosta.

6.3 Validointi

Alkuperäistä lintuhavaintodataa on siirretty digitaaliseen muotoon talkootyönä ja tässä käsittelyssä syntyy helposti erilaisia virheitä. Datasta löytyikin rakenteellisia virheitä muunnettaessa sitä linkitetyksi dataksi. Näistä virheistä osa on löytynyt muunnosvaiheessa, osa dataa visualisoidessa ja osa on selvinnyt erillisissä datan validoinneissa. Joitain rakenteellisia virheitä jäisi myös kiinni datan muunnosvaiheessa ilman erillisiä tarkastuksia, koska ne aiheuttaisivat virhetilanteen muunnosohjelmassa.

Aineistojen muunnosvaiheeseen on tehty validointeja lintuhavaintoaineiston rakenteellisten virheiden selvittämiseksi. Löydetyistä virheistä lisätään tieto havaintoon käyttäen `sdmx-attribute:nonsamplingErr` ominaisuutta. Seuraavia validointivirheitä tulee muunnosvaiheessa:

- samalla laji ja päivämäärä -yhdistelmällä on jo olemassa havainto,
- havaittu aamuvakion muuttajasumma suurempi kuin koko päivän muuttajasumma ja
- taksonilla monitulkintainen lyhenne.

Samoilla laji ja päivämäärä -yhdistelmillä on noin 350 toisteista havaintoriviä, joista suurimmassa osassa on eriävät havaintosummat. Jokaisella laji ja päivämäärä -yhdistelmällä tulisi olla datassa vain yksi havaintorivi, eikä voida tietää mikä löytyneistä toisteisista riveistä on oikea. Tätä virhettä ei ole mahdollista korjata muunnosvaiheessa vaan se tulisi korjata tarkastamalla havainnot lintuaseman alkuperäisistä paperisista havaintolomakkeista. RDF Data Cube ei salli samoja arvoja havaintojen `qb:dimension` -ominaisuuksille, joten näitä ei voida muuntaa RDF-muotoon ja jättää datan käyttäjän päätettäväksi mitä tehdä näille. Mahdollisia toimenpiteitä tälle muunnosvaiheessa on ainakin

- jättää kaikki toisteiset havainnot pois,
- jättää kaikki havainnot pois toisteisten havaintojen päivämääriltä,
- käyttää jotakin toisteisista havainnoista ja poistaa muut tai
- käyttää toisteisten havaintojen lukumäärien keskiarvoja.

Näistä jokainen aiheuttaa jonkinlaista vääristymää aineistoon. Muunnoksessa päädyttiin käyttämään toisteisista havainnoista ensimmäistä ja jättämään huomiotta loput. Näitä on kuitenkin pieni osuus ($= \frac{350}{445765} \approx 0,08\%$) koko aineistosta, eikä niiden poistaminen vaikuta aineiston käytettävyyteen.

Aamuvakion muuttajasumma on virheellisesti koko päivän muuttajasummaa suurempi joillain havaintoriveillä, vaikka aamuvakion havaittu muutto tulisi olla laskettuna mukaan koko päivän muuttoon.

Taksonin monitulkintainen lyhenne estää myös havainnon luomisen. Osa näistä on tapauksia, joissa taksonilla on epäselvästi tulkittava lyhenne, joka ei noudata lyhennesääntöjä. Esimerkiksi *Ste sp.* voi tarkoittaa joko *Sterna* tai *Stercorarius* suvun lajia, mutta varmuudella ei voi sanoa kumpaa on tarkoitettu. Myös kahden eri lajin risteymäksi ilmoitetut havainnot jätetään nyt tarkoituksella pois muunnosvaiheessa, koska on epäselvää miten nämä tulisi ilmaista ja näitä on aineistossa hyvin vähän. Puuttumaan jääviä havaintoja on kaikkiaan hyvin pieni määrä kokonaisuudesta — yhteensä 13 havaintoriviä sisältäen 8 eri nimilyhennettä ja 15 havaittua lintuyksilöä.

Näiden lisäksi tehdään joitain tarkistuksia, kuten päivämäärän muotoilun tarkistaminen, joista kaikki havainnot kuitenkin selviävät ilman virhettä. Yhdellä havaintorivillä esiintyvä lukumäärä 2+3 tulkitaan erillistapauksena luvuksi 5.

RDF Data Cube -sanasto määrittelee validointeja SPARQL-kyselyinä [CR14], joilla voidaan varmistua, että data noudattaa sanastoa. Validoinneissa on kuitenkin joitain ongelmia. Data Cube -sanaston validointi *IC-3*:ssa tulisi `qb:componentProperty` korvata `(qb:componentProperty|qb:measure)`:llä, jotta validointi kelpuuttaisi monia itse standardissa mainittuja esimerkkejä sekä linkitetyn Halias-aineiston. Standardissa määritellyn `qb:measure` ominaisuuden yläominaisuus on `qb:componentProperty`, jota alkuperäinen validointi ei huomioi.

Data Cube -sanaston validointi *IC-12* validoi sen, että jokaisella havainnolla on yksilöivät dimensiot, eli samoilla dimensioiden arvoilla ei löydy useita havaintoja. Validointi on hyödyllinen, mutta huonon skaalautuvuuden takia käyttökelvollinen ainoastaan pienillä aineistoilla. Kysely vertailee n havaintoa käsittävällä aineistolla havaintoja keskenään $n(n-1) = n^2 - n$ kertaa. Esimerkiksi lintuhavaintodatan käsittävä `hs:haliasDataSet` -aineisto sisäl-

tää 438,471 havaintoa, jolloin validointi tekisi yli 192 miljardia havaintojen vertailua.

Korjaamalla validointi *IC-3* edellä mainitulla tavalla ja jättämällä *IC-12* pois, julkaistavat linkitetyt aineistot läpäisevät validoinnit.

RDF Data Cube -muotoisten aineistojen validointiin on olemassa online-palvelu *COMPUTEX Validation Service* [LR13], joka löytyy osoitteesta `http://computex.herokuapp.com`. Julkaistut aineistot ovat palvelulle liian suuria, mutta pienten otosten validointi tällä onnistuu.

6.4 Datajulkaisu

Data on julkaistu Linked Data Finland (LDF) -alustalla [HTAM14], joka on osittain automatisoitu julkaisukanava laadukkaasti kuvailluille linkitetyille aineistoille. LDF-alustan linkitettyjä aineistoja tarjoilee Fuseki, joka on Apachen kehittämä Jena-kirjaston päällä toimiva SPARQL-palvelin [Apa14].

Datajulkaisun kuvailut ovat nähtävissä palvelussa osoitteessa `http://www.ldf.fi/dataset/halias/`. Aineisto on käytettävissä SPARQL-rajapinnan kautta, mutta aineistoon pääsyä on rajoitettu havaintoaineistojen käyttöehtojen takia niin, että SPARQL-rajapinnan käyttö vaatii kirjautumisen käyttäjätunnuksen ja salasanan avulla. Suojaus on toteutettu Apache-palvelinohjelmalla.

Julkaistujen aineistojen käyttämä Halias-skeema on kuvailtu LODE-palvelulla [PSV12], joka on sovellus OWL-ontologioiden ja muiden RDF-sanastojen automaattiseen dokumentointiin. Skeeman kuvailu löytyy osoitteesta `http://www.ldf.fi/dataset/halias/halias_schema.html`.

Vaikka julkaistut aineistot eivät ole avoimesti saatavilla, niiden saaminen tutkimuskäyttöön onnistuu pyytämällä tunnuksia datajulkaisun sivulla olevien ohjeiden mukaisesti.

7 Aineistojen visualisointi

Tässä luvussa käsitellään linkitettyjen havaintoaineistojen visualisointia.

On olemassa järjestelmiä lintuhavaintojen analysointiin, kuten eBird[8],

[8] `http://ebird.org/`

joka tarjoaa mahdollisuuden tehdä erilaisia havaintojen visualisointeja, ja jossa voi vertailla eri lajien tilastoja toisiinsa. Samankaltaisia järjestelmiä löytyy maailmalta useita, mutta visualisointimahdollisuudet ovat usein rajallisemmat. Yksi esimerkki on suomalainen Tiira[9], joka on avoin palvelu lintuhavaintojen ilmoittamista ja selaamista varten, mutta joka ei tarjoa minkäänlaisia datan visualisointeja tai muita analyysityökaluja.

Valmiita ja avoimesti käytettäviä visualisointeja löytyy aineistoille, jotka käyttävät RDF Data Cube -sanastoa [SAB+12, MHS+13]. Näiden käyttäminen ei kuitenkaan suoraan onnistu, koska visualisoitavien aineistojen SPARQL-rajapinnat vaativat salasanalla kirjautumisen, eikä tämä ole niissä tuettuna.

7.1 Visualisointipalvelu

Havaintoaineistojen visualisointia varten kehitettiin prototyyppi linkitettyjen aineistojen visualisointipalvelusta. Visualisointipalvelulla pyritään näyttämään käyttäjälle kiinnostavia aineistojen välisiä yhteyksiä.

Visualisointipalvelu on toteutettu HTML-sivuina, jotka visualisoivat WWW-selaimessa dataa käyttäen JavaScript-kirjastoja Sgvizler[10] [Ska12] ja D3[11] sekä JavaScript-ohjelmistokehystä (framework) AngularJS[12]. Visualisointeihin käytetty data haetaan SPARQL-rajapinnasta Ajax-kyselyillä. Koska SPARQL-rajapinta on suojattu salasanalla, visualisointeja ei pääse käyttämään ilman kirjautumista.

SPARQL-kyselyt käytettyihin aineistoihin ovat melko hitaita, johtuen aineistojen suuresta koosta. Tässä tutkielmassa esitettyjen visualisointien SPARQL-kyselyt tarvitsevat noin 5-70 sekuntia aikaa saadakseen visualisoinnin käyttämän datan LDF-alustan RDF-tietovarastosta. Nämä ovat ongelmallisen pitkiä aikoja visualisointipalvelun käyttäjän kannalta. Syynä hitauteen on RDF-lauseiden suuri määrä aineistoissa.

[9] `http//tiira.fi/`
[10] `http://dev.data2000.no/sgvizler/`
[11] `http://d3js.org/`
[12] `http://angularjs.org/`

7.2 Geneerinen SPARQL-visualisointi

Yksinkertaisiin visualisointeihin käytetään LDF.fi -portaalin tarjoamaa mahdollisuutta visualisoida SPARQL-kyselyiden tuloksia Sgvizlerin avulla, joka käyttää datan visualisointiin *Google Chart API*:a[13]. Visualisointi on hyvin joustava, mutta vaatii sen, että käyttäjä kirjoittaa SPARQL-kyselyitä itse. Käyttäjä voi valita useista Google Chart API:n erilaisista kuvaajatyypeistä, joista viivakuvaaja on tavallisin. Visualisoinnit ovat interaktiivisia siten, että käyttäjä näkee kuvaajan pisteiden tarkat arvot viemällä kursorin niiden päälle.

Yksinkertainen esimerkki datan visualisoinnista viivakuvaajalla on kuvassa 8, joka näyttää kuvan palvelusta valikoineen. Kuvaaja esittää kuinka monena päivänä kunakin vuonna on lintuasemalla havainnoitu. Havainnointiaktiivisuus on ollut hyvin vaihtelevaa, mutta tällä vuosituhannella jo hyvin aktiivista. Useana vuotena on havaintoja 365:nä päivänä ja karkausvuonna 2008 jopa 366:na päivänä.

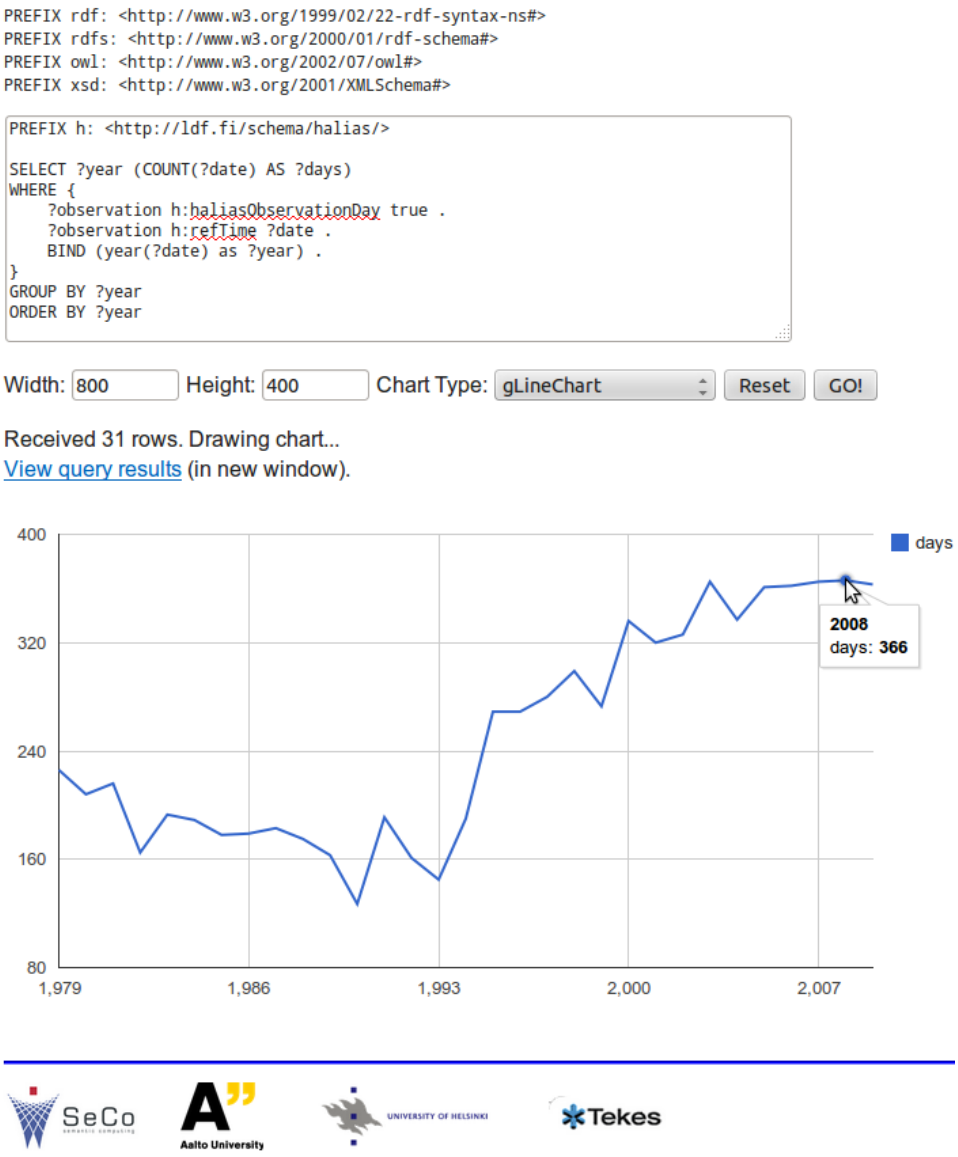
7.3 Tuulivisualisointi

Tuulivisualisointi näyttää tuuliolojen ja lintumäärien suhteen valittuna vuodenaikana. Kuvaajan idea on peräisin projektiin osallisena olleen biologin tutkimuskysymyksestä, jossa häntä kiinnosti miten paikalliset tuuliolot vaikuttavat lajien tai lajiryhmien muuton voimakkuuteen. Tätä varten kehitetty kuvaaja näyttää lajikohtaiset lintumäärät erilaisissa tuulioloissa valittuna vuodenaikana. Visualisointi piirretään D3-kirjastoa käyttävällä Circular heat chart[14] -kuvaajalla. Visualisointi käyttää AngularJS:ää sivun luomiseen ja päivittämiseen.

Käyttäjä kirjoittaa taksonin nimen (tieteellinen, englanninkielinen tai suomenkielinen) tai taksonin tieteellisen nimen lyhenteen ja valitsee haetaanko myös alemmilla taksoneilla. Ajallinen rajaus valitaan 3kk jaksoihin jaetuista vuodenajoista, missä talvi kattaa kuukaudet joulukuusta helmikuuhun. Valinnan voi tehdä vielä eri päiväkohtaisista lintumääristä: muuttajat, paikalliset vai kokonaismäärät. Käytettäessä hakua myös alemmilla takso-

[13] `https://developers.google.com/chart/`

[14] `http://prcweb.co.uk/lab/circularheat/`



Kuva 8: Kuvaaja näyttää vuosittaisten havainnointipäivien lukumäärän Hangon Lintuasemalla. Kuvan yläosassa näkyy käytetty SPARQL-kysely ja visualisointikäyttöliittymän valinnat.

neilla, haku hyödyntää hierarkista taksoniontologiaa ja hakee transitiivisesti kaikki taksonit, joista löytyy reitti valittuun taksoniin seuraamalla `taxmeon:isPartOfHigherTaxon` -ominaisuuksia.

Lintumäärät summataan jokaista kokonaislukuun pyöristettyä tuulen

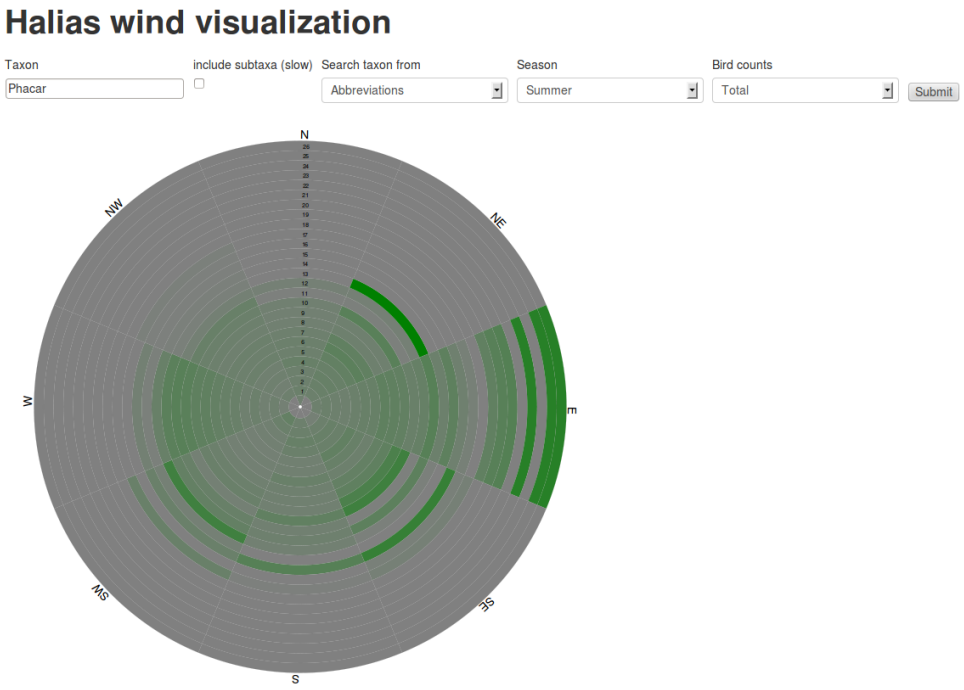
nopeuden ja tuulen suunnan yhdistelmää kohden, joilla haettua taksonia on valitulla aikarajauksella havaittu koko aineistossa. Summat normalisoidaan jakamalla ne ko. tuuliolosuhteen esiintymislukumäärällä aikarajauksen sisällä. Kuvaaja siis näyttää keskimääräistä lintumäärää niiltä päiviltä, jolloin on esiintynyt kyseistä tuulta. Tuuliolosuhteet otetaan aamuvakion ajalta, jos on valittuna vakiomuuttosummien visualisointi, tai päivän koko valoisalta ajalta muussa tapauksessa.

Tuulensuunnat esitetään pää- ja väli-ilmansuuntina, jotka on nimetty käyttäen tavallisia englanninkielisiä lyhenteitä, jotka ovat pohjoisesta alkaen myötäpäivään: N, NE, E, SE, S, SW, W ja NW. Etäisyys kuvaajan keskustasta kuvaa tuulennopeutta, jonka yksikkönä on metriä sekunnissa ja asteikko on merkittynä suoraan ylöspäin keskustasta. Kuvaajan väriskaala normalisoidaan lineaarisesti erilaisten tuulien esiintyvyyksien mukaan niin, että täysin harmaalla värillä lintuja ei ole havaittu kyseisellä tuulella ja syvin vihreän sävy kuvaa aina suurinta havaittua lintumäärää.

Kuvaajalla voidaan tarkastella etenkin sitä, minkälaisissa tuuliolosuhteissa muuttolintulajit muuttavat keväällä ja syksyllä. Kuvaajan tulkintaa hankaloittaa se, että eri tuuliolosuhteissa myös lintujen havaittavuus Hangon lintuasemalla vaihtelee. Ilmiö on erilainen eri vuodenaikoina ja eri lintulajeilla riippuen mm. lajin koosta, muuttoreiteistä ja muuttokorkeudesta. Joillain lajeilla, kuten kurjella, on kuitenkin selvästi nähtävissä muuton kannalta hyvien tuuliolosuhteiden suosiminen muuttoaikoina.

Kuva 9 visualisoi kesäkuukausina eli kesä-, heinä- ja elokuussa havaittuja merimetsoja suhteessa vallinneisiin tuuliolosuhteisiin. Lintumäärinä on käytetty päivittäistä paikallisten ja muuttavien lintujen summaa. Kuvassa on yläosassa näkyvissä visualisointipalvelun käyttöliittymän valinnat kyseiselle haulle. Datan hakemiseen käytetty SPARQL-kysely on liitteessä 1.

Kuvassa näkyvä voimakas esiintyminen itätuulilla voisi selittyä sillä, että Itämerellä pesivien merimetsojen syysmuutto on voimakasta jo elokuussa [LV00], jolloin Hangon itäpuolella sijaitsevilta pesimäalueilta mahdollisesti tulee lintuja myötätuulen siivittämänä Hankoon. Muita selkeitä trendejä kuvasta ei ole helposti nähtävissä ja erot esiintyvyydessä eri tuulilla mahdollisesti heijastelevat lähinnä eri tuuliolosuhteiden esiintyvyyttä aineistossa.



Kuva 9: Kesäinen merimetsojen esiintyminen erilaisissa tuulioloissa. Harmaalla värillä merkityillä tuuliolosuhteilla lintuja ei ole havaittu ja syvimmällä vihreällä merkityillä on havaittu suurimmat määrät.

8 Tulosten arviointi

Tässä luvussa tarkastellaan tutkielman tuloksia. Tutkielman alussa esitettyihin tutkimuskysymyksiin löytyneitä ratkaisuja esitellään aliluvuissa.

8.1 Lisäarvoa luontohavaintojen linkittämisestä

Millaista lisäarvoa lintuhavaintoihin on mahdollista saada yhdistämällä niihin säädataa ja lajitietoa?

Säätila on merkittävä yksittäisten päivien lintumuuton voimakkuutta selittävä tekijä. Yhdistämällä lintuhavaintoihin havaintoajan säätietoa, voidaan saavuttaa parempi ymmärrys itse lintuhavaintoaineistosta. Tämä mahdollistaa tarkempien päätelmien tekemisen aineiston pohjalta.

Linkittäminen taksoniontologiaan mahdollistaa taksonomisten konseptien käyttämisen sen sijaan, että käytettäisiin jokaiselle taksonille vain tieteel-

listä nimeä. Taksonomisilla konsepteilla on mahdollista ilmaista taksonien rajaukset, kun taas taksonien nimien tulkinta taksonien rajauksiksi voi olla vaikeaa. Taksoniontologian hierarkisen rakenteen takia voidaan yhdellä SPARQL-kyselyllä hakea taksonin ja sen kaikkien alempien taksonien havainnot. Taksoniontologiaan on helppo liittää lajikohtaista lisätietoa kuten uhanalaisuusluokituksia ja lajien tuntomerkkejä.

8.2 Miten julkaista havaintoaineistoja linkitettynä datana

Miten biologisista havainnoista koostuvaa tutkimusdataa kannattaisi julkaista linkitettynä datana ja rikastaa muilla aineistoilla, jotta lisäarvo on parhaiten saavutettavissa?

Biologisten havaintoaineistojen tehokas hyödyntäminen vaatii standardoituja tapoja aineistojen yhdistämiseen. Aineistojen julkaiseminen linkitettynä datana mahdollistaa paremman yhteentoimivuuden [RJS11]. Tämä lähestymistapa mahdollistaa myös valmiiden työkalujen käyttämisen datan käsittelyyn ja visualisointiin [SAB+12]. Nämä hyödyt ovat edelleen suuremmat käytettäessä aineiston rakenteen määrittelyyn esimerkiksi RDF Data Cube sanastoa.

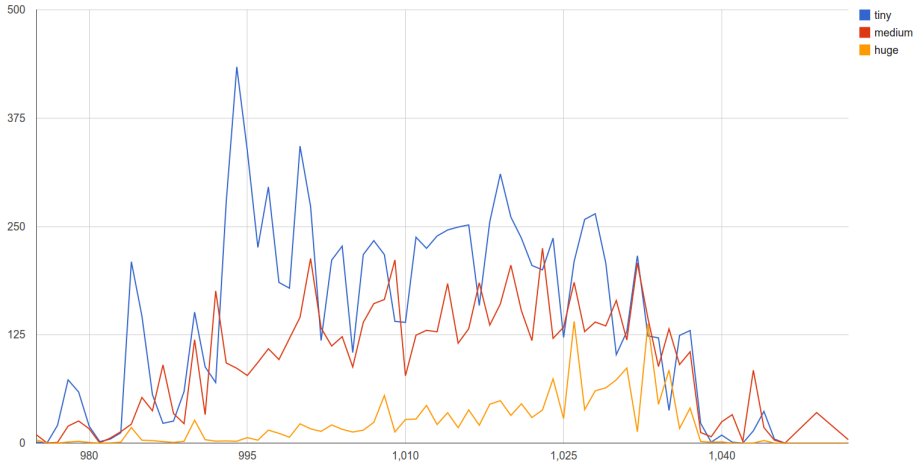
RDF Data Cube -sanaston käyttö aineistojen kuvailuun takaa sen, että aineistot ovat rakenteeltaan samankaltaisia ja rakenne on oikeanlainen visualisointeja ja aineiston analysointia varten. Lisäksi aineistojen visualisointi onnistuu käyttäen olemassaolevia tätä sanastoa varten kehitettyjä visualisointeja [MHS+13, SAB+12]. Tulevaisuudessa saatavilla tulee olemaan nykyistä parempia ja monipuolisempia avoimia visualisointityökaluja.

8.3 Aineistojen yhteyksien hahmottaminen visualisoimalla

Onko lintumuuttoaineiston päivittäisten havaintomäärien ja säähavaintoaineiston väliltä mahdollista hahmottaa yhteyksiä visualisoimalla dataa?

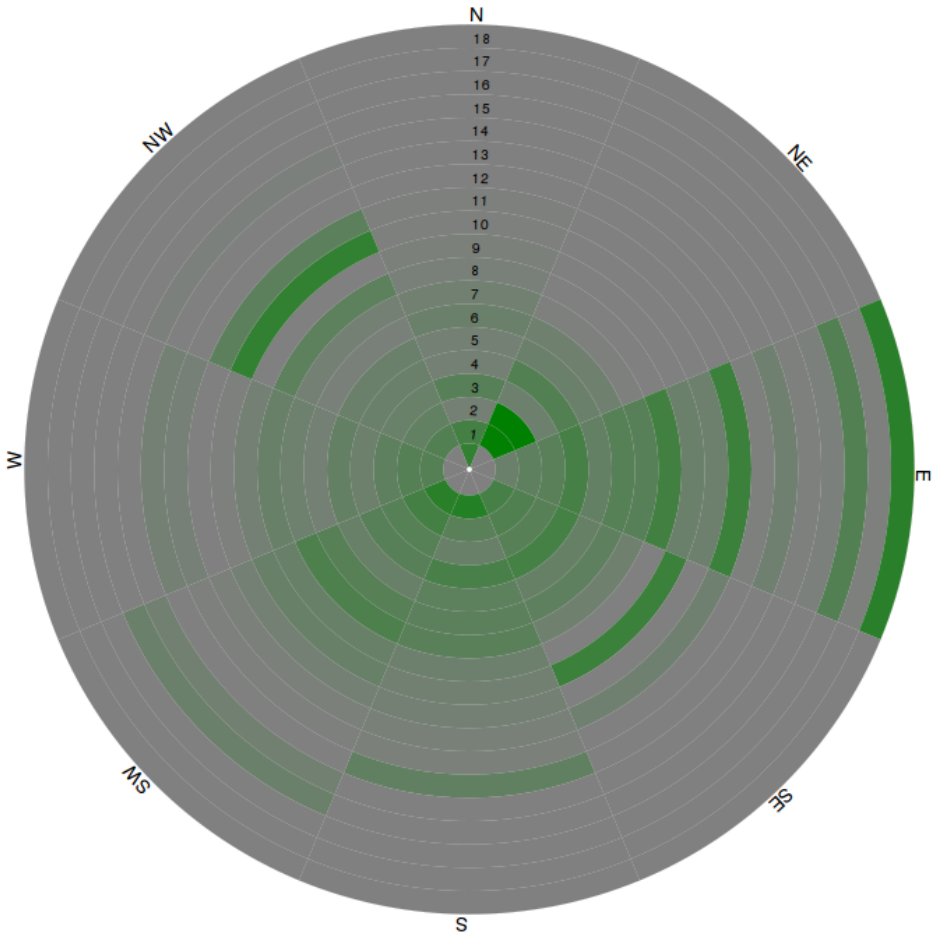
Visualisointeihin on helppo yhdistellä useita toisiinsa linkittyviä aineistoja. Voidaan käyttää lintuhavaintoaineistoa, säädataa sekä tuntomerkkiontologiaa ja visualisoida esimerkiksi sitä, millaisissa ilmanpaineoloissa eri kokoluokkien linnut muuttavat. Kuva 10 näyttää viivakuvaajan kolmen tuntomerkkiontologian kokoluokituksen omaavien lintujen muuton voimakkuuk-

sista erilaisissa ilmanpaineoloissa. Kuvaajan sininen viiva on hyvin pienet, punainen on rastaan kokoiset ja keltainen on suurikokoiset linnut. Kuvaaja näyttäisi noudattavan oletusta, että isokokoiset linnut muuttavat ”hyvällä säällä” eli korkeampien ilmanpaineiden vallitessa, koska ne ovat enemmän riippuvaisia hyvistä lento-olosuhteista. Kuvaajan SPARQL-haku jättää pois alle 975 barin ilmanpaineet, koska tätä ilmanpainetta on esiintynyt vain kahtena päivänä ja rastaan kokoisten lintujen kohdalla olisi tässä kohdin valtava piikki. Tätä pienemmiltä ilmanpaineilta ei ollut juurikaan havaintoja. SPARQL-kysely kuvaajan datan hakemiseksi on liitteessä 2.



Kuva 10: Kolmen eri kokoluokituksen lintujen muuton voimakkuus erilaisissa ilmanpaineoloissa. Pystyakselilla on havaittu normalisoitu lintumäärä ja vaaka-akselilla on ilmanpaine pascaleina.

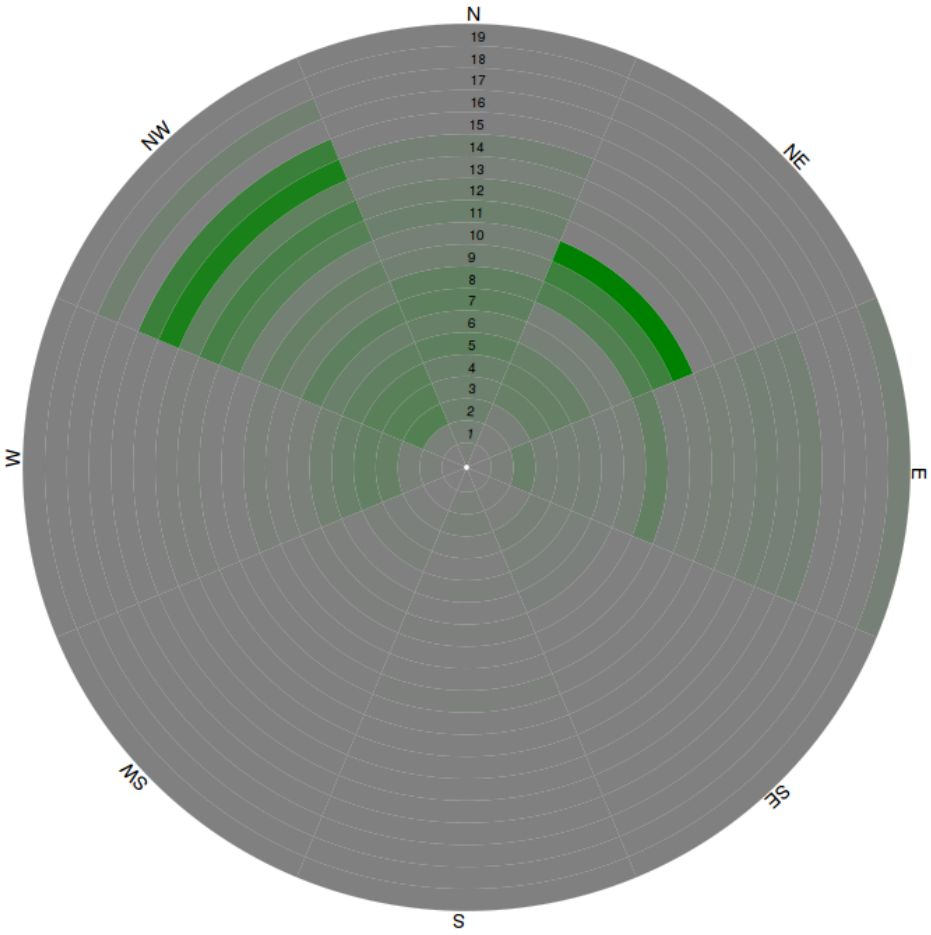
Kuva 11 esittää eri tuuliolosuhteiden esiintymistä keväisen kurkimuuton aikana koko aineiston perusteella. Lintumäärinä on käytetty päiväsummia ja tuulina vastaavasti havaintopäivien koko valoisan ajan tuulihavaintoja. Linnut pyrkivät suorittamaan muuttolentoaan mahdollisimman paljon myötätuulessa, jolloin matkanteko kuluttaa vähiten energiaa [TML14a]. Kurjen päämuuttoreitit Suomessa kulkevat keväisin laajalla rintamalla Suomenlahden ja Uudenmaan yli kohti pohjoisen pesimäalueita [TML14a, TML14b]. Kuvasta on nähtävissä, että muutto on voimakkaampaa eteläisillä tuulilla kuin pohjoisilla. Kuvan perusteella kurjet muuttavat voimakkaasti myös



Kuva 11: Keväisen kurkimuuton jakaantuminen erilaisiin tuuliolosuhteisiin.

itätuulilla, mikä voisi selittyä sillä, että itätuuli painaa Hangon itäpuolelta Suomenlahden yli muuttavia lintuja lähemmäs Hankoa, jolloin niitä havaitaan lintuasemalta. 10–11 m/s luoteistuulella näkyy kuvaajassa myös jostain syystä voimakasta muuttoa. Datan hakemiseen käytetty SPARQL-kysely keväisen kurkimuuton visualisoinnissa on liitteessä 3.

Kuva 12 näyttää vastaavasti tuuliolosuhteet syksyiselle kurkimuutolle. Syksyisin kurkimuutto keskittyy voimakkaasti läntiselle Uudellemaalle [TML14a] kurkien muuttaessa pesimäalueilta etelään talvehtimisalueille. Kuvassa on nähtävissä hyvin voimakas pohjoistuulien suosinta, eivätkä kurjet näyttäisi muuttavan käytännössä lainkaan eteläisillä tuulilla.



Kuva 12: Syksyisen kurkimuuton jakaantuminen erilaisiin tuuliolosuhteisiin.

8.4 Aineiston laadun parantaminen linkitettynä datana julkaistaessa

Voidaanko lintuhavaintoaineiston laatua parantaa julkaistaessa sitä linkitettynä datana?

Rakenteelliset virheet alkuperäisessä datassa, kuten väärän tyyppinen arvo tai virheet kardinaalisuudessa, voivat aiheuttaa ongelmia muunnettaessa dataa RDF-muotoon ja etenkin muodoltaan tiukasti määriteltyyn RDF Data Cube -muotoon. Täten aineiston muuntaminen linkitetyksi dataksi paljastaa virheitä alkuperäisessä aineistossa ja auttaa parantamaan aineiston laatua. Lisäksi muunnosvaiheessa dataa käsitellessä on helppo erikseen toteuttaa

datan oikeellisuustarkastuksia.

9 Yhteenveto

Biologisten havaintoaineistojen julkaiseminen linkitettynä datana mahdollistaa useiden aineistojen yhdistämisen toisiinsa. Aineistojen yhdistämisellä voidaan saavuttaa parempi ymmärrys aineistojen sisällöstä ja mahdollistaa näin tarkempien päätelmien tekeminen. Linkitettynä datana julkaistuja aineistoja voidaan helposti rikastaa yhä uusilla aineistoilla.

Tässä tutkimuksessa mallinnettiin Hangon lintuaseman havaintoaineisto ja Ilmatieteenlaitoksen Russarön säähavaintoaineisto linkitettynä datana sekä demonstroitiin aineistojen yhdistämisen tuomia hyötyjä. Havaintoaineistojen mallintaminen käyttäen RDF Data Cube -sanastoa parantaa aineistojen yhteentoimivuutta. Ontologioilla voidaan esittää biologisissa havaintoaineistoissa keskeisiä biologisia taksonomioita, mikä mahdollistaa aineistojen paremman yhteentoimivuuden.

Aineistojen mallintaminen ja muuntaminen RDF-muotoon voi olla melko työlästä, mutta parantaa mahdollisuuksia aineistojen jatkokäyttöön. Muunnosvaihe toimii itsessään eräänlaisena aineiston validointina, tuoden esiin mahdollisia rakenteellisia virheitä. Tämä on huomattavissa etenkin käyttämällä rakenteen tiukasti määrittelevää RDF Data Cube -sanastoa, joka mahdollistaa aineiston rakenteen validoinnin SPARQL-kyselyillä.

Yhdistettyjen aineistojen visualisoinnilla voidaan valottaa aineistojen välisiä suhteita. WWW-selaimessa toimiva visualisointi onnistuu helposti käyttäen avoimesti saatavilla olevia JavaScript-komponentteja SPARQL-rajapinnasta haettavan datan esittämiseen. Suurilla aineistoilla visualisoinnin nopeuden pullonkaulaksi muodostuu helposti RDF-tietovarasto. Tämä näkyy pitkänä viiveenä visualisoitavan datan hakemisessa SPARQL-rajapinnasta.

Tutkielmassa esitetyt menetelmät ovat yleistettävissä lintu- ja säähavaintoaineistojen lisäksi muihin rakenteeltaan samankaltaisiin havaintoaineistoihin.

Linkitettyjen aineistojen perusteella voisi olla mahdollista tehdä jatkotutkimusta biologisesti kiinnostavista tutkimuskysymyksistä. Esimerkiksi tutkimuksessa osallisena ollut biologi on kiinnostunut siitä miten tulevien päivien

sääennusteen avulla pystyttäisiin ennakoimaan näiden päivien lajikohtaisia muuttajamääriä.

Lintumuuton ja säämuuttujien havainnoista pitkältä ajalta olisi mahdollista tehdä ohjelmallista päättelyä. Jatkossa olisi tarpeellista tutkia tapoja hyödyntää sääaineistoa suoraan lintumuuton analysoinnissa ja ennustamisessa.

Yksi mahdollisuus selvittää tarkemmin yhdistettyjen aineistojen suhteita olisi soveltaa koneoppimisen menetelmiä julkaistuun dataan. Tässä tutkimuksessa käytetyillä aineistoilla muuttajamäärien ennustaminen tulevien päivien sääennusteen perusteella voitaisiin nähdä ohjatun oppimisen (supervised learning) ongelmana, jossa päivittäisiä säämuuttujia voidaan ajatella syötteenä, joka vaikuttaa lintujen muuttajamääriä ilmaisevan satunnaismuuttujan arvoihin.

Säätiedoista vielä esimerkiksi lumipeitteen paksuus ja jäätilanne voisivat olla myös hyvin muuttoa selittäviä tekijöitä. Samoin muiden säähavaintoasemien yhdistäminen mahdollistaisi tarkemman kokonaiskuvan luomisen, koska lintumuutto voi pysähtyä jossain toisaalla huonoon säähän vaikka Hangossa sää olisi muuttoa suosiva. Aineistoa olisi myös mahdollista rikastaa muun tyyppisillä ympäristömuuttujilla, joita käytetään ekologiassa selittämään eliölajien esiintymistä eri alueilla [JTv95] tai esimerkiksi muiden lintuasemien aineistoilla.

Hangon Lintuaseman aineisto toivottavasti avataan julkiseksi tulevaisuudessa, jolloin myös RDF-muotoinen aineisto sekä visualisointipalvelu saadaan julkisesti saataville.

Kiitokset

Tämä tutkielma on tehty CSC - Tieteen tietotekniikan keskuksen sekä Aalto-yliopiston perustieteiden korkeakoulun ja Helsingin yliopiston Semanttisen laskennan tutkimusryhmän (SeCo) yhteistyönä osana Tutkimuksen tietoaineistot -hanketta (TTA).

Haluan kiittää professori Eero Hyvöstä työn ohjaamisesta. Kiitokset Jouni Tuomiselle ja Miika Aloselle tuesta visualisointipalvelun toteuttamisessa. Lisäksi kiitän Aleksi Lehikoista avusta lintuhavaintoaineiston käytössä sekä

kaikkia Hangon lintuaseman havainnoijia, joita ilman tämä työ ei olisi ollut mahdollista.

Lähteet

[Abl73] Able, K. P.: *The Role of Weather Variables and Flight Direction in Determining the Magnitude of Nocturnal Bird Migration.* Ecology, 54(5):1031–1041, 1973.

[ADML12] Auer, S., Demter, J., Martin, M. ja Lehmann, J.: *LODStats–an extensible framework for high-performance dataset analytics.* Teoksessa *Knowledge Engineering and Knowledge Management*, sivut 353–362. Springer, 2012.

[AH11] Allemang, D. ja Hendler, J.: *Semantic web for the working ontologist: effective modeling in RDFS and OWL.* Elsevier, 2011.

[Ale11] Alerstam, T.: *Optimal bird migration revisited.* Journal of Ornithology, 152(1):5–23, 2011.

[Apa14] *Apache Jena - Fuseki: serving RDF data over HTTP*, 2014. `http://jena.apache.org/documentation/serving_data/` [02.09.2014].

[B+06] Berners-Lee, T. *et al.*: *Tabulator: Exploring and analyzing linked data on the semantic web.* Teoksessa *Proceedings of the 3rd International Semantic Web User Interaction Workshop, ISWC 2006*, Georgia, Yhdysvallat, marraskuu 2006.

[Ber06] Berners-Lee, T.: *Linked data - Design Issues.* 2006. `http://www.w3.org/DesignIssues/LinkedData.html` [23.07.2014].

[BGM14] Brickley, D., Guha, R. V. ja McBride, B.: *RDF Schema 1.1.* W3C Recommendation, 2014. `http://www.w3.org/TR/2014/REC-rdf-schema-20140225/` [27.02.2014].

[BHB09] Bizer, C., Heath, T. ja Berners-Lee, T.: *Linked data - the story so far.* International journal on semantic web and information systems, 5(3):1–22, 2009.

[BHL01] Berners-Lee, T., Hendler, J. ja Lassila, O.: *The semantic web.* Scientific american, 284(5):28–37, 2001.

[C+11] Cyganiak, R. *et al.*: *Vocabulary of Interlinked Datasets (VoID)*, 2011. `http://vocab.deri.ie/void` [31.07.2014].

[Coo88] Cooke, W. W.: *Report on bird migration in the Mississippi Valley in the years 1884 and 1885*, nide 2. US Department of Agriculture, Division of Economic Ornithology and Mammalogy, 1888.

[Cox13a] Cox, S.: *Geographic information — Observations and measurements*. OGC Abstract Specification, Open Geospatial Consortium, 2013. `http://portal.opengeospatial.org/files/?artifact_id=41579` [01.07.2014].

[Cox13b] Cox, S.: *OWL representation of ISO 19156 (Observation model)*, 2013. `http://def.seegrid.csiro.au/static/isotc211/iso19156/2011/observation.html` [01.07.2014].

[CR14] Cyganiak, R. ja Reynolds, D.: *The RDF Data Cube Vocabulary*. W3C Recommendation, 2014. `http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/` [28.02.2014].

[CWL+14] Cyganiak, R., Wood, D., Lanthaler, M., Klyne, G., Carroll, J. J. ja McBride, B.: *RDF 1.1 Concepts and Abstract Syntax*. W3C Recommendation, 2014. `http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/` [27.02.2014].

[Dat11] *Data Cube Vocabulary - Government Linked Data (GLD) Working Group Wiki*, 2011. `http://www.w3.org/2011/gld/wiki/Data_Cube_Vocabulary` [28.08.2014].

[Dod05] Dodds, L.: *Introducing SPARQL: Querying the Semantic Web*. 2005. `http://www.xml.com/lpt/a/1628` [11.07.2014].

[DR11] Dadzie, A. ja Rowe, M.: *Approaches to visualising linked data: A survey*. Semantic Web, 2(2):89–124, 2011.

[DS05] Dürst, M. ja Suignard, M.: *Internationalized resource identifiers (IRIs)*. tekninen raportti, The Internet Society, 2005.

[Ehr07] Ehrig, M.: *Ontology Alignment: Bridging the Semantic Gap*. Springer Science+Business Media, LLC, 2007.

[EMLA13] Ermilov, I., Martin, M., Lehmann, J. ja Auer, S.: *Linked open data statistics: Collection and exploitation*. Teoksessa *Knowledge Engineering and the Semantic Web*, sivut 242–249. Springer, 2013.

[FCH04] Franke, A., Caelli, T. ja Hudson, R.: *Analysis of movements and behavior of caribou (Rangifer tarandus) using hidden Markov models*. Ecological Modelling, 173:259–270, 2004.

[Fra11] Franz, N. M.: *Biological taxonomy and ontology development: scope and limitations*. Biodiversity informatics, 7(1), 2011.

[FZ13] Fernández, A. ja Zarrabeitia, A. S.: *Implementation of a Linked Open Data Solution for the Statistics Agency of Cantabria's Metadata and Data Bank*. DCMI International Conference on Dublin Core and Metadata Applications, 2013. `http://dcpapers.dublincore.org/pubs/article/view/3681` [19.10.2014].

[G+93] Gruber, T. R. *et al.*: *A translation approach to portable ontology specifications*. Knowledge acquisition, 5(2):199–220, 1993.

[GFH+04] Graham, C. H., Ferrier, S., Huettman, F., Moritz, C. ja Peterson, A. T.: *New developments in museum-based informatics and applications in biodiversity analysis*. Trends in Ecology & Evolution, 19(9):497 – 503, 2004.

[GOS09] Guarino, N., Oberle, D. ja Staab, S.: *What Is an Ontology?* Teoksessa *Handbook on Ontologies*, International Handbooks on Information Systems, sivut 1–17. Springer Berlin Heidelberg, 2009.

[Gru95] Gruber, T. R.: *Toward principles for the design of ontologies used for knowledge sharing?* International Journal of Human-Computer Studies, 43(5—6):907–928, 1995.

[H+13] Hawke, S. *et al.*: *W3C Semantic Web Activity Homepage*, 2013. `http://www.w3.org/2001/sw/` [05.02.2014].

[HAKT13] Hyvönen, E., Alonen, M., Koho, M. ja Tuominen, J.: *BirdWatch—Supporting Citizen Scientists for Better Linked Data Quality for Biodiversity Management.* Teoksessa *Proceedings of the first international Workshop on Semantics for Biodiversity (S4BioDiv), ESWC 2013.* CEUR Workshop Proceedings, Montpellier, France, toukokuu 2013.

[Han13a] *Hangon lintuaseman julkaisuluettelo 1980-*, 2013. `http://www.tringa.fi/hangon-lintuasema/julkaisut/` [15.05.2014].

[Han13b] *Hangon lintuaseman yleisohjeet*, 2013. `http://www.tringa.fi/wp-content/uploads/2014/04/Yleisohjeet-20130716.pdf` [31.07.2014].

[HAV14] Hyland, B., Atemezing, G. ja Villazón-Terrazas, B.: *Best Practices for Publishing Linked Data.* W3C Working Group Note, 2014. `http://www.w3.org/TR/2014/NOTE-ld-bp-20140109/` [25.07.2014].

[HB11] Heath, T. ja Bizer, C.: *Linked Data: Evolving the Web into a Global Data Space (1st edition)*, nide 1 sarjassa *Synthesis Lectures on the Semantic Web: Theory and Technology.* Morgan & Claypool, 2011. `http://linkeddatabook.com/editions/1.0/` [19.10.2014].

[HCF+07] Hochachka, W., Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D. ja Kelling, S.: *Data-Mining Discovery of Pattern and Process in Ecological Systems.* Wildlife management, 71(7):2427–2437, syyskuu 2007.

[HHR+09] Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L. ja Ayers, D.: *SCOVO: Using Statistics on the Web of Data.* Teoksessa *The Semantic Web: Research and Applications*, nide 5554 sarjassa *Lecture Notes in Computer Science*, sivut 708–722. Springer–Verlag, 2009.

[Hog14] Hogan, A.: *Linked Data & the Semantic Web Standards.* Teoksessa *Linked Data Management: Principles and Techniques*, Series on Emerging Directions in Database Systems and Applications. CRC Press, 2014.

[HSP14] Harris, S., Seaborne, A. ja Prud'hommeaux, E.: *SPARQL 1.1 Query Language.* W3C Recommendation, W3C, 2014. `http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/` [11.07.2014].

[HTAM14] Hyvönen, E., Tuominen, J., Alonen, M. ja Mäkelä, E.: *Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets*, Heraklion, Kreikka, toukokuu 2014.

[JTv95] Jongman, R. H. G., Ter Braak, C. J. F. ja van Tongeren, O. F. R.: *Data analysis in community and landscape ecology.* Cambridge University Press, 1995.

[JW14] Jacobs, I. ja Walsh, N.: *Architecture of the World Wide Web, Volume One.* W3C Recommendation, 2014. `http://www.w3.org/TR/2004/REC-webarch-20041215/` [25.07.2014].

[KH11] Kämpgen, B. ja Harth, A.: *Transforming statistical linked data for use in OLAP systems.* Teoksessa *I-Semantics '11: Proceedings of the 7th international conference on Semantic systems*, sivut 33–40, New York, Yhdysvallat, 2011. ACM.

[KHKP06] Kennedy, J., Hyam, R., Kukla, R. ja Paterson, T.: *Standard data model representation for taxonomic information.* OMICS: A Journal of Integrative Biology, 10(2):220–230, 2006.

[KHL+07] Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C. ja Giannopoulou, E.: *Ontology visualization methods—a survey.* ACM Computing Surveys (CSUR), 39(4):10, 2007.

[KHL14] Koho, M., Hyvönen, E. ja Lehikoinen, A.: *Ornithology Based on Linking Bird Observations with Weather Data.* Teoksessa *Proceedings of the 4th Workshop on Semantic Publishing (SePublica),*

ESWC 2014. CEUR Workshop Proceedings, Heraklion, Kreikka, toukokuu 2014.

[Kje98] Kjellén, N.: *Annual variation in numbers, age and sex ratios among migrating raptors at Falsterbo, Sweden from 1986–1995*. Journal für Ornithologie, 139(2):157–171, 1998.

[L+08] Lehikoinen, A. *et al.*: *Lintukantojen kehitys Hangon lintuaseman aineiston mukaan 1979–2007*. Tringa, 35:313–321, 2008.

[LBH+12] Lefort, L., Bobruk, J., Haller, A., Taylor, K. ja Woolf, A.: *A Linked Sensor Data Cube for a 100 Year Homogenised Daily Temperature Dataset*. Teoksessa *Proceedings of the 5th International Workshop on Semantic Sensor Networks (SSN2012), ISWC 2012*, sivut 1–16. CEUR Workshop Proceedings, Boston, Yhdysvallat, marraskuu 2012.

[LR13] Labra Gayo, J. E. ja Rodrıguez, J. M. A.: *Validating statistical index data represented in RDF using SPARQL queries*. W3C RDF Validation Workshop, Practical Assurances for Quality RDF Data, Cambridge, Yhdysvallat, syyskuu 2013. `http://www.w3.org/2001/sw/wiki/images/d/d4/ValidatingStatisticalIndexData.pdf` [19.10.2014].

[LSB+10] Lehikoinen, A., Saurola, P., Byholm, P., Lindén, A. ja Valkama, J.: *Life history events of the Eurasian sparrowhawk Accipiter nisus in a changing climate*. Journal of avian biology, 41(6):627–636, 2010.

[LV00] Lehikoinen, A ja Vähätalo, A: *Phenology of bird migration at the Hanko Bird Observatory, Finland, in 1979–1999*. Tringa, 27:150–224, 2000.

[LVG14] Lepage, D., Vaidya, G. ja Guralnick, R.: *Avibase – a database system for managing and organizing taxonomic concepts*. ZooKeys, 420:117–135, 2014. `http://zookeys.pensoft.net/articles.php?id=3906` [19.10.2014].

[MHS+13] Mutlu, B., Hoefler, P., Sabol, V., Tschinkel, G. ja Granitzer, M.: *Automated Visualization Support for Linked Research Data.* Teoksessa *Proceedings of the I-SEMANTICS 2013 Posters & Demonstrations Track*, sivut 40–44. CEUR Workshop Proceedings, Graz, Itävalta, syyskuu 2013.

[MR05] Maimon, O. ja Rokach, L. (toimittajat): *The data mining and knowledge discovery handbook.* Springer, 2005.

[Muo04] Muona, J.: *Systematiikka, taksonomia, fylogenia ja luokittelu - sisältöä sanoihin*, 2004. `http://koivu.luomus.fi/elaintiede/hyonteiset/tietoa/systematiikka.htm` [18.10.2014].

[Mv04] McGuinness, D. L. ja van Harmelen, F.: *OWL Web Ontology Language Overview.* W3C Recommendation, 2004. `http://www.w3.org/TR/2004/REC-owl-features-20040210/` [25.07.2014].

[ND68] Nisbet, I. C. T. ja Drury Jr., W. H.: *Short-term effects of weather on bird migration: A field study using multivariate statistics.* Animal Behaviour, 16(4):496–530, 1968.

[Ohj14] *Ohjeita miehittäjille*, 2014. `http://www.tringa.fi/web/lintuasemat/hangon-lintuasema/ohjeita-miehitt%C3%A4jille.html` [03.04.2014].

[Pan09] Pan, J. Z.: *Resource Description Framework.* Teoksessa Staab, S. ja Studer, R. (toimittajat): *Handbook on Ontologies*, International Handbooks on Information Systems, sivut 71–90. Springer Berlin Heidelberg, 2009. `http://dx.doi.org/10.1007/978-3-540-92673-3_3`.

[PMP13] Petrou, I., Meimaris, M. ja Papastefanatos, G.: *Towards a methodology for publishing Linked Open Statistical Data.* The Share-PSI 2.0 Workshop, 2013. `http://www.w3.org/2013/share-psi/wiki/images/e/e2/LinkedStatistics_SharePSI2.0.pdf` [25.07.2014].

[PSV12] Peroni, S., Shotton, D. ja Vitali, F.: *The Live OWL Documentation Environment: a tool for the automatic generation of ontology documentation.* Teoksessa *Knowledge Engineering and Knowledge Management*, sivut 398–412. Springer, 2012.

[RJS11] Reichman, O. J., Jones, M. B. ja Schildhauer, M. P.: *Challenges and opportunities of open data in ecology.* Science, 331(6018):703–705, 2011.

[Rod06] Roderic, P.: *Taxonomic names, metadata, and the Semantic Web.* Biodiversity Informatics, 3, 2006. `https://journals.ku.edu/index.php/jbi/article/view/25` [19.10.2014].

[SAB+12] Salas, P. E. R., Auer, S., Breitman, K. K., Casanova, M. A. ja Martin, M.: *Publishing Statistical Data on the Web.* International Journal of Semantic Computing, 6(4):373–388, 2012.

[SEK08] Sheldon, D., Elmohamed, M. A. Saleh ja Kozen, D.: *Collective Inference on Markov Models for Modeling Bird Migration.* Advances in Neural Information Processing Systems, 20:1321–1328, 2008.

[Ska12] Skaeveland, M.: *Sgvizler: A JavaScript Wrapper for Easy Visualization of SPARQL Result Sets.* Teoksessa *Proceedings of the ESWC 2012*. Springer–Verlag, Heraklion, Kreikka, toukokuu 2012.

[SSB08] Schulz, S., Stenzhorn, H. ja Boeker, M.: *The ontology of biological taxa.* Bioinformatics, 24(13):i313–i321, 2008. `http://bioinformatics.oxfordjournals.org/content/24/13/i313.full` [19.10.2014].

[Sta09] Statistical Data and Metadata eXchange: *SDMX User Guide - Version 2009.1*, 2009. `http://sdmx.org/wp-content/uploads/2009/02/sdmx-userguide-version2009-1-71.pdf` [30.07.2014].

[Sve78] Svensson, S. E.: *Efficiency of two methods for monitoring bird population levels: breeding bird censuses contra counts of migrating birds.* Oikos, 31:373–386, 1978.

[TLH11] Tuominen, J., Laurenne, N. ja Hyvönen, E.: *Biological Names and Taxonomies on the Semantic Web – Managing the Change in Scientific Conception.* Teoksessa *Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011)*. Springer–Verlag, Heraklion, Kreikka, kesäkuu 2011.

[TLKH13] Tuominen, J., Laurenne, N., Koho, M. ja Hyvönen, E.: *The Birds of the World Ontology AVIO.* Teoksessa Cimiano, P. *et al.* (toimittajat): *The Semantic Web: ESWC 2013 Satellite Events*, nide 7955 sarjassa *Lecture Notes in Computer Science*, sivut 300–301. Springer–Verlag, 2013.

[TML14a] Toivanen, T., Metsänen, T. ja Lehtiniemi, T.: *Lintujen päämuuttoreitit Suomessa*, 2014. `http://www.ymparisto.fi/download/noname/%7BFA98FD1F-987F-4546-84F7-93BDC1F0CE06%7D/100332` [11.07.2014].

[TML14b] Toivanen, T., Metsänen, T. ja Lehtiniemi, T.: *Lintujen päämuuttoreitit Suomessa: karttaliite*, 2014. `http://www.ymparisto.fi/download/noname/%7B31868315-3213-4C2E-ADB6-75A7BBF693F2%7D/100333` [11.07.2014].

[Uus06] Uusivuori, P.: *Suositus kenttähavaintojen merkitsemiseksi.* 2006. `http://www.birdlife.fi/lintuharrastus/suositus_kenttahavaintojen_merkitsemiseksi.pdf` [03.02.2014].

[VAR09] Villa, F., Athanasiadis, I. N. ja Rizzoli, A. E.: *Modelling with knowledge: A review of emerging semantic approaches to environmental modelling.* Environmental Modelling & Software, 24(5):577–587, 2009.

[W+13] Wieczorek, J. *et al.*: *Darwin Core*, 2013. `http://rs.tdwg.org/dwc/` [08.08.2014].

[WBG+12] Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, Ma., Giovanni, R., Robertson, T. ja Vieglais, D.: *Darwin Core: An evolving community-developed biodiversity data standard.* PLoS One, 7(1):e29715, 2012.

[YWH10] Yu, J., Wong, W. ja Hutchinson, R. A.: *Modeling experts and novices in citizen science data for species distribution modeling.* Teoksessa *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, sivut 1157–1162. IEEE, Sydney, Australia, joulukuu 2010.

Liite 1 SPARQL-kysely tuulivisualisointiin

Esimerkki SPARQL-kyselystä, jolla haetaan tuulivisualisointiin dataa. Haku hakee kaikkien kesien merimetsohavainnot ja näinä päivinä esiintyneet tuulet sekä normalisoi lintumäärät tuulten esiintymismäärän mukaan.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX h: <http://ldf.fi/schema/halias/>
PREFIX qb: <http://purl.org/linked-data/cube#>

SELECT ?order ?speed ((SUM(?cnt)/?days) AS ?normalized)
WHERE {
    ?observation h:observedSpecies ?species .
    ?species h:abbreviation ?label .
    FILTER(lcase(str(?label)) = 'phacar') .
    ?observation h:season h:summer .
    ?observation h:countTotal ?cnt .
    ?observation h:refTime ?date .

    ?wind h:windSpeed ?speed .
    ?wind h:windDirection ?dir .
    ?dir h:order ?order .
    ?weatherobs h:refTime ?date .
    ?weatherobs h:windDay ?wind .
    {
        SELECT ?wind (COUNT(DISTINCT ?date) AS ?days)
        WHERE {
            ?weatherobs h:refTime ?date .
            ?weatherobs h:windDay ?wind .
            ?weatherobs h:season h:summer .
            ?weatherobs h:haliasObservationDay ?observed .
            filter (?observed)
        }
        GROUP BY ?wind
    }
```

```
}
GROUP BY ?order ?speed ?days
ORDER BY ?speed
```

Liite 2 SPARQL-kysely ilmanpainevisualisointiin

Liitteen SPARQL-kysely palauttaa eri kokoluokan lintujen lukumääriä suhteutettuna ilmanpaineeseen. Kyselyn tulokset voidaan suoraan visualisoida esimerkiksi kolmena viivakuvaajana.

```
PREFIX h: <http://ldf.fi/schema/halias/>
PREFIX bc: <http://ldf.fi/halias/bird-characteristics/>

SELECT ?pressure ?tiny ?medium ?huge
WHERE {
{
SELECT ?pressure ((SUM(?cnt)/?days) AS ?tiny)
WHERE {
    ?observation h:observedSpecies ?species .
    ?species h:hasCharacteristic bc:hyvinpieni .
    ?observation h:refTime ?date .
    ?observation h:countMigration ?cnt .
    ?weather h:refTime ?date .
    ?weather h:airPressure ?pressure .
{
SELECT ?pressure (COUNT(DISTINCT ?date) AS ?days)
WHERE {
?weather h:refTime ?date .
?weather h:airPressure ?pressure .
}
GROUP BY ?pressure
}
}
GROUP BY ?chara ?pressure ?days
ORDER BY ?pressure
}

{
SELECT ?pressure ((SUM(?cnt)/?days) AS ?medium)
```

```
WHERE {
    ?observation h:observedSpecies ?species .
    ?species h:hasCharacteristic bc:rastas .
    ?observation h:refTime ?date .
    ?observation h:countMigration ?cnt .
    ?weather h:refTime ?date .
    ?weather h:airPressure ?pressure .
{
SELECT ?pressure (COUNT(DISTINCT ?date) AS ?days)
WHERE {
?weather h:refTime ?date .
?weather h:airPressure ?pressure .
}
GROUP BY ?pressure
}
}
GROUP BY ?chara ?pressure ?days
ORDER BY ?pressure
}

{
SELECT ?pressure ((SUM(?cnt)/?days) AS ?huge)
WHERE {
    ?observation h:observedSpecies ?species .
    ?species h:hasCharacteristic bc:valtava .
    ?observation h:refTime ?date .
    ?observation h:countMigration ?cnt .
    ?weather h:refTime ?date .
    ?weather h:airPressure ?pressure .
{
SELECT ?pressure (COUNT(DISTINCT ?date) AS ?days)
WHERE {
?weather h:refTime ?date .
?weather h:airPressure ?pressure .
```

```
}
GROUP BY ?pressure
}
}
GROUP BY ?chara ?pressure ?days
ORDER BY ?pressure
}
FILTER (?pressure > 974)
}
```

Liite 3 SPARQL-kysely kurkimuuton tuulivisualisointiin

SPARQL-kysely tuulivisualisointia varten, joka hakee kurkien muuttajamääriä keväiltä ja normalisoi lintumäärät tuulten esiintymismäärän mukaan.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX h: <http://ldf.fi/schema/halias/>
PREFIX qb: <http://purl.org/linked-data/cube#>

SELECT ?order ?speed ((SUM(?cnt)/?days) AS ?normalized)
WHERE {
    ?observation h:observedSpecies ?species .
    ?species rdfs:label ?label .
    FILTER(lcase(str(?label)) = 'grus grus') .
    ?observation h:season h:spring .
    ?observation h:countMigration ?cnt .
    ?observation h:refTime ?date .

    ?wind h:windSpeed ?speed .
    ?wind h:windDirection ?dir .
    ?dir h:order ?order .
    ?weatherobs h:refTime ?date .
    ?weatherobs h:windDay ?wind .
    {
        SELECT ?wind (COUNT(DISTINCT ?date) AS ?days)
        WHERE {
            ?weatherobs h:refTime ?date .
            ?weatherobs h:windDay ?wind .
            ?weatherobs h:season h:spring .
            ?weatherobs h:haliasObservationDay ?observed .
            filter (?observed)
        }
        GROUP BY ?wind
```

```
    }
}
GROUP BY ?order ?speed ?days
ORDER BY ?speed
```