LAJITIETOKESKUKSEN KOKONAISARKKITEHTUURI

Tämä on Lajitietokeskuksen kokonaisarkkitehtuuridokumentin versio 1.0

Versiohistoria

2015-08-19: Ensimmäinen versio 0.0 luotu olemassaolevista materiaaleista pohjaksi jatkokehitykselle. TL
2015-09-09: Muokattu sisältöä Hanna Koivulan kanssa käytyjen keskustelujen pohjalta. TL
2015-09-09: Lisätty Liite 1. TL
2015-09-10: Jatkettu Liitteen 1 tekstin muokkausta. TL
2016-01-18: Siirretty dokumentaatio Word-dokumentista wikiin. TL
2016-01-26: Muokattu tietoarkkitehtuurin dokumentaatiota. Luotu sivusto schema.laji.fi, jossa julkinen dokumentaatio luokista ja attribuuteista. TL ja V-MR
2016-11-20: Muokattu dokumentaation sisältöä useasta eri kohdasta. TL
2016-12-05: Toiminta-arkkitehtuurin ja tietoarkkitehtuurin dokumentaation muokkausta. TL
2016-12-15: Korjattu pikkuvirheitä käsitteistössä ja kieliopissa. TL
2016-12-16: Muokattu dokumenttia monesta eri kohdasta. TL
2016-12-19: Poistettu Liite 4 (Lajitietokeskuksen sidosryhmät), jonka sisältö liitetty osaksi dokumenttia
2017-02-09: Muodostettu wikissä olevasta sisällöstä KA-dokumentin versio 1.0. TL

JOHDANTO

Valtionvarainministeriö myönsi Suomen ympäristökeskukselle (SYKE) vuosille 2015-2017 määrärahan ympäristö- ja luonnonvaratietojen avaamista, yhtenäistämistä ja käyttöä edistävään Envibase-hankkeeseen (http://www.ymparisto.fi/envibase). Osana Envibase-hanketta toteutetaan valtakunnallinen Lajitietokeskus, jonka kehitystyön koordinoinnista vastaa Helsingin yliopiston alainen Luonnontieteellinen keskusmuseo (LUOMUS).

Lajitiedolla tarkoitetaan eliölajien luokitteluun, nimistöön, ominaisuuksiin, esiintymiseen, tutkimukseen, suojeluun ja hyväksikäyttöön liittyvää informaatiota. Lajitietokeskuksen tavoitteena on luoda Suomeen virtuaaliorganisaatio, jonka jäseninä ovat kaikki keskeiset lajitietoa tuottavat ja hyödyntävät tahot. Lajitietokeskuksen avulla lajidatan tuottamiseen, jakeluun ja säilyttämiseen liittyvät kansalliset prosessit pyritään yhdenmukaistamaan ja yksinkertaistamaan käyttäen apuna modernia informaatio- ja kommunikaatioteknologiaa. Myös Suomen kansainväliset velvoitteet eurooppalaisessa ja globaalissa mittakaavassa otetaan huomioon Lajitietokeskuksen toiminnan määrittelyssä.

Tässä dokumentissa kuvataan Lajitietokeskuksen kokonaisarkkitehtuuri käyttäen vakiintuneita kokonaisarkkitehtuurin kuvausmenetelmiä. Kuvaus tehdään neljällä abstraktiotasolla, jotka ovat toiminta-arkkitehtuuri, tietoarkkitehtuuri, tietojärjestelmäarkkitehtuuri ja teknologia-arkkitehtuuri. Kullakin tasolla kuvataan ensin lajitiedon hallinnan nykytila. Tämän jälkeen kuvataan lajitiedon hallinnan tavoitetila, johon pyritään Lajitietokeskuksen toiminnan käynnistämisellä ja vakiinnuttamisella.

Lajitietokeskus ei konseptina ole kansainvälisessä mittakaavassa missään määrin uusi ja ainutlaatuinen. Esimerkiksi Britanniassa vastaavanlainen organisaatio, Biological Records Centre, perustettiin jo vuonna 1964 (http://www.brc.ac.uk). Ruotsissa kansallinen ArtDatabanken aloitti toimintansa vuonna 1990 (http://artdatabanken.se). Yksi 2000-luvun kehittyneimmistä järjestelmistä on Atlas of Living Australia (http://www.ala.org.au/), joka avattiin julkiseen käyttöön vuonna 2010. Myös Hollannissa toimiva Naturalis Biodiversity Center (http://www.naturalis.nl/en/) on hyvä esimerkki modernista, Lajitietokeskuksen periaatteita vastaavasta luonnontieteellisestä museosta, joka yhdistää perinteiseen museotoimintaan nykyaikaisen biodiversiteetti-informatiikan näkökulman.

Suomessa lajitietojen koordinointi kansallisella tasolla alkaa siis verrattain myöhään moniin muihin kehittyneisiin maihin verrattuna. Lajitiedon keruuta ja jakelua koordinoivan organisaation puuttuminen on johtanut suomalaisen lajitiedon pirstoutumiseen lukuisiin erillisiin, toistensa kanssa epäyhteensopiviin tietojärjestelmiin, mikä on lisännyt päällekkäistä kehitys- ja ylläpitotyötä ja manuaalisen työn määrää tietojen yhdistelyssä ja analysoinnissa. Lajitietokeskushankkeen käynnistyminen muita maita myöhemmin antaa toisaalta mahdollisuuden

käyttää hyväksi muiden maiden kokemuksia ja rakentaa kansallinen lajitiedon hallinnan järjestelmä uusimman käytettävissä olevan ICT-arkkitehtuurin varaan.

TOIMINTA-ARKKITEHTUURI

Nykytila

Lajitiedolla tarkoitetaan nykyisin tai menneisyydessä eläneiden (fossiilit) eliölajien ominaisuuksiin ja esiintymiseen liittyvää informaatiota. Lajitiedon hallinnan perustana on eliölajien biologinen luokittelu ja nimeäminen, joka perustuu Carl von Linnén jo 1700-luvulla määrittelemiin periaatteisiin.

Lajitiedon tarve on lähtöisin jo ihmiskunnan esihistoriasta, jolloin menestyksekäs ravinnonhankinta ja pedoilta ja taudeilta suojautuminen edellytti todenmukaista tietoa ihmisten elinympäristössä esiintyvistä eliölajeista ja niiden ominaisuuksista. Tämä lajitiedon käyttötapa on edelleen keskeisessä roolissa suomalaisessa maa- ja metsätalouden harjoittamisessa. Myös lajitiedon käyttö alkuperäisimmässä muodossaan riista- ja kalatalouden perustana on edelleen ajankohtaista, joskin sen merkitys ihmisten toimeentulossa on nykyään vähäisempi kuin erätalouden aikoina.

Näiden perinteisten, eliölajien hyödyntämiseen ihmisten toimeentulon perustana liittyvien lajitiedon käyttötapojen rinnalle on nyky-yhteiskunnassa kehittynyt monia luonnon- ja ympäristönsuojeluun sekä vapaa-ajan harrastuksiin liittyviä lajitiedon käyttötapoja. Lisäksi biologinen perustutkimus luonnollisesti tuottaa ja analysoi lajitietoa osana tutkimustoimintaa. Kansalaisten terveyden ja hyvinvoinnin kannalta hyödylliset ja haitalliset lajit kiinnostavat koko väestöä.

Lajitietoa tuotetaan ja käytetään päätöksentekoon, tutkimuksessa, kaupallisissa yhteyksissä sekä harrastustoiminnassa. Päätöksenteon tarpeet hajaantuvat pääasiassa kahden eri ministeriön, Ympäristöministeriön ja Maa- ja metsätalousministeriön alaisuuteen sekä lisäksi usealle hallinnontasolle ministeriöistä ely-keskusten kautta aina kuntatasolle asti. Kansainväliset sitoumukset asettavat lisäksi lajitiedon hyödyntämiselle omat vaatimuksensa.

Paras taksonominen asiantuntemus Suomessa on Luonnontieteellisen keskusmuseon ja muiden luonnontieteellisten museoiden henkilökunnalla. Museoiden asiantuntemus ei kuitenkaan ole kaikkien eliöryhmien osalta kattava, ja täydentävää taksonomista asiantuntemusta löytyy muista valtion laitoksista (mm. yliopistojen biologian laitoksista, Suomen ympäristökeskuksesta ja Luonnonvarakeskuksesta). Eräiden ryhmien osalta jopa yksityiset harrastajat edustavat parasta taksonomista asiantuntemusta.

Taksonomisen tutkimuksen käyttämien tieteellisten nimien lisäksi lajitiedon käsittelyssä tarvitaan lajien ja lajiryhmien suomen- ja ruotsinkielisiä nimiä. Suomenkielisten lajinimistöjen ylläpitoa on koordinoinut biologian seuran Vanamo, jonka asettamat työryhmät ovat määritelleet suomenkielisiä nimiä eräiden eliöryhmien pääasiassa Suomessa esiintyville edustajille. Suomessa käytettävät lajien ruotsinkieliset nimet poikkeavat jonkin verran Ruotsissa käytetyistä nimistä. Lisäksi eräissä

eliöryhmissä (mm. hyönteiset ja hämähäkkieläimet) yksityishenkilöt ja yksityiset harrastajaryhmät ovat julkaisseet suomenkielisiä lajiluetteloita.

Lajien esiintymistä koskevaa havaintotietoa keräävät Suomessa useat eri organisaatiot, joiden välillä ei ole mitään valtakunnallista koordinaatiota. Julkishallinnon organisaatioissa lajitietoa kerätään tutkimuksen, suojelun, luonnonvarojen hyödyntämisen ja maankäytön suunnittelun tarpeisiin. Harrastajayhteisöt ovat keränneet havaintotietoa omiin havaintoarkistoihinsa jo vuosikymmenten ajan, ja 2000-luvulla tiedonkeruuta on myös automatisoitu verkkopalveluiden avulla. Havaintomäärältään merkittävimpiä harrastajayhteisöjen käyttämiä havaintotietokantoja ovat Birdlife Suomen ylläpitämä lintuharrastajien Tiira-järjestelmä ja hyönteisharrastajien käyttämä Suomen perhostutkijain seuran hyönteistietokanta.

Huomattava osa, eräiden arvioiden mukaan jopa yli 90%, lajihavainnoista on peräisin vapaaehtoisilta harrastajilta, jotka joko täysin vapaaehtoisesti tai enintään nimellisen korvauksen vastineeksi tuottavat havaintoja lajien esiintymisestä maan eri osissa. Harrastajien keräämille havaintoaineistoille on tyypillistä taksonominen, maantieteellinen ja ajallinen epätasaisuus, mistä seuraa ongelmia aineistojen tulkinnalle. Parempilaatuista aineistoa saadaan kerättyä useissa eri seurantahankkeissa, joissa harrastajien tiedonkeruuta ohjataan jonkin koordinoivan organisaation kautta.

Tavoitetila

Lajitietokeskuksen tavoitteena on koordinoida suomalaisten eliötietojen keruuta, analysointia ja raportointia. Lajitietokeskuksen jäseninä on sekä julkishallinnon organisaatioita että yksityisiä kansalaisjärjestöjä, jotka keräävät ja käsittelevät eliöiden esiintymiseen ja ominaisuuksiin liittyvää tietoa.

Lajitietokeskus toteutetaan virtuaaliorganisaationa, joka rakentuu jäsenten olemassaolevan osaamisen ja infrastruktuurin varaan. Panostusta tarvitaan yhteistyön koordinointiin ja teknisten rajapintojen luomiseen ja ylläpitoon. Tätä tavoitetta edistetään kuvaamalla lajitietokeskuksen toiminta kokonaisarkkitehtuurina (tämä dokumentti).

Osoitteessa http://laji.fi on Lajitietokeskuksen suurelle yleisölle tarkoitettu julkinen portaali, jonka kautta voi hakea ja tallettaa tietoja lajeista ja niiden ominaisuuksista. Keskeinen osa Lajitietokeskuksen toimintaa on kuitenkin tarjota myös mahdollisuuksia lajitiedon käyttämiseen koneluettavassa muodossa kolmansien osapuolten kehittämien sovellusten tarpeisiin. Tähän tarkoitukseen käytettävät rajapinnat toteutetaan ja dokumentoidaan osoitteessa http://api.laji.fi.

Lajitietokeskuksen toiminnan lähtökohtana on avoin data. Kaikki lajitietokeskuksen käsittelemä data on oletusarvoisesti avointa ja vapaasti käytettävissä sekä ei-kaupallisiin että kaupallisiin tarkoituksiin. Mikäli tietojen käyttöä rajataan, siihen tulee olla aina joko lainsäädännöstä johdetut tai muutoin hyvät perustelut (esim. Julkisuuslaki tai havainnoijien/kerääjien halu salata omia tietojaan).

Lajitietokeskuksen käsittelemiä aineistoja hallinnoidaan tieteellisin periaattein. Tämä tarkoittaa sitä, että aineistojen laatua arvioidaan kriittisesti parhaan käytettävissä olevan tiedon perusteella, ja käyttäjien aineiston laatuun kohdistamat huomautukset arvioidaan asianmukaisesti. Tavoitteena on tuottaa, ylläpitää ja jakaa mahdollisimman hyvälaatuista aineistoa sekä kotimaisiin että kansainvälisiin käyttötarkoituksiin.

Lajitietokeskuksen tavoitteena on muodostaa eliölajistoa koskeva, ajallisesti kattava kansallinen tietoaineisto, jota voidaan käyttää mm. poliittisen päätöksenteon ja maankäytön suunnittelun tieteellisenä perustana. Yhdistämällä tämä kansallinen aineisto kansainvälisiin vastaavanlaisiin aineistoihin on mahdollista luoda myös ennusteita eliöstön tulevista kehityslinjoista.

Lajitietokeskuksen sidosryhmät

Lajitietokeskuksen kansallisia sidosryhmiä ovat muun muassa seuraavat tahot:

- Luonnontieteellinen keskusmuseo ja muut luonnontieteelliset museot
- Suomen ympäristökeskus (Syke)
- Luonnonvarakeskus (Luke, ent. Metla, MTT, RKTL)
- Metsähallitus
- Biologian sekä maatalous- ja metsäalan oppi- ja tutkimuslaitokset
- Evira
- Riistakeskus
- ELY-keskukset
- Kunnalliset ympäristövirannomaiset
- Vanamo-seura
- Societas pro Fauna et Flora Fennica
- Suomen luonnonsuojeluliitto
- Luonto-Liitto
- Etelä-Karjalan allergia- ja ympäristöinstituutti
- BirdLife Suomi ja sen jäsenjärjestöt
- Suomen hyönteistieteellinen seura
- Helsingin hyönteistieteellinen yhdistys
- Suomen perhostutkijain seura
- Suomen nisäkästieteellinen seura
- Lepakkoseura
- Suomen kasviplanktonseura
- Tampereen kasvitieteellinen yhdistys
- Kalastusalan järjestöt
- Metsästysjärjestöt

- Puutarha-alan järjestöt
- Maanviljelysjärjestöt
- Helcom
- Luontoportti
- Partio
- 4H-kerhot
- Yksityishenkilöt

Lajitietokeskuksen biodiversiteetti-informattiikkaan liittyviä kansainvälisiä sidosryhmiä ja verkostoja ovat:

- GBIF: Global Biodiversity Information Facility (http://www.gbif.org)
- GEO BON: Group on Earth Observations Biodiversity Observation Network (http://geobon.org)
- IPBES: Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (http://www.ipbes.net)
- CBD: Convention on Biological Diversity (http://www.cbd.int)
- IUCN: Internation Union for Conservation of nature (http://www.iucn.org)
- RAMSAR: Convention of Wetlands (http://www.ramsar.org)
- EEA: European Environment Agency (http://www.eea.europa.eu)
- EASIN: European Alien Species Information Network (http://easin.jrc.ec.europa.eu/)

Lajitietokeskuksen avoimen datan ja avoimen tutkimuksen edistämiseen liittyviä sidosryhmiä ovat:

- Avoin tiede ja tutkimus (http://avointiede.fi/)
- RDA: Research Data Alliance (https://www.rd-alliance.org/)
- Open Knowledge International (https://okfn.org/) ja Open Knowledge Finland (http://fi.okfn.org/)
- Linked Data (http://linkeddata.org/) ja Linked Data Finland (http://www.ldf.fi/)

Lajitietokeskuksen palvelut ja tuotteet

Lajitietokeskuksen ydinpalveluja ovat kansallisen lajiluettelon ylläpito ja lajien esiintymistä koskevan havaintoaineiston keruu ja jakelu tietojen käyttäjille. Kansallisessa lajiluettelossa ylläpidetään nimistön lisäksi myös tietoja lajien keskeisistä biologisista ja hallinnollisista ominaisuuksista kuten rauhoituksista ja uhanalaisuudesta. Yleiskäyttöisen web-portaalin lisäksi Lajitietokeskuksen palveluja voi käyttää koneluettavien rajapintojen kautta.

Näiden ydintoimintojen tukena Lajitietokeskus ylläpitää myös henkilö- ja aineistoluetteloita sekä valtakunnallista data-arkistoa biologisten havaintoaineistojen pitkäaikaissäilytystä varten.

Lajitietokeskus tuottaa lajitiedon hallintaan ja käyttöön liittyvää koulutusmateriaalia oppilaitosten, alan järjestöjen ja yksityisten harrastajien tarpeisiin.

Lajitietokeskuksen prosessit ja organisaatio

Lajitietokeskus toimii lajitietoa tuottavien ja käyttävien tahojen välisenä virtuaaliorganisaationa, jolla ei ole omaa henkilöstöä eikä rahoitusta. Lajitietokeskuksen toimintaa koordinoi Luonnontieteellinen keskusmuseo osana omaa toimintaansa. Keskeisenä yhteistyötä koordinoivana toimintamuotona ovat vapaamuotoiset työryhmät, jotka kokoontuvat tarpeen mukaan neuvottelemaan Lajitietokeskuksen toimintaan liittyvistä kysymyksistä.

TIETOARKKITEHTUURI

Nykytila

Lajitietokeskuksen tietoarkkitehtuuri perustuu Luonnontieteellisen keskusmuseon määrittelemään arkkitehtuuriin, jossa aluksi yhtenäistettiin tieteellisten kokoelmien sekä havainto-ja seuranta-aineistojen tietorakenteita siten, että niiden tarkasteleminen samoilla visualisointi- ja analyysivälineillä tuli mahdolliseksi. Alunperin luonnontieteellisten museoiden tiedonhallintaan suunniteltu arkkitehtuuri on Lajitietokeskuksen yhteydessä laajennettu kattamaan kaikki lajitietoa tuottavat ja käyttävät tahot.

Sanastot

Lajitietokeskuksen tiedonhallinta perustuu World Wide Web -konsortion (W3C) määrittelemän **Resource Description Framework** (RDF) -spesifikaation mukaisiin ontologiarakenteisiin. Tiedonhallintaa varten on kuvattu joukko luokkia ja niiden ominaisuuksia, joihin viitataan globaalisti yksikäsitteisillä HTTP URI -tunnisteilla.

Tiedonhallinnan perusyksikkönä RDF-spesifikaatiossa on resurssi, joka on jonkin ontologiarakenteessa kuvatun luokan edustaja eli instanssi. Jokaisella resurssilla on yksikäsitteinen HTTP URI -tunniste ja joukko ominaisuuksia, jotka määräytyvät resurssin luokan perusteella.

Lajitiedon yhteiskäytön kannalta on oleellista, että kaikki lajitietoa hallinnoivat organisaatiot käyttävät samoista asioista samoja käsitteitä. Tietojärjestelmien tasolla yhteiskäyttö varmistetaan käyttämällä yhteisiä URI-tunnisteita tietojen hallinnassa ja välityksessä. Tätä tavoitetta palvelee eliöiden taksonomiaan sekä paikkaan, aikaan ja henkilöihin liittyvä ydintieto (*master data*).

Lajitietokeskuksen käyttämät sanastot on dokumentoitu osoitteessa http://schema.laji.fi.

Luonnontieteellisten museoiden kokoelmatietojen hallinta (Kotka)

Lajitietokeskus ylläpitää osoitteessa https://kotka.luomus.fi Kotka-kokoelmanhallintajärjestelmää, joka on tarkoitettu Suomessa toimivien luonnontieteellisten museoiden kokoelmatietojen hallintaan. Järjestelmä on ollut tuotantokäytössä vuodesta 2011, ja se on käytössä Luonnontieteellisen keskusmuseon lisäksi noin kymmenessä muussa museossa.

Vaikka Kotka-järjestelmän kehitys alkoi ennen Lajitietokeskushankkeen käynnistämistä, se on arkkitehtuuriltaan täysin yhteensopiva Lajitietokeskuksen tietojärjestelmien kanssa, ja sitä kehitetään tekniikaltaan yhteensopivaksi havaintotietojen käsittelyyn tarkoitetun Vihkon (ks. alla) kanssa.

Tavoitetila

Taksonomisen ydintiedon ylläpito

Eliötietojen käsittelyn perustana ovat lajiluettelot, joiden avulla lajien ominaisuuksiin ja esiintymiseen liittyviä tietoja voidaan kytkeä toisiinsa. Eliönimistöjen hallinta sekä kansallisella että kansainvälisellä tasolla on vaativa tehtävä. Lajitietokeskuksen tehtävänä on ylläpitää ja jakaa käyttäjille uusimpaan tieteelliseen tietoon perustuvia valtakunnallisia eliönimistöjä, joiden pohjalla eliöiden ominaisuuksiin ja esiintymiseen liittyviä aineistoja voidaan käsitellä yhdenmukaisella tavalla. Nimien lisäksi kullekin taksonille annetaan globaalisti uniikki HTTP URI -tunniste, jonka avulla lajitietoja voidaan yhdistää kansainvälisiin tietojärjestelmiin.

Lajitietokeskus ottaa vastuun lajien suomenkielisten nimien ylläpidon koordinoinnista. Toistaiseksi eri eliöryhmien nimistötyöryhmät ovat toimineet toisistaan riippumatta, mikä on aiheuttanut samojen nimien käyttöä eri ryhmissä (esimerkiksi "kiitäjät" ja "kehrääjät"). Tavoitteena on, että kukin suomenkielinen lajinimi on yksikäsitteinen.

Suomessa käytettävät lajien ruotsinkieliset nimet pyritään yhtenäistämään Ruotsissa käytettävien nimien kanssa. Pohjana on Ruotsin kansallinen Dyntaxa-nimistötietokanta.

Nimien lisäksi Lajitietokeskuksen ylläpitämään taksonitietokantaan tallennetaan lajien biologisiin ja hallinnollisiin ominaisuuksiin liittyviä tietoja. Tällaisia tietoja ovat mm. uhanalaisuus, rauhoitukset, riistalajit ja lajien statustiedot (tulokaslajit, vakiintuminen, tulotapa).

Yksityiskohtaisempi kuvaus taksonomisten ydintietojen ylläpidosta Lajitietokeskuksen organisaatiossa on kuvattu liitteessä 1.

Eliöseurantojen tiedonhallinnan koordinointi

Lajitietokeskuksen tehtävänä on koordinoida eri eliöryhmiä koskevien valtakunnallisten seurantojen tiedonhallintaa. Seurannoissa kertyvät primaariaineistot voidaan arkistoida joko seurannasta vastaavien organisaatioiden omiin tietokantoihin tai Lajitietokeskuksen data-arkistoon. Tavoitteena on saattaa seurantojen tulokset mahdollisiman nopeasti tutkijoiden ja muiden aineistoista kiinnostuneiden henkilöiden käyttöön avoimena datana, jonka käytölle ei aseteta tarpeettomia rajoituksia.

Havaintotietojen keruu ja jakelu

Lajitietokeskus ylläpitää valtakunnallista tietovarastoa, jonne kopioidaan jäsenorganisaatioiden omista tietokannoista eliölajien esiintymiseen liittyvää havaintotietoa. Havaintotiedon tärkeimpiä yksityiskohtia ovat lajinimet, paikkatiedot (paikannimet ja koordinaatit), havaintoaika sekä havainnoijien nimet sekä esiintymän alkuperä ja tila.

Valtakunnallisesta tietovarastosta havaintotietoa jaetaan käyttäjille sekä alkuperäisessä muodossa että jalostettuina luetteloina, tilastoina, karttoina ja diagrammeina. Uhanalaisten ja muutoin suojelua tarvitsevien lajien salaustarpeet otetaan huomioon tietojen keruussa ja jakelussa. Havaintodatan laatu tarkastetaan ja havaitut virheet korjataan ennen jakelua.

Tietovarasto toimii myös maailmanlaajuisen GBIF-hankkeen Suomen solmuna, jonka kautta jaetaan valtaosa suomalaisista havaintoaineistoista kansainväliseen käyttöön.

Valtakunnallinen havaintojärjestelmä (Vihko)

Lajitietokeskuksen laji.fi-portaalin yhteyteen kehitetään Vihko-niminen havaintopäiväkirja, jonka kautta palveluun rekisteröityneet käyttäjät voivat kirjata eliöhavaintoja sekä Suomesta että ulkomailta. Lajivihkon kehitykseen liittyviä tavoitteita ja näkökohtia ovat:

- Korvaa Hatikan, Löydöksen ja Hyönteistietokannan sekä joukon linnustonseurannan tietojärjestelmiä
- Havaintojen annotointi ja laadunvalvonta
- Havaintojen automaattinen arviointi ja poikkeamien poiminta tarkistukseen
- Havaintotietojärjestelmän mobiiliversio
- Virtalan kuvapankin integrointi osaksi havaintojärjestelmää
- Ulkomaisten havaintojen tallennustekniikat

Data-arkisto

Lajitietokeskus perustaa valtakunnallisen biologisten havaintoaineistojen data-arkiston, jonne tutkijat voivat arkistoida aineistonsa pitkäaikaissäilytystä varten. Lajitietokeskus tarjoaa myös monipuolisia analysointi- ja visualisointipalveluja arkistoidulle datalle.

Datan arkistoinnissa käytettäviä tiedostoformaatteja ei ole vielä määritelty. Oleellista on datan säilyminen luku- ja tulkintakelpoisena pitkien ajanjaksojen yli. Arkistoitavan datan bittitason säilytyksestä vastaa Opetusministeriön alaisena toimiva tieteen tietotekniikan keskus CSC, joka on määritellyt datan pitkäaikaissäilytykseen tarkoitetut PAS-ratkaisut.

Lajitietokeskuksen data-arkistoa kehitetään osana Opetusministeriön Avoin tiede ja tutkimus (ATT) -hanketta.

Datan laadunvalvonta

Lajitietokeskuksen kautta välitettävän näyte- ja havaintoaineiston yhteyteen on tarvetta liittää tietojen luotettavuutta kuvaavia merkintöjä. Näillä annotaatiomerkinnöillä tietoja voidaan rikastaa käyttökelpoisempaan muotoon eri tarkoituksiin.

Annotaatioiden hallintaa varten Lajitietokeskus ylläpitää tietokantaa, johon tallennetaan havaintoihin ja näytteisiin liittyviä merkintöjä. Keskeisiä annotaatioita ovat havaintojen lajinmäärityksen, paikannuksen, aikatietojen tai muiden tietojen luotettavuuteen liittyvät merkinnät, joiden avulla tietoja voidaan suodattaa eri käyttötarkoituksiin.

Annotaatioissa kohteena olevaan näytteeseen tai havaintoon viitataan yksikäsitteisellä URI-tunnisteella. Myös annotaatiomerkinnän tekijällä on uniikki URI-tunniste. Annotaation varsinainen tietosisältö koostuu rajatusta arvojoukosta valitusta luotettavuusmerkinnästä (myös URI) sekä vapaamuotoisesta selitetekstistä, jossa kuvataan tarkemmin annotaation taustaa ja perusteluja.

Aineistojen metatietojen hallinta

Lajitietoa sisältävien aineistojen metatiedot näytetään yhdessä paikassa. Lajitietokeskus tarjoaa välineen metatietojen ylläpitoon niille, joilla ei ole omaa vastaavaa järjestelmää. Muista järjestelmistä metatiedot kerätään lajitietokeskuksen portaaleihin näytettäväksi rajapintojen avulla. Rajapintojen kautta näytettävien metatietojen "rikastaminen" lajitietokeskuksen kannalta relevanteilla tiedoilla voisi olla mahdollista/tarpeellista, mutta tämä ominaisuus on vielä suunnittelematta.

Henkilötietojen hallinta

Eliöhavaintojen luotettavuuden arvioinnissa on tärkeää tietää, kuka havainnon on tehnyt. Lajitietokeskus koordinoi jäsenorganisaatioiden omien tietokantojen henkilötietojen hallintaa, jotta eri lähteistä peräisin olevien aineistojen havainnoijatietoja voidaan käsitellä yhtenäisellä tavalla. Valtaosa (noin 99%) lajitietokeskuksen hallinnoimasta havaintoaineistosta on julkista. Pieni osa (alle 1%) havaintoaineistosta sisältää tietoja, joiden julkisuudelle on asetettava eriasteisia rajoituksia mm. uhanalaisten lajien suojelun, maankäytöllisten ym. syiden vuoksi. Henkilötietojen hallintajärjestelmän avulla määritellään kullekin käyttäjälle oikeudet nähdä ja käyttää suojattua aineistoa omiin tarkoituksiinsa.

Henkilötietojen hallinnan periaatteet on kuvattu tarkemmin liitteessä 3.

Paikkatietojen hallinta

Lajitietokeskus pyrkii yhtenäistämään havaintopaikkojen kuvauksessa käytettäviä paikannimien ja koordinaattien esitysmuotoja, jotta eri lähteistä peräisin olevien aineistojen keskinäinen vertailu olisi mahdollisimman helppoa.

- Yhteydet tausta-aineistoihin (ilmasto, maaperä, topografia, maankäyttö)
- Kuntien ja eliömaakuntien rajojen ylläpito ja jakelu
- YKJ-karttapalvelu

- Koordinaattimuunnokset

Aineistopolitiikka

Lajitietokeskus pyrkii edistämään eliötietojen keruuta ja avointa jakelua määrittelemällä jäsenorganisaatioille suositukset yhtenäisestä aineistopolitiikasta. Valtaosa eliöhavainnoista (yli 99%) voidaan katsoa avoimeksi dataksi, joihin ei liity käyttöä koskevia rajoituksia. Pieni osa havainnoista vaatii suojausta eri syistä.

Lajitietokeskuksen aineistopolitiikkaa ylläpidetään osana kokonaisarkkitehtuuridokumentaatiota. Aineistopolitiikan ensimmäinen versio on julkaistu 26.10.2015, ja se on saatavilla PDF-dokumenttina osoitteesta http://cms.laji.fi/wp-content/uploads/2015/11/Suomen-Lajitietokeskuksen-Aineistopolitiikka.pdf.

Säädöspohja

Toimintaan vaikuttavaa lainsäädäntöä löytyy mm. Lajitietokeskuksen aineistopolitiikasta ja Kansallisen Digitaalisen Kirjaston (KDK) kokonaisarkkitehtuurista sekä Ympäristöministeriön selvityksestä. Lajitietoon vaikuttavilta osilta myös muita mukanaolevia organisaatioita koskeva lainsäädäntö sekä EU-tason säädökset.

TIETOJÄRJESTELMÄARKKITEHTUURI

Nykytila

Kaikilla lajitietoa hallinnoivilla organisaatioilla on omia tietojärjestelmiä, jotka yksinkertaisimmillaan ovat Excel-taulukoita vapaamuotoisissa formaateissa. Isompien organisaatioiden hallinnoima lajitieto on yleensä tallennettu tietokantoihin, joiden tietorakenteiden välillä on paljon eroja.

Lajien esiintymistä koskevien havaintotietojen käsittelyä varten monet organisaatiot ylläpitävät lajiluetteloita, joiden taksonominen ja maantieteellinen kattavuus vaihtelee organisaation tarpeiden mukaan. Näissä luetteloissa lajeihin viitataan yleensä tieteellisen nimen avulla, mutta nimien merkitystä ei ole täsmällisemmin dokumentoitu.

Lajitiedon vaihto organisaatioiden välillä perustuu yleensä ihmistyönä tuotettavien vientitiedostojen lähettämiseen verkon välityksellä tai sopivan tallennusmedian avulla vastaanottajalle. Tiedonvaihto perustuu yleensä kahdenkeskisiin sopimuksiin ja standardoimattomiin dataformaatteihin.

Tavoitetila

Lajitietokeskuksen pyrkii toiminnallaan vakiinnuttamaan seuraavien periaatteiden noudattamisen lajitiedon käsittelyn yhteydessä:

Primaari- ja sekundaarijärjestelmät

Lajitietoa käsittelevät järjestelmät erotellaan primaari- ja sekundaarijärjestelmiin. Kullekin tietoalkiolle määritellään täsmälleen yksi `primaarijärjestelmä`, missä kyseistä tietoa ylläpidetään.

Lajitietokeskus ylläpitää valtakunnallista lajitiedon tietovarastoa, jonne tiedot replikoidaan jäsenorganisaatioiden primaarijärjestelmistä erikseen määritellyn aikataulun mukaisesti. Tietovarastoa voidaan käyttää kansallisena lajitiedon analysointi- ja raportointipalveluna. Tietovarasto on `sekundaarijärjestelmä`, eli sen tietosisältöä ylläpidetään joissakin muissa järjestelmissä.

Primaarijärjestelmistä lajitieto kopioidaan (replikoidaan) määrävälein tietovarastoon ja tarvittaessa myös muihin sekundaarijärjestelmiin. Replikoinnin aikaväli voi vaihdella lähes reaaliaikaisesta päivän, viikon tai jopa kuukausien välein tehtävään replikointiin.

Kukin Lajitietokeskuksen jäsenorganisaatio voi ylläpitää omia primaarijärjestelmiään tai käyttää Lajitietokeskuksen tarjoamia primaaritiedon hallintapalveluita. Näistä tärkeimmät ovat Lajitietokeskuksen uusi havaintotietojärjestelmä `Vihko` ja luonnontieteellisten kokoelmien hallintaan tarkoitettu `Kotka`.

Lajitietokeskuksen ylläpitämistä primaarijärjestelmistä tieto replikoidaan automaattisesti Lajitietokeskuksen tietovarastoon.

Tunnisteet

Kussakin Lajitietokeskuksen yhteydessä toimivassa primaarijärjestelmässä tietoalkioille tulee määritellä pysyvä, globaalisti yksikäsitteinen URI-tunniste, jonka avulla tietoalkioon on mahdollista viitata. Mikäli mahdollista, tunnisteiden tulisi olla ns. HTTP URI -tunnisteita, jolloin niitä voi käyttää myös alkiota kuvailevan tiedon hakemiseen Web-tekniikoilla. Tunnisteen on tarkoitus olla pysyvä, ja sen vuoksi ns. "tyhmä" (informaatiota sisältämätön) ja globaalisti uniikki. Globaalisti uniikiksi tunniste saadaan käyttämällä domain-nimeä nimiavaruuden edessä. Kerran annettu tunniste pysyy samana (alkuosa mukaan lukien) aina, huolimatta sijainnin tai omistajuuden muutoksista. Sen avulla muutetut tiedot ja kaikki historia voidaan säilyttää juuri tunnisteen yhteydessä.

Aikaleimat

Primaarijärjestelmissä kullekin tietoalkiolle tulee kirjata syntyhetki ja viimeisin muokkaushetki vähintään sekunnin tarkkuudella.

Historiatieto

Primaarijärjestelmien tulee kyetä tallentamaan historiatiedot tietoalkioiden koko elinkaaren ajalta. Yksittäisiin tietoalkioihin saatetaan viitata tietyllä ajanhetkellä. Jos tietoalkio viittauksen jälkeen muuttuu, historiatietojen avulla on kyettävä palauttamaan näkyville tietoalkio siinä muodossa kuin se oli viittaushetkellä.

TEKNOLOGIA-ARKKITEHTUURI

Nykytila

Tietokannat

Lajitietokeskuksen primaaristen ydintietojen (sanastot, taksonomia, henkilötiedot) hallinta on toteutettu Helsingin yliopiston tietotekniikkakeskuksen ylläpitämään Oracle-tietokantaan, jonka tietosisällöstä vastaa Luonnontieteellisen keskusmuseon ICT-tiimi. Myös Lajitietokeskuksen Vihko-havaintotietokanta on toteutettu samaan Oracle-ympäristöön.

Lajitietokeskuksen sekundaarisen havainto- ja näytedatan sisältävä tietovarasto on toteutettu Tieteen tietotekniikan keskus CSC:n ylläpitämän Pouta-pilvialusta päällä toimivaan HP Vertica-analyysitietokantaan. Verticasta on toistakseksi käytössä ilmainen Community Edition -versio, mikä rajoittaa tietokannan maksimikoon yhteen teratavuun.

Lajitietokeskuksen jäsenorganisaatioilla on omia primaaritietokantoja, joiden arkkitehtuuria ei ole tarpeen kuvata tässä dokumentissa. Jäsenorganisaatioiden tiedonvaihto Lajitietokeskuksen kanssa toimii alempana kuvattujen rajapintapalvelujen kautta.

WWW-palvelut

Lajitietokeskuksen www-palvelut on toteutettu CSC:n Pouta-pilvialustan päällä toimivien Linux-virtuaalikoneiden avulla. Näihin kuuluvat lajitietokeskuksen www-portaali osoitteessa http://laji.fi sekä datan viennissä ja tuonnissa käytettävät HTTP-protokollaan perustuvat rajapintapalvelut osoitteessa http://api.laji.fi.

GIS-palvelut

Lajitietokeskuksen paikkatietopalveluita varten on CSC:n Pouta-pilvialustalla toimiva Geoserver-paikkatietopalvelin. Sen kautta toteutetaan mm. lajien levinneisyyteen kuuluvia INSPIRE-palveluita.

Tavoitetila

- yliopiston IT-palvelut
- CSC:n pilvipalvelut
- CSC:n data-arkisto
- LUOMUKSEN tietojärjestelmät
- Muiden jäsenorganisaatioiden tietojärjestelmät

LIITE 1. TAKSONOMINEN YDINTIETO

Taustaa

Lajiluettelot ovat Lajitietokeskuksen toiminnassa ydintietoja (*master data*), joiden yhtenäistäminen jäsenorganisaatioiden kesken mahdollistaa automatisoidun lajitiedon vaihdon. Lajiluetteloiden pohjana on biologinen taksonomia, jota on kuitenkin sovellettava modernien tietojärjestelmäarkkitehtuurien asettamien vaatimusten viitekehyksessä.

Biologisen taksonomian perusyksikkö on laji, jolle on kuvauksen yhteydessä annettu nimistösääntöjen mukainen tieteellinen nimi (esim. *Parus major*). Tieteellisen nimen lisäksi lajilla voi olla eri kielissä annettuja yleiskielisiä nimiä (esim. talitiainen, talgoxe, Great Tit). Lajit ryhmitellään oletettujen kehityshistoriallisten sukulaisuussuhteiden mukaan ylemmän tason taksoneiksi kuten suvuiksi, heimoiksi, lahkoiksi ja luokiksi. Ylimmällä tasolla luokittelu kattaa kaikki tunnetut elävät eliöt. Myös lajinsisäisiä taksonomisia yksiköitä kuten alalajeja, muunnoksia ja muotoja on nimetty osalle lajeista.

Lajitiedon hallinnan tarpeisiin biologinen nimistö soveltuu kuitenkin huonosti. Samaan lajiin saatetaan viitata usealla tieteellisellä nimellä, ja toisaalta sama nimi saattaa olla käytössä useassa eri merkityksessä. Lisäksi tieteellistä nimistöä sääteleviä kansainvälisiä säännöstöjä on useita, minkä vuoksi sama tieteellinen nimi voi viitata kahteen eri taksoniin. Esimerkiksi sukujen nimet *Liparis* ja *Prunella* ovat käytössä sekä kasvi- että eläinkunnan nimistössä. Tieteelliset nimet eivät niin ollen sovellu taksonien tunnisteiksi, koska nimet eivät ole pysyviä eivätkä yksikäsitteisiä.

Lajitietokeskuksen lajitiedon hallinta rakentuu Luonnontieteellisen keskusmuseon koordinoiman taksonitietokannan varaan. Taksonitietokannassa ylläpidetään valtakunnallista lajiluetteloa (*master checklist*), joka pyrkii kattamaan kaikki Suomessa tavatut lajit sekä tarvittavan osajoukon globaalista lajistosta.

Taksonitietokannassa kullekin taksonille on määritelty globaalisti yksikäsitteinen tunniste, joka muodoltaan on ns. URI-tunniste (*Uniform Resource Identifier*). Täysimittainen taksonin URI-tunniste on esim.

http://tun.fi/MX.34567

joka samalla on myös web-osoite taksonin kuvaustietoihin (HTTP URI). Tunnisteesta voi eri yhteyksissä käyttää myös lyhennettyä muotoa (Q*ualified Name* eli *QName*) MX.34567. Jopa pelkkä numerokoodi 34567 kelpaa taksonin tunnisteeksi, jos asiayhteydestä muutoin käy ilmi, että kyse on nimenomaan lajitietokeskuksen taksonitunnisteesta. URI-tunnisteen alkuosa

http://tun.fi/MX.

on sama kaikilla taksoneilla, ja se voidaan ohjelmallisesti generoida lyhennetyn muodon perusteella.

Tietokannassa taksonitunnisteeseen kytketään käytössä olevat tieteelliset ja yleiskieliset nimet ominaisuustietoina. Taksonitunniste säilyy muuttumattomana niin kauan kun sen taustalla oleva taksonirajaus ei muutu. Esimerkiksi pelkkä tieteellisen tai yleiskielisen nimen muuttuminen ei muuta taksonitunnistetta, sen sijaan taksonin jakaminen kahdeksi tai useammaksi taksoniksi tai yhdistäminen johonkin toiseen taksoniin luo uuden taksonikäsitteen, jolle annetaan oma tunniste.

Taksonitunnisteiden yhtenäistäminen

Lajitietokeskuksen jäsenorganisaatioilla on omia tietokantoja, joissa ylläpidetään lajiluetteloita kiinnostuksen kohteena olevista eliöryhmistä. Näissä tietokannoissa kullekin taksonille on yleensä annettu jokin järjestelmän sisäinen tekninen tunniste, jolla ei ole informaatiosisältöä kyseisen järjestelmän ulkopuolella.

Lajitiedon vaihdossa organisaatioiden välillä taksonin tunnisteena on perinteisesti käytetty tieteellistä (joskus myös yleiskielistä) nimeä. Tällöin edellämainittujen syiden vuoksi datan siivoamiseen joudutaan käyttämään manuaalista työaikaa.

Lajitietokeskuksen ylläpitämien taksonitunnisteiden tarkoituksena on toimia henkilötunnuksen kaltaisena yksilöllisenä tunnisteena, joka mahdollistaa tietojen automaattisen yhdistämisen useista eri tietolähteistä. Tämän tavoitteen saavuttamiseksi joudutaan kussakin Lajitietokeskukseen liittyvässä organisaatiossa tekemään kertaluontoinen manuaalinen operaatio, jossa organisaation omat taksonitunnisteet kytketään Lajitietokeskuksen määrittelemiin tunnisteisiin.

Yksinkertaisimmillaan kytkentä voidaan toteuttaa lisäämällä organisaation omaan lajiluetteloon ylimääräinen kenttä, johon tallennetaan joko taksonin täysimittainen URI-tunniste, sen lyhennetty muoto (*qualified name*) tai pelkkä numerokoodi. Tiedonvaihdossa kaksi jälkimmäistä muotoa voidaan automaattisesti muuntaa täysimittaiseksi URI-tunnisteeksi.

Käytännössä taksonitunnisteiden yhtenäistäminen on järkevää tehdä eliöryhmittäin. Ensimmäisessä vaiheessa organisaation oman ja Lajitietokeskuksen lajiluettelon välillä ajetaan tieteellisten nimien perusteella täysi ulkoliitos. Tämän vertailun tuloksena saadaan luettelo nimistä, jotka esiintyvät

a) molemmissa luetteloissa
b) vain organisaation luettelossa
c) vain Lajitietokeskuksen luettelossa

Tämän automaattisen vertailun jälkeen kyseisen ryhmän asiantuntijan on käytävä vertailu läpi ja tarkistettava, mitä toimenpiteitä ryhmiin b ja c kuuluville taksoneille vaaditaan. Lisäksi asiantuntijan on varmistettava, että kaikki ryhmään a kuuluvat taksonit ovat yhteneviä myös lajikäsitteen tasolla.

Lajiluetteloiden vertailu on tarkoitus tehdä yhteistyössä jäsenorganisaation ja Luonnontieteellisen keskusmuseon asiantuntijoiden kanssa. Vertailun tuloksena syntyy luettelo, jossa organisaation omat taksonitunnisteet on kytketty Lajitietokeskuksen URI-tunnisteisiin. Tämä luettelo voidaan sitten viedä organisaation tietokannassa ylläpidettävän lajiluettelon osaksi.

Taksonitunnisteiden käyttö ja ylläpito

Kun organisaation taksonitunnisteet on yhtenäistettu ylläkuvatulla tavalla, niitä voidaan käyttää lajitiedon vaihdossa Lajitietokeskuksen kanssa molempiin suuntiin. Käytännön arjessa lajeihin liittyvää tietoa (esimerkiksi havaintotietoa) kerätään ja käsitellään jatkossakin tieteellisen tai yleiskielisen nimen perusteella. Tietojärjestelmien välisen tiedonsiirron yhteydessä nimien tulkinta taksonitunnisteiksi pyritään kuitenkin tekemään ensisijaisesti ennen tiedonvaihtoa.

Luonnontieteellinen keskusmuseo pyrkii ylläpitämään taksonitietokannassa kattavaa luetteloa taksonien synonyymiminimistä, joita voidaan käyttää apuna nimien muuntamisessa taksonitunnisteiksi. Tietokannasta on myös mahdollista hakea luettelo niistä tieteellisistä tai yleiskielisistä nimistä, joiden merkitys ei ole yksikäsitteinen. Tämän luettelon avulla käyttäjiä voidaan ohjata välttämään epäselvien taksoninimien käyttöä lajitiedon ylläpidossa.

Taksonomisen tutkimuksen edistyessä Lajitietokeskuksen lajiluetteloon joudutaan aika ajoin tekemään muutoksia. Näiden muutosten automaattinen synkronointi jäsenorganisaatioiden omiin järjestelmiin täytyy suunnitella erikseen. Lajitietokeskus säilyttää myös pääluettelosta poistuneet taksonitunnisteet tietokannassa ja kuvaa niiden suhteet korvaaviin taksonitunnisteisiin.

LIITE 2. PAIKKAAN JA AIKAAN LIITTYVÄ YDINTIETO

Tähän liitteeseen kuvataan Lajitietokeskuksen käyttämät paikka- ja aikatietoihin liittyvät käytännöt.

LIITE 3. HENKILÖIHIN LIITTYVÄ YDINTIETO

Henkilötietoja tarvitaan lajitietojen hallinnassa monessa eri roolissa. Lajien esiintymistä koskeva tausta-aineisto perustuu valtaosaltaan luontoharrastajien sekä biologian ja ympäristöalan ammattilaisten tekemiin havaintoihin, joista osaan saattaa liittyä kokoelmiin tallennettu näyte. Havainnoijien ja näytteiden kerääjien tietojen säilyttäminen ja julkaiseminen havaintojen yhteydessä on vakiintunut käytäntö, jonka avulla aineiston kerääjät saavat asianmukaisen tunnustuksen työstään. Sen lisäksi henkilötietojen avulla voidaan arvioida havaintojen uskottavuutta, koska eri henkilöiden taidot lajintuntemuksen ja muiden havainnointiin liittyvien taitojen suhteen vaihtelevat suuresti.

Henkilötietojen hallintaa tarvitaan myös havaintojen jatkokäsittelyssä. Havaintoaineiston laadunvalvonta perustuu paljolti auktorisoitujen henkilöiden havaintoihin lisäämiin annotaatioihin, joiden perusteella havaintojen käyttöä eri yhteyksissä voidaan säädellä. Julkisuuslain säädösten mukaisesti osa havaintoaineistosta on lisäksi luokiteltu salassapidettäväksi. Tällöin salattuja havaintoja pääsevät näkemään ainoastaan auktorisoidut henkilöt.

Henkilötietojen hallinnan haasteena on se, että sama henkilö saattaa käyttää useita erillisiä Lajitietokeskuksen yhteydessä toimivia tietojärjestelmiä, joissa kussakin voi olla oma erillinen käyttäjätietojen hallintajärjestelmä. Saman käyttäjän eri järjestelmissä olevat tiedot on tarve yhdistää samaan henkilöön monestakin eri syystä.

Henkilöihin liittyvien ydintietojen hallinnan koordinointi tehdään Luonnontieteellisen keskusmuseon ontologiakirjaston kautta. Kullekin yksilöllisesti tunnistetulle henkilölle määritellään URI-tunnus, joka on muotoa http://tun.fi/MA.123. Tähän tunnukseen voidaan kytkeä kyseisen henkilön identiteettitiedot sekä henkilön eri tietojärjestelmissä käyttämät käyttäjätunnukset.