Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Ökoloogia ja maateaduste instituut

Botaanika osakond

Merilin Radvilavicius

SUURANDMETE KASUTUS MAKROÖKOLOOGIAS

Bakalaureusetöö

Bioloogia ja elustiku kaitse

12 EAP

Juhendaja: prof Meelis Pärtel

Tartu 2024

INFOLEHT

Suurandmete kasutus makroökoloogias

Suurandmed ja andmeteaduse meetodid võimaldavad pakkuda vastuseid küsimustele, millega on bioloogid juba aastakümneid rinda pistnud. Käesoleva töö eesmärk on anda kirjanduse põhjal ülevaade, mida kujutavad suurandmed endast ette makroökoloogia kontekstis. Selleks teen esmalt jalutuskäigu suurandmete olemusse ja selle võidukäiku makroökoloogiasse, misjärel kirjeldan suurandmeid just makroökoloogia kontekstis: mis need on ja kust need tulevad. Annan ka ülevaate suurandmete kasutamise võimalustest, kitsaskohtadest ja võimalikest lahendustest, mis võivad ette tulla erinevates suurandmetega töötamise etappides. Kõige selle juures illustreerin töö eesmärke graafilise analüüsiga, kus annan ülevaate makroökoloogiliste uuringute temaatilisest rühmitumisest suurandmete kontekstis.

Märksõnad: makroökoloogia, suurandmed, masinõpe, andmehaldus, andmebaasid

CERCS teaduseriala kood: B270 Taimeökoloogia; B280 Loomaökoloogia; B320 Süstemaatiline botaanika, zooloogia, zoogeograafia; P175 Informaatika, süsteemiteooria

The use of big data in macroecology

Big data and data science methods can provide answers to questions that biologists have been struggling with for decades. This thesis aims to provide a literature-based overview of what big data represent in the context of macroecology. To do this, I will first take a walk through the nature of big data and its triumphal march into macroecology, whereupon I describe the big data specifically in the context of macroecology - what are they and where they come from. In addition, the thesis provides an overview of the opportunities, challenges, and possible solutions to the use of big data that can occur at different stages of working with big data. I illustrate the aims of the work with a graphical analysis of the thematic clustering of macroecological research in the context of big data.

Keywords: macroecology, big data, machine learning, data management, databases

CERCS research field code: B270 Plant ecology; B280 Animal ecology; B320 Systematic zoology, zoogeography; P175 Informatics, systems theory

Sisukord

1. SISSEJUHATUS

Makroökoloogia (ingl *macroecology*) on võrdlemisi noor teadusharu, mille formaalne sünd nägi ilmavalgust Browni ja Maureri (1989) avaldatud uuringus. Sellest ajast alates on makroökoloogiat kui terminit proovitud korduvalt defineerida. Siinkohal toon välja ühe õnnestunud definitsiooni – makroökoloogia on statistiline lähenemine ökoloogiale, mis uurib suurel ruumiskaalal liikide, populatsioonide ja ökosüsteemide vahel esinevaid ökoloogilisi protsesse ja mustreid (Beck *et al*., 2012; Smith *et al*., 2008). Ruumiskaalaline ülesus on see, mis eristab makroökoloogiat traditsioonilisest ökoloogiast – lokaalsel tasandil tehtud uuringuid on võimalik kasutada suureskaalalistes uuringutes, samal ajal saab suurel skaalal tehtud uuringutega seletada lokaalsel tasandil toimuvaid ökoloogilisi protsesse (Laanisto & Pärtel, 2019). Piiripealse valdkonnana on makroökoloogia lähedaseks naabriks teistele valdkondadele nagu biogeograafia (ingl *biogeography*) ja ökoinformaatika (ingl *ecoinformatics*; Michener & Jones, 2012).

Aina enam nähakse potentsiaali kasutada suurandmeid (ingl *big data*) just makroökoloogilistes uuringutes (Wüest *et al*., 2020). Teadmised suurandmete kohta on ajas märgatavalt täienenud, mis annab hea võimaluse neid rakendada ka makroökoloogias. Siiski ripub õhus palju küsimusi, kuna makroökoloogias pole suurandmete kasutamisel juurdunud veel ühtseid arusaamasid ja tavasid, vaid paljuski toimub omaette nokitsemist (Michener & Jones, 2012). Makroökoloogia ja suurandmed kui terminid on vaid loetud aastakümned vanad, mistõttu pole nende kontseptsioonid rahvusvahelises ega eestikeelses teadusringkonnas veel selgelt välja kujunenud. Seetõttu annab töö panuse emakeelsesse teadusterminoloogiasse, samal ajal on tööga seotud olulisematele terminite esmamainimisel toodud välja ka nende inglisekeelsed vasted.

Töö eesmärgiks on tuua kokku bioloogia ja andmeteadus selgitamaks, kuidas pakub viimane usaldusväärseid vastuseid küsimustele, millega on bioloogid juba aastakümneid maadelnud. Töös annan ülevaate, mida kujutavad suurandmed endast ette makroökoloogia kontekstis: millised need on ja kust need tulevad. Sellele lisaks annan läbilõike suurandmete kasutamise võimalustest, kitsaskohtadest ja võimalikest lahendustest, mis tulevad ette erinevates andmete eluringi etappides. Töö eesmärke illustreerin graafilise analüüsiga, andmaks ülevaadet makroökoloogiliste uuringute temaatilisest rühmitumisest suurandmete kontekstis. Töös koondasin otsingutega võimalikult täieliku nimestiku teemakohastest teadusartiklitest, mille seas tegin kriitilise valiku, kuna sel teemal toimuvad teaduses endiselt aktiivsed arutelud.

2. SUURANDMED

Suurandmeid kui terminit kasutati esimest korda Coxi ja Ellsworthi (1997) poolt, ent mõistagi on suurandmeid produtseeritud ja kasutatud juba varem. Termini vormumisest alates on üritatud pakkuda sellele mitmesuguseid definitsioone eri valdkonna spetsialistide poolt, ent kuldset keskteed pole siiani kujunenud (Chen *et al*., 2014). Ebakindlusele vaatamata on terminit "suurandmed" kasutatud laialdaselt kõikvõimalikes valdkondades, mis võib muuta suurandmete kontseptsiooni veelgi ebamäärasemaks (Farley *et al*., 2018; Favaretto *et al*., 2020). Suurandmete määratlemise teeb keeruliseks ka asjaolu, et tehnoloogia ja analüütilise võimekuse areng sunnib iga edusammu järel suurandmete olemusele uuesti otsa vaatama (LaDeau *et al*., 2017). Siiski võib öelda, et suurandmete iseloomustamiseks on aja jooksul kujunenud kaks domineerivat raamistikku (Farley *et al*., 2018). Kaht raamistikku eristab nende konkreetsus suurandmete defineerimisel – kui üks piirdub kirjeldamisel ühesõnaliste omadustega ja on võrdlemisi rangete piiridega, siis teine kasutab selle asemele laiemat kirjeldust ja vabamat käsitlust, puudutades siiski kõiki suurandmete omadusi (Farley *et al*., 2018).

2.1. Suurandmed kui mitme omadusega fenomen

Suurandmete kirjeldamisel tahetakse piirduda vaid nende mahuga (ingl *volume*), ent sellest ei piisa suurandmete sügavama olemuse iseloomustamiseks (LaDeau *et al*., 2017). Varastel 2000. aastatel tutvustas IMB kontseptsiooni, mis kirjeldab suurandmeid läbi kolme erineva inglise keeles V-tähega algava omaduse: maht, mitmelaadilisus (ingl *variety*) ja kiirus (ingl *velocity*; Laney, 2001). Aja jooksul on nimekirja täiendatud veel mitmete omadustega (Kitchin & McArdle, 2016), ent kõige enam kasutatakse neist tõepärasust (ingl *veracity*) ja väärtust (ingl *value*; Farley *et al*., 2018; Wüest *et al*., 2020).

Andmete **mahu** all mõistetakse andmehulga suurust arvuti infoühikutes, enamasti on mahud terabaitide ja eksabaitide (1 EB ≈ 1 000 000 TB) piires, kusjuures terabaidised personaalarvutid said kättesaadavaks alles 2000. aastate keskel (Farley *et al*., 2018; LaDeau *et al*., 2017). **Mitmelaadilisus** kirjeldab andmete endi mitmekesisust ja nendevahelisi keerulisi seoseid, mis on saadud kombineerituna eri allikatest pärit andmetega, sisaldades lisaks arvulistele väärtustele ka näiteks teksti, pilti, videot jms (Farley *et al*., 2018; LaDeau *et al*., 2017; Wüest *et al*., 2020). **Kiirus** väljendab andmete tootmise, kogumise ja analüüsimise kiirust, viidates nende reaalajas vastavusele kui ka kättesaadavusele (Farley *et al*., 2018; LaDeau *et al*., 2017; L'Heureux *et al*., 2017; Wüest *et al*., 2020). **Tõepärasuse** all

mõistetakse seda, kuivõrd tõesed ja kvaliteetsed on andmed ning andmeallikad (Farley *et al*., 2018; L'Heureux *et al*., 2017; Wüest *et al*., 2020). Viimaks – **väärtus** näitab, milliseid teadmisi on neist andmetest võimalik saada, kusjuures väärtus sõltub kontekstist ja püstitatud eesmärkidest ning andmeid saab väärindada läbi oskusliku analüüsimise (Gandomi & Haider, 2015; Wüest *et al*., 2020).

Kuigi suurandmete omaduste loendil põhinev raamistik on juba võrdlemisi omaks võetud ja leidnud laialdast kasutust, leidub sellele ka kriitikuid. Väljaspool andmeteaduse valdkonda peetakse seda raamistikku liialt tehniliseks (Favaretto *et al*., 2020). Mainitud viiele V-tähega algavale omadusele (ingl *The Five Vs of Big Data*) on lisaks pakutud veel omadusi nagu valiidsus (ingl *validity*), visualiseerimine (ingl *visualisation*) ja haavatavus (ingl *vulnerability*; Al-Mekhlal & Ali Khwaja, 2019), ent need ei iseloomusta enam suurandmete olemust, vaid on rõhuasetusega neid puudutavatele kitsaskohtadele (Kitchin & McArdle, 2016).

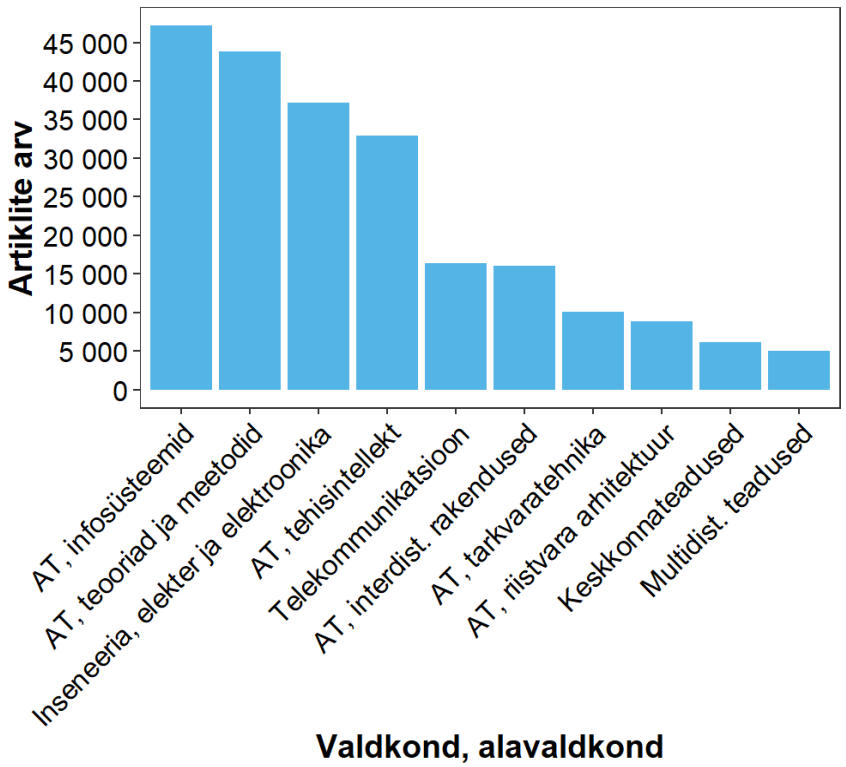
2.2. Suurandmed kui kompleksne suuremahuline andmekogum

Teise raamistiku põhjal ei läheneta suurandmete kirjeldamisel läbi ühesõnaliste omaduste, vaid rõhuasetusega, et suurandmed on ulatuslik kompleksne andmestik, mille analüüsimisel jääb tavapärastest analüüsimeetoditest ja arvutusvõimekusest väheks (Chen *et al*., 2014; Hampton *et al*., 2013; Snijders *et al*., 2012; Tiit & Tooding, 2019). Üheks peamiseks suurandmete erinevuseks traditsiooniliste uurimisandmetega on asjaolu, et suurandmetega soovitakse püstitatud uurimishüpoteeside asemele otsida hoopis neis esinevaid võimalikke mustreid (Favaretto *et al*., 2020; Tiit & Tooding, 2019). Tasub ka välja tuua, et suurandmed tekivad ja muutuvad ajas pidevalt (Kitchin & McArdle, 2016; Tiit & Tooding, 2019), millele viitab ka nende kiiruse omadus. Suurandmed on oma sisult ja vormilt paindlikumad, olles sageli mitmemõõtmelised (Fan *et al*., 2014; Kitchin & McArdle, 2016; Tiit & Tooding, 2019), mida iseloomustab ka nende maht ja mitmelaadilisus. Suurandmete analüüsimine ja töötlemine on ressursimahukas ning metodoloogiliselt spetsiifiline (Tiit & Tooding, 2019), kuna suurandmete kvaliteeti mõjutavad nii vead kui ka kallutatused (Kitchin & McArdle, 2016; Wüest *et al*., 2020).

3. SUURANDMETE TULEK MAKROÖKOLOOGIASSE

Suurandmed on alguse saanud arvutiteadusest (LaDeau *et al*., 2017), ent sellele lisaks on suurandmed leidnud laia kasutust ka teistes valdkondades nagu (personaal)meditsiin (Hampton *et al*., 2013; Wüest *et al*., 2020), psühholoogia (Harlow & Oswald, 2016), majandus (Fan *et al*., 2014) ja loodusteadused (Farley *et al*., 2018). Ajas kasvanud digitaliseeritud andmete hulk on sillutanud tee ka uute valdkondadeni nagu arvutuslik sotsiaalteadus (Lazer *et al*., 2009) ja digihumanitaaria (Ewing *et al*., 2016).

Selleks, et uurida suurandmete kasutust, on üheks võimaluseks vaadata, millise valdkonna teadusartiklites on märksõna "suurandmed" kasutatud. Selle uurimiseks teostasin 24. jaanuaril 2024. aastal kirjanduse andmebaasis Web of Science päringu. Päringust selgus, et kümnest enim suurandmeid käsitlevast Web of Science'i valdkonnast on kuus mõnest arvutiteaduse alavaldkonnast (Joonis 1). Umbes 190 000 avaldatud suurandmeid käsitlevatest teadusartiklitest moodustavad arvutiteaduses avaldatud pea 85%. Neljas alavaldkonnas – "infosüsteemid", "teooriad ja meetodid", "tehisintellekt" (kõik arvutiteadus) ning "elekter ja elektroonika" (inseneeria) – ilmunud teadusartiklid moodustavad ligi kolmandiku kõikidest ilmunud suurandmeid käsitlevatest teadusartiklitest.



Joonis 1. *Suurandmeid käsitlevate teadusartiklite arv kümnes enimlevinud kirjanduse andmebaasi Web of Science alavaldkonnas 24. jaanuaril 2024. aastal, AT - arvutiteadus. Päringuga* ALL=*(“*big data*”* OR *“*big-data*”* OR bigdata*) otsiti üle kõikide väljade märksõna “suurandmed”, hõlmates selle kõiki ingliskeelseid kirjapilte. Päringu teostamise metoodika detailid on toodud tekstis.*

Loodusteadustes on pikema suurandmete kasutamise traditsioonidega näiteks geograafid (Kitchin, 2013), geneetikud (Fan *et al*., 2014), bioinformaatikud (Jones *et al*., 2006) ja klimatoloogid (Schnase *et al*., 2017). (Makro)ökoloogias hakkavad suurandmetel põhinevad uuringud ja selleks tarvilike analüüsimeetodite kasutamine leidma aina laiemat kasutust, sest ka neis valdkondades on kättesaadavate andmete hulk kümnendite jooksul kasvanud (Wüest *et al*., 2020). Suurandmed võimaldavad paremini liikuda makroökoloogia fundamentaalse eesmärgi – mõista elusloodus eri ruumiskaaladel esinevaid interaktsioone ja mustreid – poole (Bruelheide *et al*., 2018; Farley *et al*., 2018; Leonelli, 2019). Suurandmed lubavad leida lahendusi küsimustele, milleni traditsiooniline ökoloogia sageli ei küündi. Suurandmetega on näiteks võimalik paremini uurida liikide ohustatust (Bachman *et al*., 2024; McCleery *et al*., 2023; Shipley *et al*., 2018), liikide levikut (Morueta-Holme & Svenning, 2018), inimmõju ökosüsteemidele (McCleery *et al*., 2023), keskkonna- ning globaalmuutusi (Andrew *et al*., 2017; Franklin *et al*., 2017) ja teostada globaalsete mustrite sünteese (Cheruvelil *et al*., 2017; Heberling *et al*., 2021; Keppel *et al*., 2021).

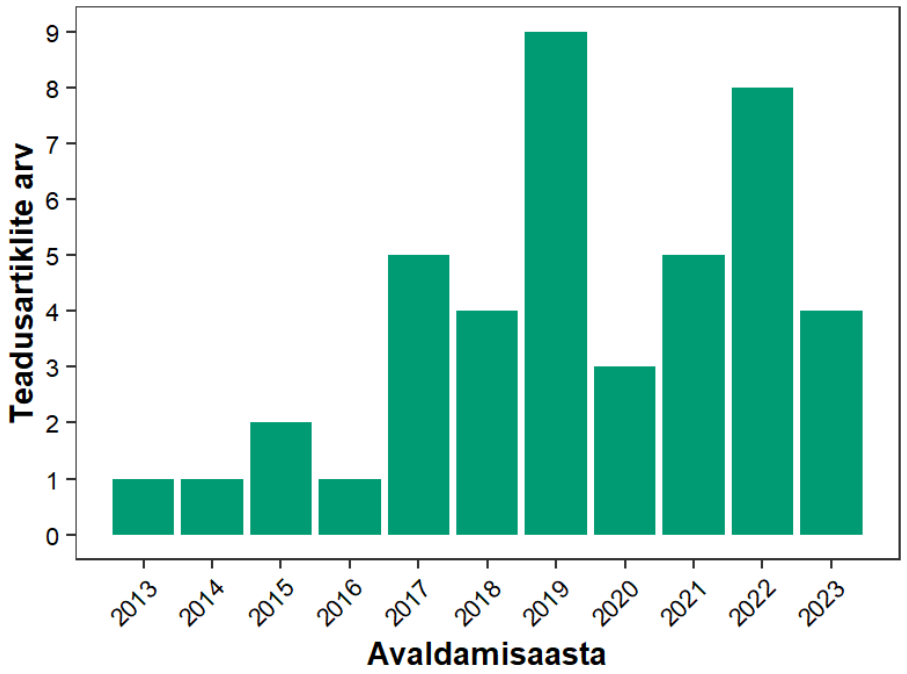
3.1. Suurandmeid käsitlevad makroökoloogilised uuringud

Saamaks ülevaadet suurandmeid käsitlevatest makroökoloogilistest uuringutest, teostasin 24. jaanuaril 2024. aastal Web of Science'is päringu, kus üle kõikide väljade (teema, pealkiri, lühikokkuvõte jms; tähistatud *ALL*) otsisin suurandmete koosesinemist makroökoloogiaga:

ALL=((*macroecolog* OR* "*macro-ecolog**") *AND* ("*big data*" *OR* "*big-data*" *OR bigdata*)),

kus arvestasin kõikvõimalike kirjapiltide (kokku- ja lahkukirjutatuna ning sidekriipsu kujul) ja tüvelõppudega (*) nagu makroökoloogi**a** (*macroecolog**y***) ja makroökoloogi**line** (*macroecolog**ical***). Samuti kasutasin Boole'i loogikaoperaatoreid. *AND*-loogikaoperaator ("ja") piirab päringut, sundides kaasama mõlemaid loogikalause pooli, antud juhul nii makroökoloogiat kui ka suurandmeid. *OR*-loogikaoperaator ("või") laiendab päringut, lubades kaasata nii üht kui teist või kõiki märksõnu, antud juhul võis päring kaasata kas üht, mõnd või kõiki märksõna "suurandmed" kirjapilte. Selleks, et päringuga saaks saavutatud ka soovitud tulemused, tuli päringus paigutada sulud õigete sõnakomplektide ümber. Viimaks pandi märksõnad jutumärkidesse tagamaks, et neid otsitaks kõrvuti olevatena ehk nende vahel poleks teisi sõnu.

Päringule sain vasteks 43 teadusartiklit, mille avaldamised jäävad ajavahemikku 2013. – 2023. aasta (Joonis 2). Üheksa teadusartiklit on ilmunud 2019. aastal ja pea sama palju ehk kaheksa teadusartiklit 2022. aastal. Päringu tegemise hetkel polnud käesoleval aastal avaldatud veel ühtegi teemakohast teadusartiklit. Väikesele valimile vaatamata võib Pearsoni korrelatsioonanalüüsi põhjal öelda, et suurandmete kaasamine makroökoloogiasse on kasvutrendis ($r = 0{,}66$, $p = 0{,}027$) – peale 2016. aastat on senise ühe-kahe teadusartikli asemel ilmunud aastas kolm kuni üheksa teadusartiklit.



Joonis 2. *Suurandmeid käsitlevate makroökoloogiliste teadusartiklite arv avaldamisaastati kirjanduse andmebaasis Web of Science 24. jaanuaril 2024. aastal. Pearsoni korrelatsioonikordaja r = 0,66 viitab võrdlemisi tugevale seosele. Nullhüpoteesi (H_0) korral, kus teadusartiklite arvu muutust aastatel 2013 – 2023 ei ole, lükkan olulisuse tõenäosuse p = 0,027 ($p < \alpha = 0,05$) juures ümber nullhüpoteesi ja toetuse saab alternatiivne hüpotees (H_1) – aastatel 2013 – 2023 avaldatud teadusartiklite arvud kasvavad ajas statistiliselt oluliselt.*

3.2. Märksõnade analüüs

Illustreerimaks suurandmeid käsitlevate makroökoloogiliste uuringute temaatilist jaotust, kasutasin Web of Science'ist saadud kirjete märksõnu, mis omakorda said kaasatud graafilisse analüüsi VOSvieweri tarkvarasse (Eck & Waltman, 2009). Päringu koostasin samade põhimõtete järgi (kirjapildid, tüvelõpud, loogikaoperaatorid, sulud ja sõnade koosesinemine), nagu sai eelnevalt kirjeldatud. Kuna makroökoloogia pole selgepiiriline teadus, kaasasin päringusse ka teisi piiripealseid valdkondi nagu biogeograafia ja ökoinformaatika. Lisaks kaasasin päringusse makroökoloogia põhilisema uurimisobjekti ehk elurikkuse (ingl *biodiversity*). Teostasin 24. jaanuaril 2024. aastal Web of Science'is päringu, kus üle kõikide väljade otsiti suurandmete ja/või andmemahukuse (ingl *data-intensive*) koosesinemist makroökoloogia, ökoloogia (ingl *ecology*), elurikkuse, ökoinformaatika ja/või biogeograafiaga:

ALL=((“*big data*” *OR* “*big-data*” *OR bigdata OR* “*data-intensive*” *OR* “*data intensive*” *OR dataintensive*) *AND* (*macroecolog* OR* “*macro-ecolog**” *OR ecolog* OR biodiversity OR* “*bio-diversity*” *OR ecoinformatic* OR* “*eco-informatic**” *OR biogeograph* OR* “*bio-geograph**”)).

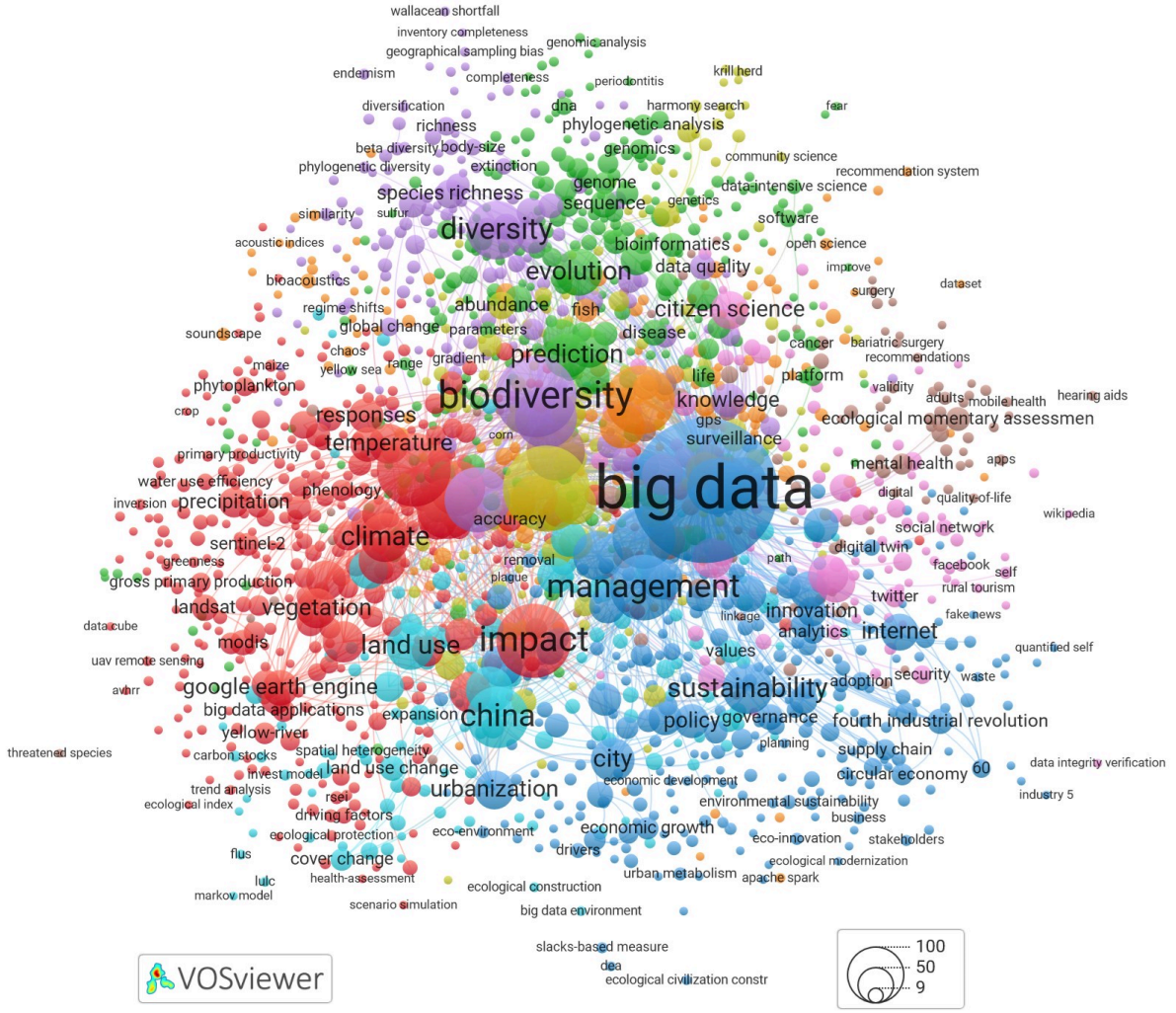
Päringule vastas 3511 teadusartiklit. Selleks, et saada kirjed VOSviewerisse, eksportisin need lihttekstifailina (.txt). Kuna soovisin kaasata kõiki teadusartikleid ja viidatud allikaid, siis Web of Science’i iseärasustest tulenevalt tuli lihttekstifaile eksportida 500 kirje kaupa, tehes lõpuks kokku kaheksa lihttekstifaili. Bibliograafilistel andmetel põhineva analüüsi tegemisel lasin VOSvieweril lugeda kõiki saadud kirjete lihttekstifaile.

Enne lõplikku analüüsi tuli teostada märksõnade puhastamine, vältimaks samatähenduslike märksõnade kordumisi. Selleks laadisin VOSvieweri analüüsi ettevalmistamise viimasel etapil alla lihttekstifaili, mis sisaldas Web of Science’i vastete põhjal koostatud märksõnu, nende esinemiste sagedusi ja seoste tugevusi. Tabelis, kus teostasin puhastamist, oli kaks tulpa: esimeses esialgsed märksõnad (*Label*) ja teises vastavalt vajadusele asendav märksõna (*Replace by*), mis pärines esimesest tulbast. Tulpade pealkirjad olid inglise keeles, sest VOSviewer on programmeeritud just selles keeles aru saama. Vastavalt oma parimale teadmisele teostasin märksõnade puhastamise, järgides allolevaid põhimõtteid:

1. Eelistasin ainsust mitmusele (näiteks “*area*” ja “*areas*”).
2. Eelistasin lahkukirjutamist sidekriipsuga ja kokkukirjutatud variandile (näiteks “*big data*” ja “*big-data*” ja “*bigdata*”).
3. Eelistasin täispikkuses nimetust selle akronüümile (näiteks “*support vector machine*” ja “*svm*”).
4. Eelistasin Briti inglise keelt Ameerika inglise keelele (näiteks “*digitalisation*” ja “*digitalization*”).
5. Koondasin omavahel sünonüümid või muidu sarnase tähendusega sõnad (näiteks märksõnad “*modelling/modeling*” viisin kokku märksõnaga “*model*”, märksõna “*analytics*” märksõnaga “*analysis*” ja “*microbiota*” märksõnaga “*microbiome*”).
6. Ei koondanud alamärksõnu suurema tähendusega märksõna alla kokku (näiteks “*drough*”, “*drough stress*” ja “*drough tolerance*”), et vältida liigset lihtsustumist.

Peale märksõnade puhastamist andsin sellesama sünonüümide faili sisu VOSviewerile lihttekstifaili kujul ette, kus olid alles jäetud vaid need read märksõnadest, kus olid välja toodud asendamised. Selle tulemusena jäi algsest 16 131 märksõnast järgi 15 896. Lisaks

määrasin märksõna esinemise sageduseks kolm või enam korda üle kõikide Web of Science'i kirjete. Selle lävendi ületas lõpuks 1717 märksõna, mis said analüüsi kaasatud ja rühmitatud (Joonis 3). Rühmitamisel määrasin klastri minimaalseks suuruseks 100 teadustööd, et saaksid eristatud peamised suured uurimisteemade valdkonnad. VOSvieweri rühmitamise aluseks on koosesinemise maatriksi pealt kalkuleeritud sarnasusmaatriks, mille peal rakendatakse omakorda VOS-i kaardistamise tehnikat (Eck & Waltman, 2009). Viimaks optimiseeritakse saadud tulemust, sealjuures rakendades peakomponentanalüüsi (Eck & Waltman, 2009).



Joonis 3. *Üldmulje suurandmetega seotud märksõnadest makroökoloogilistes uuringutes. Klastrite värv eristab erinevate uurimisvaldkondade rühmi ja ruumiline paigutus väljendab terminite kooskasutamist ning uurimisvaldkondade omavahelist seotust. Seejuures väljendab iga üksiku märksõna ruumiline paigutus seotust oma ja ka naaberklastri märksõnadega. Palli suurus väljendab märksõnade kasutuskorda teadusartiklites. Omavahel tugevamini seotud märksõnad on ühendatud joonega. Käesolev joonis annab seostest üldmulje, suurema lahutusega, kus tekstid on paremini loetavad, versiooni leiab Lisast 1.*

Kokku moodustus üheksa klastrit (Joonis 3). Klastrite koondumine joonise keskele näitab, et moodustunud valdkonnad omavad omavahel tugevaid seoseid ja kattuvusi. Visuaalselt hõlmavad suuremal hulgal märksõnu tumesinine, punane ja tumeroheline klaster. **Tumesinisesse** klastrisse koonduvad suurandmete ja tehnoloogia valdkonnaga seotud märksõnad, nende hulgas ka töökeskne märksõna "suurandmed". Sellele lisaks on tumesinises klastris esil märksõnad nagu "majandamine" (ingl *management*), "jätkusuutlikkus" (ingl *sustainability*) ja "linnastumine" (ingl *urbanization*). **Punane** klaster hõlmab endas keskkonna, kliima ja seirega seonduvat, kus oli laialdaselt kasutatud märksõnu nagu "mõju" (ingl *impact*), "kliima" (ingl *climate*) ja "temperatuur" (ingl *temperature*). **Tumeroheline** klaster puudutab evolutsiooni ja geneetikaga seonduvaid märksõnu nagu "prognoos" (ingl *prediction*), "bioinformaatika" (ingl *bioinformatics*) ja "järjestus" (ingl *sequence*).

Mainitud klastrite vahele jäävad ka teised nagu helesinine ja lilla klaster, mille märksõnad asuvad kohati kõrvaloleva(te) klastri(te) piirimail (Joonis 3). **Helesinine** klaster hõlmab endas maakasutusega (ingl *land use*) seotud märksõnu nagu "laienemine" (ingl *expansion*) ja "maakasutuse muutus" (ingl *land use change*), aga üllatuslikult esineb ka märksõna "Hiina" (ingl *China*). **Lillasse** klastrisse on koondunud erinevad looduskaitset (ingl *conservation*), mitmekesisust (ingl *diversity*) ja selle mustreid väljendavad märksõnad nagu "elurikkus", "liigiline mitmekesisus" (ingl *species richness*) ja "fülogeneetiline mitmekesisus" (ingl *phylogenetic richness*).

Viimaks on roosa, pruun, oranž ja kollakasroheline klaster, milles on rühmitunud temaatilised märksõnad, ent need on väga tugevalt seotud ka naaberklastritega (Joonis 3). **Roosa** klaster hõlmab endas kommunikatsiooni ja kaasatusega seonduvaid märksõnu nagu "informatsioon" (ingl *information*), "sotsiaalmeedia" (ingl *social media*) ja "harrastusteadus" (ingl *citizen science*). **Pruuni** klastrisse on koondunud eelkõige meditsiini ja tervisega seotud märksõnad nagu "vaimne tervis" (ingl *mental health*) ja "depressioon" (ingl *depression*), seejuures on klastris enim kasutuses märksõna "masinõpe" (ingl *machine learning*). **Oranž** klaster hõlmab endas ökoloogia ja rändega (ingl *migration*) seotud märksõnu nagu "liikumine" (ingl *movement*), "võrgustik" (ingl *network*) ja "GPS" (ingl *Global Positioning System*). Viimases, **kollakasrohelises**, klastris on mudeldamisega seotud märksõnad nagu "optimeerimine" (ingl *optimization*) ja "algoritm" (ingl *algorithm*), aga ka "GIS" (ingl *geographic information system*).

Märksõnade ruumilisest paigutusest on märgata temaatiliste märksõnade koondumist (Joonis 3). Joonise paremale küljele on koondunud erinevad meditsiini ja tervisega ning vasakule kliima ja keskkonnaga seonduvad märksõnad. Joonise ülaosas on elurikkuse ja evolutsiooniga ning alaosas jätkusuutliku majandamise ja linnastumisega seonduvad märksõnad. Keskele on koondunud andmeteaduse ja selle meetoditega seonduvad märksõnad. On märgata ka erinevate valdkondade sulandumisi, näiteks on joonise paremasse ülanurka koondunud ökoloogiliste andmete ja selle meetoditega seonduvad märksõnad.

Jooniselt 3 on näha, et suurandmete klastris (tumesinine) on bioloogilistest märksõnadest näiteks "ökosüsteemide talitlused" (ingl *ecosystem functions*), nende majandamine ja "säilenõtkus" (ingl *resilience*). Lisaks on suurandmete klastriga seotud ka teised klastrid koos nende bioloogiliste märksõnadega. Näiteks on tumerohelisest klastrist seotud "evolutsioon" ja "bioinformaatika", lillast "elurikkus" ja "looduskaitse", punasest "globaalmuutused kliimas" (ingl *climate change*) ja "taimkate" (ingl *vegetation*) ning helesinisest "maakasutus".

Nagu graafiline analüüs illustreerib, siis on suurandmete käsitlus makroökoloogia ja seda hõlmavates piiriteadustes väga laiahaardeline, kus saavad puudutatud evolutsioon, kliimamuutused, elurikkus, bioinformaatika, looduskaitse jms. Seega võib öelda, et suurandmeid leidub kõikjal ja makroökoloogial on palju potentsiaali, et uurida nende abil seniteadmata ökoloogilisi protsesse ja mustreid ning lahendada põletavaid keskkonnaprobleeme.

4. MAKROÖKOLOOGILISTE SUURANDMETE TÜÜBID JA ALLIKAD

Nagu sai mainitud, siis suurandmed on mitmelaadilised, mis tähendab, et suurandmete sisu ja vorm on varieeruv ning need pärinevad erinevatest allikatest. Seejuures pole makroökoloogiliste suurandmete liigitamisel üht selget tava välja kujunenud, ent siiski on neid võimalik liigitada näiteks statistikas kasutatud lähenemise põhjal. Süsteeme, kust makroökoloogilised suurandmed tulla võivad, on laias laastus kolm: vaatlused, seired ja loodusteaduslikud kogud. Seejuures võivad vaatlused olla kogutud läbi harrastusteaduslike projektide. Et need andmed oleksid lihtsasti kättesaadavad, koondatakse sarnase sisuga andmed andmebaasidesse (ingl *database*).

4.1. Makroökoloogiliste suurandmete tüübid

Iga tabeli kujul andmekogum sisaldab endas tunnuseid (näiteks lehe pikkus), mis sisaldavad väärtusi (näiteks lehe pikkus millimeetrites). Tunnuste liigitamine tüüpideks sõltub, millisest näitajast lähtutakse: väärtuse hulga omadused, andmete saamisviis või nende sisuline roll (Farley *et al*., 2018; Tiit & Tooding, 2019). Levinuim tunnuste liigitus läheneb statistikute vaatenurgast, mis põhineb väärtuste loogiliselt lubatavatel matemaatilistel tehetel, liigitades tunnused neljaks tüübiks: nominaal- (ingl *nominal*), järjestus- (ingl *ordinal*), vahemikskaala (ingl *interval*) ja suhteskaala tunnused (ingl *ratio*; Tiit & Tooding, 2019). Näitlikustatuna on nominaaltunnuseks õite värv, järjestustunnuseks linnupoegade arv pesakonnas, vahemikskaala tunnuseks temperatuur ning suhteskaala tunnuseks kaal ja pikkus. Siiski oleneb ka selle liigituse puhul paljustki lahendatava ülesande kontekstist (Tiit & Tooding, 2019), näiteks võib liigi ladinakeelne nimetus olla olukorrast sõltuvalt kas nominaal- (järjekord pole oluline) või järjestustunnus (tähestikuline järjekord). Lisaks saab andmebaasides või andmetabelites olevaid andmeid liigitada oma struktureerituse astmelt kaheks: struktureeritud (ingl *structured data*) ja struktureerimata (ingl *unstructured data*) andmeteks (Lee, 2017). Struktureeritud andmete näiteks on ridade ja veergudega tabel ning struktureerimata andmete näideteks pilt, video, heli ja tekst (Lee, 2017). Siiski viitab juba suurandmete olemus sellele, et need on enamasti struktureerimata (Fan *et al*., 2014).

Alljärgnevalt toon ühe võimaliku suurandmete tüüpide liigitamise viisi, mis võtab arvesse selle, kas need andmed on algsed – primaarsed andmed (ingl *primary data*) – või on need esmaste andmete pealt tuletatud – sekundaarsed andmed (ingl *secondary data*; Deakin University Library, 2023). Siiski ei saa väita, justkui oleks primaarseteks ja sekundaarseteks andmeteks liigitamine alati üheselt mõistetav (Deakin University Library, 2023). Primaarsed

andmed, mida võib pidada ka kõige üldisemateks globaalse elurikkuse suurandmeteks (Cornwell *et al*., 2019), on sellised andmed, mis käivad konkreetse vaatluse kohta kindlas ajas ja ruumis (Troia & McManamay, 2016). Primaarsete andmete põhjal saab analüüsida ja kirjeldada ökoloogilisi protsesse ja mustreid (Troia & McManamay, 2016), tekitades nõnda omakorda sekundaarseid andmeid. Siinkohal annan ette valiku primaarsetest andmetest, mis pakuvad makroökoloogidele huvi (Beck *et al*., 2012; Costello *et al*., 2013; Flantua *et al*., 2023; Mascarenhas *et al*., 2020; Michener & Jones, 2012; Schiller *et al*., 2021; Wüest *et al*., 2020):

- taksonoomilised andmed;
- fülogeneetilised andmed;
- liigi leiuandmed (geograafilised koordinaadid);
- liigi tunnused;
- DNA sekveneerimise andmed;
- biogeograafilised andmed;
- liigiline koosseis proovis, katsealal;
- keskkonnaandmed;
- liigi olemasolu dokumenteerimine pildi, video ja heliga;
- paleontoloogilised andmed.

Sageli kaasatakse makroökoloogilistesse uuringutesse keskkonnaandmeid (ilmaandmed, muld jne), mille varalt produtseeritakse omakorda sekundaarseid andmeid nagu erinevad kliima- ja mullakaardid (Franklin *et al*., 2017). Pildi, video ja heli kujul talletatud andmed pakuvad makroökoloogidele aina enam uusi võimalusi, nõudes sellega ka spetsiifilisemaid oskusi (Hoekendijk *et al*., 2021; Lee, 2017; Robeva *et al*., 2020; Schiller *et al*., 2021). Samuti pakuvad makroökoloogias palju võimalusi erinevad paleontoloogilised andmed, võimaldades heita pilk liikide evolutsioonile ja ökoloogiale (Beck *et al*., 2012; Flantua *et al*., 2023).

Sekundaarsed andmed on andmed, mis on tuletatud juba olemasolevate primaarsete andmete analüüsist (Andrews *et al*., 2012; Deakin University Library, 2023). Sekundaarsed andmed on enamasti primaarsete andmete pealt tuletatud erinevat sorti kaardid, modelleeringud, rekonstruktsioonid, analüüsid ja hinnangud liigi ning populatsiooni eri aspektidele (levik, käitumine, dünaamika, ohustatus jms). Seejuures võivad need kasutatavad primaarsed andmed olla kogutud kellegi teise poolt mõnel teisel eesmärgil (Johnston, 2014). Sekundaarsete andmetega on võimalik uurida uusi teadusküsimusi, vaadelda andmeid uue

nurga alt, üldistada seaduspärasusi ja kontrollida olemasolevaid seisukohti (Andrews *et al*., 2012). Võimalikud makroökoloogidele huvipakkuvad sekundaarsed andmed on (Beck *et al*., 2012; Costello *et al*., 2013; Hampton *et al*., 2013; Vilela & Villalobos, 2015):

- kliimakaardid;
- mullakaardid;
- liikide leviku kirjeldused (sh kaardid);
- elurikkuse kaardid;
- populatsioonide dünaamika mudelid;
- liikide punased nimestikud;
- fülogeneesipuu rekonstruktsioonid.

4.2. Makroökoloogiliste suurandmete allikad

Lisaks suurandmete sisule ja vormile võib suurandmeid liigitada ka nende päritolu põhjal (Farley *et al*., 2018). Makroökoloogias on palju erinevaid suurandmete süsteeme, mis panustavad reaalajas suurandmete pidevale tekkele läbi automatiseeritud kogumiste näiteks vaatluste ja seirete näol (Beck *et al*., 2012; Farley *et al*., 2018; Hampton *et al*., 2013; Michener & Jones, 2012). Seejuures võivad andmed olla kogutud harrastusteaduslike projektide käigus (Hampton *et al*., 2013). Lisaks on väga olulised ka muuseumides, raamatukogudes, erakogudes jms olevad loodusteaduslikud kogud, mille digiteerimine (ingl *digitisation*) panustab suurandmete mahu kasvamisse ja tagab seni digitaalsena kättesaamatutena olnud andmete ligipääsetavuse (Drew *et al*., 2017; Moritz & Agudo, 2013). Suurandmetesse annavad oma panuse ka juba varasemalt lühi- või pikaajaliste uuringute käigus kogutud ja avalikustatud andmed (Farley *et al*., 2018). Kõiki neid andmeid on vaja koondada ja loogiliselt hoiustada – selle tarbeks luuakse erisuguseid andmebaase (Hortal *et al*., 2015).

4.2.1. Vaatlused ja seired

Aina enam hakkavad makroökoloogias kanda kinnitama automatiseeritud vaatlused ja seired, võimaldades seda teha aina väiksemate kuludega (Hampton *et al*., 2013; Michener & Jones, 2012; Zipkin *et al*., 2021). Automatiseeritud andmete kogumisega on võimalik tõsta suurandmete kogumise kiirust, olles aina enam reaalajas kättesaadavad (Farley *et al*., 2018). Makroökoloogilistes uuringutes on hakatud kasutama eri sorti sensoreid – kaamerad, satelliit- ja raadiotelemeetrid ning erinevad mõõturid (Zipkin *et al*., 2021) –, mis võimaldavad koguda aastas petabaitide (1 PB ≈ 1 000 TB) viisi andmeid (Michener & Jones, 2012).

Siiski ei piisa ainuüksi seadmete olemasolust – vaatlus- ja seireandmete kogumisele annavad kuju erinevad organiseeritud vaatlussüsteemid ja seirejaamad, millest FLUXNET, International Long Term Ecological Research networks (LTER), National Ecological Observatory Network (NEON) ja Terrestrial Ecosystem Research Network (TERN) on ühed tuntumad (Baldocchi *et al*., 2001; Michener & Jones, 2012). Kaugseires (ingl *remote sensing*) on satelliitide, lennukite ja droonidega võimalik koguda erisuguseid kliimaandmeid ja aerofotosid (Farley *et al*., 2018). Seiramine on avanud uksed juba mitmes ökoloogia vallas, neist muuhulgas loomade elupaikade seires ning taimestiku ja biomassi produktsiooni seires (Farley *et al*., 2018; Wüest *et al*., 2020).

4.2.2. Harrastusteadus

Harrastusteadus kujutab endas vabatahtlike, enamasti mitte-teadlaste, panust teadusesse vaatluste ja andmete sisestamise näol (Dimson *et al*., 2023). Aina enam nähakse potentsiaali kasutada harrastusteaduse käigus kogutud andmeid teadustöödes, kuna harrastusteadlased on võimelised koguma andmeid intensiivsemalt, laiemal ruumi- ja ajaskaalal ning kuluefektiivsemalt kui oma ala eksperdid teaduslikel välitöödel (Dimson *et al*., 2023; Hampton *et al*., 2013; Hochachka *et al*., 2012; Zipkin *et al*., 2021).

Harrastusteaduslike andmete kooskasutamine näiteks satelliitandmetega annab laiema pildi toimuvatest ökoloogilistest protsessidest (Dimson *et al*., 2023; Hampton *et al*., 2013). Nõnda on näiteks võimalik kaardistada taimede funktsionaalseid tunnuseid (Wolf *et al*., 2022), invasiivseid taimeliike (Dimson *et al*., 2023) ja seirata üldisi ökoloogilisi muutusi (Mahecha *et al*., 2021). Näiteid harrastusteaduslikest projektidest, mille raames koguti andmeid sihipäraselt teaduse heaks, leidub ka Eestist. Tartu Ülikooli ja Eestimaa Looduse Fondi poolt korraldatud projekti “Eesti otsib nurmenukke” käigus koguti nurmenukkude morfoloogilisi

andmeid, et analüüsida Eesti nurmenuku populatsiooni õietüüpide (nn L- ja S-tüüp) esinemist eri maastikel (Aavik *et al*., 2020). Samuti leiab mitmeid näiteid suureskaalalistest harrastusteaduslikest võrgustikest nagu iNaturalist ja eBird (Augustine *et al*., 2024) ning ka kodumaise eElurikkuse koos PlutoF-i liidesega, kajastades infot Eesti elurikkuse kohta (Abarenkov *et al*., 2010).

Mõistagi võivad harrastusteadlaste poolt kogutud andmetel olla omad puudused, mida tuleb nende andmete analüüsil ja saadud tulemuste tõlgendamisel arvesse võtta (Hampton *et al*., 2013; Johnston *et al*., 2023). Näiteks tuleb harrastusteaduslike andmete analüüsil arvestada, et andmed on enamasti kogutud tihedama inimasustusega piirkondadest või on vaadeldud vaid huvipakkuvaid ja/või sagedamini esinevaid liike (Daru *et al*., 2018; Dimson *et al*., 2023; Johnston *et al*., 2023). Samuti on harrastusteaduslike andmete kvaliteet kõikuv, kuna andmekogujate määramise oskused ja kogemused on erinevad, mis võivad endaga kaasa tuua valemäärangud (Johnston *et al*., 2023).

4.2.3. Loodusteaduslikud kogud

Aegade jooksul on loodusteadlased kogunud ekspeditsioonidelt erinevaid proove ja organisme, kirjeldanud ja joonistanud üles uusi liike, täheldanud avastused reisipäevikutesse ja koondanud uued teadmised kokku raamatutesse, teinud herbaariume ja kogusid. Kõik need loodusteaduslikud kogud on nüüd leidnud oma koha (loodus)muuseumides, raamatukogudes, eraisikute ja ülikoolide kogudes (Edwards *et al*., 2000). Sellised füüsilised kogud annavad hea võimaluse kiigata liikide leviku minevikku ja võrrelda seda olevikuga, uurida liikide vastupanu kliimamuutustele ja sekveneerida eksemplaridelt DNA-d (Lang *et al*., 2019; Moritz & Agudo, 2013; Zipkin *et al*., 2021).

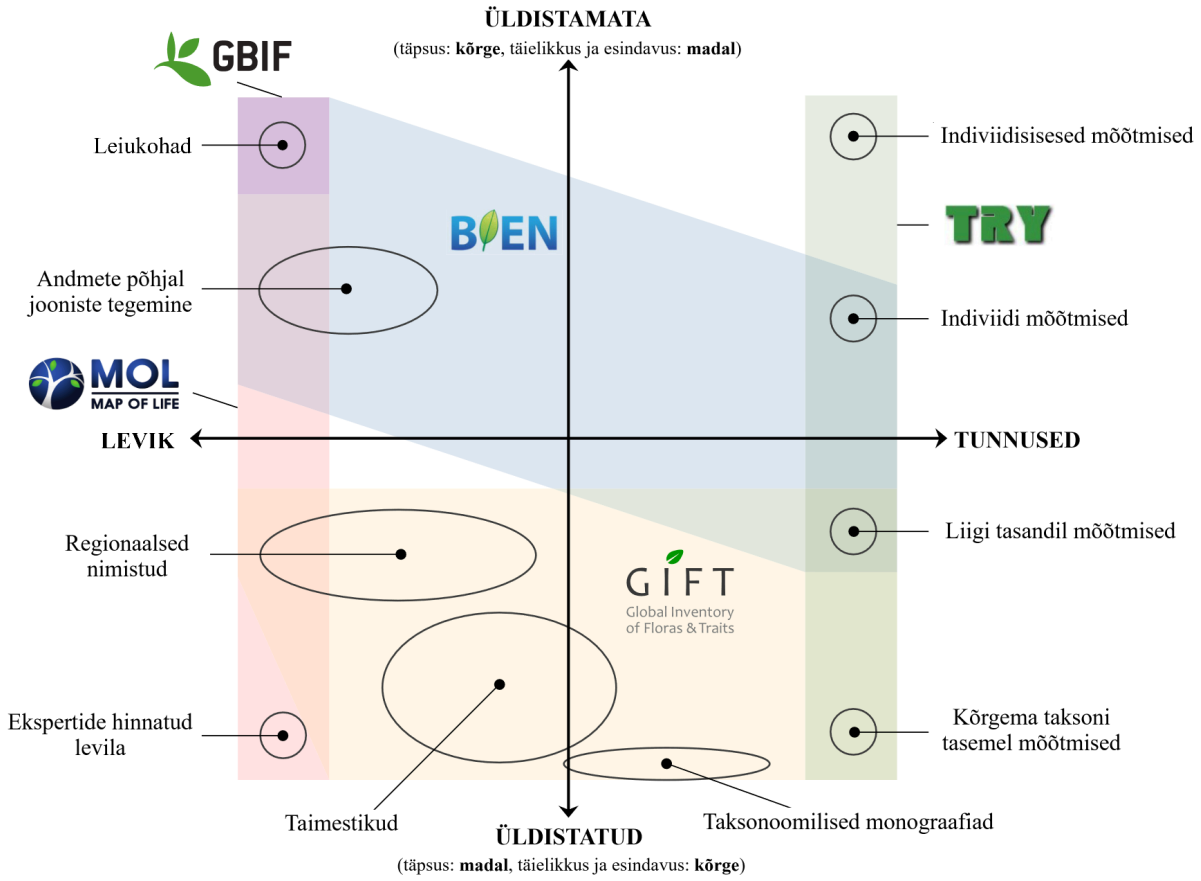
Selleks, et loodusteaduslikud kogud oleksid laiemalt kättesaadavad, on hakatud neid digiteerima (Ball-Damerow *et al*., 2019; Drew *et al*., 2017; Lang *et al*., 2019). Digiteerimine on ajakulukas ja kallis ettevõtmine, kuna digiteerimist vajavaid eksemplare on üle maailma kokku sadu miljoneid (Lang *et al*., 2019). Digiteerimise hõlbustamiseks arendatakse erinevaid digiteerimise lahendusi objektide tuvastamiseks (Lang *et al*., 2019). Kogude digiteerimine panustab teadmiste lünkade täitmisesse. Näiteks on olnud juhtumeid, kus on teada kõigest liigi nimetus, ent tänu kogude digiteerimisele on leitud liigile juurde lisainformatsiooni nagu leiukoht ja määramistunnused (Cornwell *et al*., 2019).

Loodusteaduslikud kogud annavad suure panuse suurandmetesse, ent siiski tuleb neist saadavatesse andmetesse suhtuda teatava ettevaatlikkusega. Loodusteaduslike kogude tõepära ja mitmelaadilisuse määrab ära see, millised taksonoomilised ja ajalis-ruumilised määrangud on need kogujatelt saanud, kes ei pruugi sageli olla enam elavate hulgas (Daru *et al*., 2018; Farley *et al*., 2018; Lang *et al*., 2019; Yesson *et al*., 2007). Samuti tuleb arvesse võtta teatavat andmete kallutatust (Prendergast *et al*., 1993). See võib olla tingitud nii kogujate endi eelistustest ja kogumise võimalustest (Daru *et al*., 2018) kui ka vaatluste ajaliselt lühikesest vaatest minevikku (Lang *et al*., 2019). Digiteerimises endas võib ka omakorda esineda kallutatusi, näiteks võivad mõne piirkonna kogud olla digitaalsel kujul alaesindatud (Nelson & Ellis, 2018). Viimaks on palju arenguruumi ka digiteerimise kvaliteedis (Nelson & Ellis, 2018).

4.2.4. Andmebaasid

Selleks, et andmetel oleks mingi väärtus, peavad need olema leitavad ja kasutatavad, mistõttu peavad need olema korrastatult andmebaasides. Suurandmete laialdasem kasutus on loonud viljaka pinnase erinevate andmebaaside loomiseks (Farley *et al*., 2018; Wüest *et al*., 2020). Bioloogiliste andmebaaside ülesandeks on kajastada elurikkuse andmeid ja korrastada andmeid nõnda, et neid oleks teadusküsimustele vastamiseks võimalik ka teiste andmetega koos kasutada (Cornwell *et al*., 2019). Selline makroökoloogiliste suurandmete koondamine ühte kohta aitab paremini mõista toimuvaid ökoloogilisi protsesse ja mustreid (Farley *et al*., 2018). Lisaks pakuvad andmebaasides olevad andmed alternatiivi kulukatele proovide kogumisele ja selle tarbeks vajaminevatele reisimistele (Harris *et al*., 2023).

Bioloogilisi andmebaase saab tinglikult jagada kaheks: kindla temaatikaga ja laiapõhjalised andmebaasid (Beck *et al*., 2012; König *et al*., 2019). Viimased aitavad vähendada eri andmebaasidest andmete otsimisele ja haldamisele kuluvat aega (Beck *et al*., 2012). Sõltumata andmebaasi tüübist, saab selles olevaid andmeid jaotada mitmemõõtmeliselt. Näiteks on näidatud taimede makroökoloogias olulisemates taimeandmebaasides olevate andmete jaotust kahes olulises aspektis: 1) üldistatuse tase (ingl *aggregation*) ja 2) levik ning funktsionaalsed tunnused (Joonis 4). Seejuures ei ole andmebaasis olevad andmed läbinisti vaid ühe põhimõtte järgi, vaid need pigem varieeruvad mitmemõõtmeliselt. Nõnda erinevad näiteks TRY andmebaasi andmed üldistatuse tasemelt indiviidisisestest kuni kõrgemate taksonite tasemel üldistusteni välja.



Joonis 4. *Valitud taimeandmebaaside liigitus vastavalt andmete 1) üldistatuse tasemele ja 2) geograafilisele paigutusele ruumis ning tunnustele. Võib öelda, et andmebaasid pole ühe põhimõtte järgi rangelt piiritletud, vaid varieeruvad mitmemõõtmeliselt. BIEN, Botanical Information Network and Ecology Network; GBIF, Global Biodiversity Information Facility; GIFT, Global Inventory of Floras and Traits; MOP, Map of Life; TRY. Joonis on kohandatud König* et al. *(2019) järgi.*

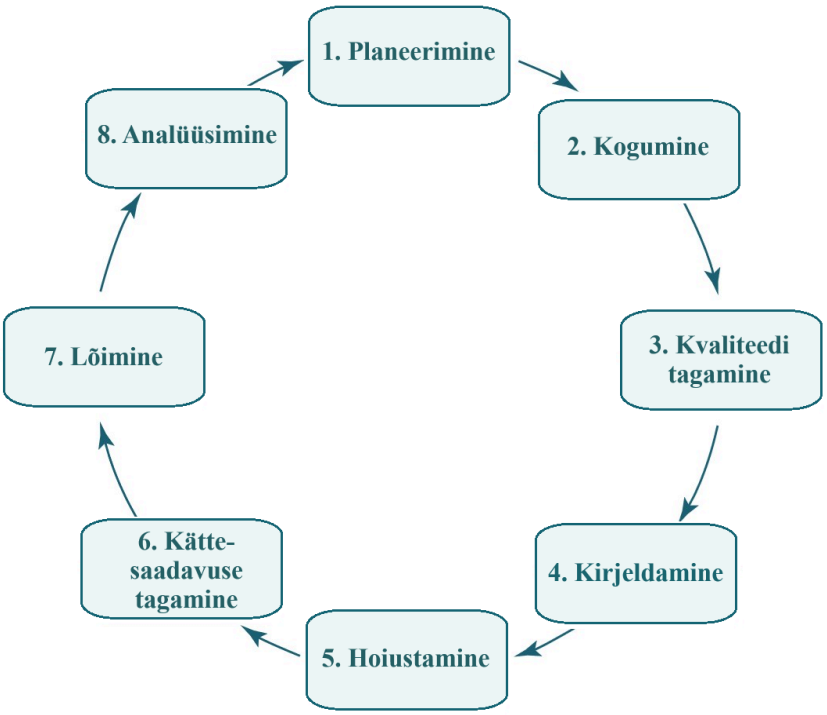
Üle kõikide andmebaaside tasub eraldi välja tuua GenBank-i, GBIF-i ja TYR-i kui ühed hetkel kõige terviklikumad andmebaasid (König *et al*., 2019). Samuti mainivad väärimist Eestis arendatavad UNITE andmebaas, mis keskendub eukarüootide tuuma ribosoomide ITS-regioonile (Abarenkov *et al*., 2024), ja MaarjAM andmebaas, mis hõlmab endas spetsiifiliselt arbuskulaarse mükoriisa DNA-järjestuste infot (Öpik *et al*., 2010). Lisa 2 pakub kokkuvõtlikku ülevaadet makroökoloogilistest andmebaasidest.

5. MAKROÖKOLOOGILISTE SUURANDMETE PROBLEEMID JA LAHENDUSED

Vaatamata suurandmete laialdastele kasutusvõimalustele ja potentsiaalile pakkuda vastuseid keerulistele ökoloogilistele küsimustele, lasuvad nende ümber mitmed spetsiifilised kui ka üldised metodoloogilised, tehnilised ja töökultuurilised probleemid (Fan *et al*., 2014). Sageli on need põhjustatud suurandmete endi olemusest (Farley *et al*., 2018): andmeid on palju, andmed on mitmelaadilised, andmeid tekib kiiresti, andmed pole alati täielikult tõesed ja kõik andmed pole antud kontekstis ka väärtuslikud. Lisaks makroökoloogiast leitavatele lahendustele on võimalik võtta lahendusi üle teistest külgnevatest valdkondadest, sest tihti peitub heade lahenduste taga võimalus rakendada neid universaalselt (Farley *et al*., 2018). Selle tarbeks on sageli vaja teha oma või teiste valdkondadega koostööd. On ka mitmeid teisi aspekte, nagu suurandmete privaatsus ja turvalisus, mis väärivad lähitulevikus makroökoloogia kontkestis põhjalikumat käsitlemist, kuna universaalseid meetodeid pole neil teemadel veel välja kujunenud (Lee, 2017). Selleks, et oleks lihtsam mõista, millised on võimalikud ettetulevad probleemid makroökoloogiliste suurandmetega töötamisel, esitlen probleeme vastavalt Micheneri ja Jonesi (2012) välja toodud andmete eluringi (ingl *data life cycle*) skeemile, kus saavad hõlmatud nii mõnedki andmehaldusega (ingl *data management*) seotud aspektid.

5.1. Andmete eluring ja planeerimine

Varem mainitud suurandmete omadused panevad proovile traditsioonilised andmehalduse ja -analüüsi lähenemised (Farley *et al*., 2018), sestap on ka makroökoloogias suur nõudlus selgete andmehalduse tavade järele (Augustine *et al*., 2024). Hea uuringu aluseks on andmehalduse plaan, nõudes läbimõeldud põhimõtteid alates andmete kogumisest kuni nende analüüsini välja (Farley *et al*., 2018; Hampton *et al*., 2013; Wüest *et al*., 2020). Andmete eluringis saab eristada kokku kaheksa etappi: planeerimine, kogumine, kvaliteedi tagamine, kirjeldamine, hoiustamine, kättesaadavuse tagamine, lõimimine ja analüüsimine (Joonis 5). Lõpuks töötab see eluring siiski pigem spiraalina – iga tsükli lõpp on uue alguseks. Igas etapis kannavad andmed erinevaid rolle ja nende juurde käivad erinevad tööprotsessid.



Joonis 5. *Andmete eluring läbi kaheksa etapi. Joonis on kohandatud Micheneri ja Jonesi (2012) järgi.*

Andmetega töötamise juures on kõige aluseks hea uuringu plaan (Farley *et al*., 2018). Hea uuringu plaan aitab hoida kokku uuringule kuluvat aega ja ressursse, parandada uuringu kvaliteeti ja efektiivsust ning võimaldab katsete ja analüüside korratavust, tagades nõnda andmete usaldusväärsuse (Michener & Jones, 2012; Moore & McCarthy, 2016). Korraliku andmehalduse korral on võimalik pidada järge iga andmete eluringi etapi ja seal ettetulevate probleemidega (Reichman *et al*., 2011), mille jälgimist lihtsustavad erinevad rakendused, tarkvarad, algoritmid, tehnoloogilised lahendused jms abivahendid (Michener & Jones, 2012). Erinevate abivahendite kasutamisel tasub tähele panna, et nende populaarsus on ajas muutuv (Jones *et al*., 2006).

Suurandmete kasvava tähtsuse valguses on oluline, et nende käitlemisel järgitaks kindlaid standardeid, mille üheks näiteks on Darwin Core (Augustine *et al*., 2024; Jones *et al*., 2006). Standard hõlmab endas infot andmete ja metoodika kohta, tagades sätestatud miinimumnõuete kinnipidamist (Plesser, 2018; Zurell *et al*., 2020). Hea standard võimaldab uuringute korratavust ja süstemaatilist lähenemist uuringutele (Plesser, 2018; Zurell *et al*., 2020). Alati ei vasta erinevad makroökoloogilised andmestikud olemasolevatele standarditele (Parsons *et al*., 2011) või on need loodud spetsiifiliselt konkreetse uuringu tarbeks ja pole ülekantavad sama valdkonna uuringutele (Jones *et al*., 2006; Wüest *et al*., 2020).

Standardite loomine on kahtlemata vajalik samm, et saada aina kvaliteetsemaid taaskasutatavaid andmeid, ent samal ajal on see kahe teraga mõõk. Liiga palju erinevaid standardeid muudab pildi väga kirjuks, mille tõttu on standardite omavaheline lõimimine ja samaaegne järgimine keeruline (Jones *et al*., 2006). Samuti on standardite endi detailsus varieeruv (Jones *et al*., 2006). Võidakse piirduda vaid etteantud tunnustega ja nõnda võib jääda tähelepanuta mõnele liigile iseloomulik tunnus, mida polnud standardis ette nähtud (Harris *et al*., 2023; Hortal *et al*., 2015). Ideaalis võiks uute standardite loomise asemel olemasolevaid arendada ja täiustada, ent lõpuks jääb teatav mitmelaadilisus paratamatult kõikidest andmetest kumama (Farley *et al*., 2018). Selleks tuleb leida tasakaal eksisteerivate andmete standardimisel ja paremate vaatlusmeetodite loomise vahel (Farley *et al*., 2018). Seega tuleb tõdeda, et ühe ja ainsa standardi loomine on sisuliselt võimatu missioon, kuna makroökoloogilised andmed ise on juba väga mitmelaadilised (Higgins *et al*., 2002).

5.2. Andmete kogumine

Suurandmete kvaliteeti saab mõjutada juba nende kogumisest alates, kus kogutud andmete kvaliteedi määrab nende kogus ja olemus (Michener & Jones, 2012). Sageli määrab kogutud andmete kvaliteedi ka kasutatava standardi detailsus (Harris *et al*., 2023). Samuti on andmete kogumisel oluline märkida üles võimalikud mõõtevea piirid, et sellega osatakse andmete käitlemise edasistes etappides arvestada (Zipkin *et al*., 2021). Samas pole ka liialt paljude tunnuste kogumine alati hea, kuna paljude tunnuste kogumine võib viia andmete tahtmatu korreleerumiseni müraga (Fan *et al*., 2014). Muidugi ei saa üle ega ümber ka andmete sisestamisel tekkivatest vigadest (König *et al*., 2019).

Tagamaks võimalikult hea andmete kvaliteedi juba nende kogumisel, tasub enne andmete kogumist teha selgeks, milliste piirkondade, tunnuste, liigirühmade jms kohta on puudulikud või kallutatud andmed (König *et al*., 2019; Meyer *et al*., 2016). Nõnda on võimalik täita teadmistelünki ja osata sättida paremini uuringu fookust (Meyer *et al*., 2016). Samuti lihtsustavad andmete kogumist, sisestamist ja sisestatu kontrollimist erinevad tarkvarad ja nende paketid (Boyle *et al*., 2013; Cayuela *et al*., 2012; König *et al*., 2019). Näiteks on erinevate tarkvarapakettidega võimalik kontrollida sisestatud taimeliikide nimetuste õigekirja ja sünonüümide kasutust (Boyle *et al*., 2013; Cayuela *et al*., 2012). Sisestatu kontrollimiseks sobib ka eri isikute poolt läbiviidav manuaalne, ent see-eest süsteemne kontroll (König *et al*., 2019; Michener & Jones, 2012). Üle ega ümber ei saa standardite järgimisest – on oluline, et neid rakendatakse juba alates andmete kogumise etapist (Wüest *et al*., 2020).

5.3. Andmete kvaliteedi tagamine

Andmete kvaliteedi tagamine on makroökoloogilistest uuringutes muutumas aina olulisemaks (Augustine *et al*., 2024). Cybernetica AS (2023a) kohaselt "näitab andmete kvaliteet, mil määral andmekarakteristikud rahuldavad teadaolevaid või eeldatavaid vajadusi ettemääratud tingimustes kasutamisel". Kasutatavate andmete kvaliteet pole sageli teada ja tihti polegi need uuringus kasutamiseks piisavalt kvaliteetsed (Beck *et al*., 2012). Selline seis viib „rämps sisse, rämps välja" olukorrani – kui sisendiks olevad andmed ei ole kvaliteetsed, siis ei ole seda ka nende andmete pealt tehtavad järeldused (Hortal *et al*., 2015). Kvaliteedi kõikumine teeb mudelid ebatäpseks ja võib viia eksitavate hinnanguteni (Dormann *et al*., 2008).

Andmete kvaliteet on laiaulatusliku mõjuga mitmemõõtmeline kontseptsioon (Michener & Jones, 2012; Pipino *et al*., 2002). Kvaliteedi hindamine on ühe andmestiku piires keeruline, sest kvaliteet sõltub konkreetse uuringu kontekstist (Cornwell *et al*., 2019). Võib öelda, et kvaliteedi hindamine on katkematu protsess ja seda ei saa kunagi lõppenuks lugeda (Cornwell *et al*., 2019). Andmete kvaliteeti on võimalik tõsta nende oskusliku puhastamisega näiteks erinevate algoritmide ja tarkvaradega, kus jäetakse kõrvale eeldatavad vigased ja kallutatud kirjed (Augustine *et al*., 2024; Michener & Jones, 2012).

Makroökoloogias mõjutavad enim andmete kvaliteeti puuduvad andmed (Cornwell *et al*., 2019), mis teevad elurikkuse ja selles toimuvate protsesside hindamise keeruliseks (Beck *et al*., 2012). Andmestikust puuduvad andmed võivad viia andmestikus olevate andmete nihkumiseni (Cornwell *et al*., 2019). Andmete kvaliteeti vähendavad ka vigased andmed, mis võivad viia väärade arusaamadeni (Cornwell *et al*., 2019). Andmete kvaliteeti mõjutavad ka duplikaadid, mis tähendab, et sama vaatlust tekib andmebaaside omavahelise info vahetamise käigus mitu korda (Harris *et al*., 2023). Universaalne lahendus on puuduvaid ja vigaseid andmeid imputeerida (ingl *data imputation*), mis tähendab, et puuduv või vigane väärtus asendatakse andmestiku põhjal hinnatud väärtusega (König *et al*., 2019; Tiit & Tooding, 2019).

Sõltuvalt makroökoloogilistest suurandmetest (leiukoht, liiginimetus, liigitunnused jne) on kirjeldatud nendega seonduvad vajakud (ingl *shortfall*; Beck *et al*., 2012; Hortal *et al*., 2015; Wüest *et al*., 2020). Uurituim neist on Wallace'i vajak (ingl *Wallacean shortfall*), mis on oma nime saanud ühe biogeograafia alusepanija Alfred Russel Wallace'i (1823 – 1913) järgi (Camerini, 2024; Hortal *et al*., 2015). Selle vajaku puhul pole täielikku ülevaadet liikide geograafilise leviku kohta (Hortal *et al*., 2008; Hortal *et al*., 2015; Lobo *et al*., 2007), mistõttu

on keeruline jälgida ajas toimuvaid liikide leviku muutusi (Huisman & Millar, 2013) ja väljasuremismustreid (Lobo, 2001). Võrdlemisi uuritud on ka Linné' vajak (ingl *Linnean shortfall*), mis on oma nime saanud süstemaatika teerajaja Carl von Linné (1707 – 1778) järgi (Hortal *et al*., 2015; Müller-Wille, 2024). Selle puudujäägi kohaselt pole omavahel kooskõlas kirjeldatud liikide arv ja tegelikult eksisteerivate liikide arv organismi- või ka taksonirühmiti (Hortal *et al*., 2015).

Vähemuuritud, kuid aina olulisemaks muutuvaid vajakuid on veelgi. Näiteks on evolutsiooniteooria alusepanija Charles Darwini (1809 – 1882) järgi nime saanud Darwini vajak (ingl *Darwinian shortfall*), mille puhul on lünklikud teadmised liikide suguluse ja tunnuste evolutsiooni kohta (Desmond, 2024; Diniz-Filho *et al*., 2013; Hortal *et al*., 2015). Teiseks kasvava tähtsusega vajakuks on funktsionaalse ökoloogia alguse juures olnud Christen C. Raunkiæri (1860 – 1938) järgi nimetatud Raunkiæri vajak (ingl *Raunkiæran shortfall*), kus on puudulikud teadmised liikide tunnuste ja nende funktsioonide kohta (Gonçalves-Souza *et al*., 2023; Hortal *et al*., 2015; JSTOR Global Plants, 2013).

5.4. Metaandmete kirjeldamine

Metaandmed (ingl *metadata*) on täpne ja struktuurne kirjeldus olemasolevatest andmetest – teisisõnu andmed andmetest (Jones *et al*., 2006). Metaandmed on vajalikud, et mõista kasutatava andmestiku ülesehitusest ja andmete kogumist (Alves *et al*., 2018; Jones *et al*., 2006). Metaandmed peavad praegusel suurandmete ajastutel olema loetavad lisaks inimestele ka arvutitele (Jones *et al*., 2006). Sellele lisaks peavad metaandmed olema koostatud kindla standardi järgi ja olema kättesaadavad (Hampton *et al*., 2013; Harris *et al*., 2023; Jones *et al*., 2006).

Nagu ikka, leiab lahenduse standardite rakendamises – ikka selleks, et tagada metaandmete ühtne vorm ja hõlbus kasutamine (Michener & Jones, 2012). Leidub nii universaalseid, näiteks Dublin Core, kui ka spetsiifiliselt (makro)ökoloogia jaoks kohandatud vorme, näiteks Ecological Metadata Language (EML) ja Darwin Core (Jones *et al*., 2006). Samuti on metaandmete haldamiseks võimalik kasutada tarkvarasid, kus saab neid samu standardeid mugavalt rakendada (Hampton *et al*., 2013). Siiski ei saa maha laita standardite rohkust, kuna makroökoloogilised suurandmed on väga mitmekesist päritolu, sestap ongi keeruline luua üht ja ainsat standardit metaandmete kirjeldamiseks (Higgins *et al*., 2002). Vaatamata standardite kasutamisele, võib metaandmete kvaliteet olla siiski kõikuv, kuna ei järgita täielikult kehtestatud standardit (Harris *et al*., 2023).

5.5. Andmete hoiustamine

Suurandmete hulk kasvab ajas kiirenevalt, mis nõuab nende oskuslikku haldamist ja kättesaadavaks tegemist (Augustine *et al*., 2024). Kogutud andmeid tuleb organiseerida ja hoiustada, et neile oleks nii praegu kui ka edaspidi võimalik ligi pääseda. Sageli saab takistuseks juba palju mainitud suurandmete maht ja kogumise kiirus (Farley *et al*., 2018), mistõttu on vaja paremat arvutusvõimsust. Selleks, et uuringusiseselt oleks kasutatavaid andmeid lihtne jagada, tasub panustada paremate andmete, analüüside ja standardite hoiustamise ning haldamise infrastruktuuridesse, näiteks ühiste andmehoidlate loomisesse (Farley *et al*., 2018; Hampton *et al*., 2013; Reichman *et al*., 2011). Ka on võimalik kasutada pilvtöötluse (ingl *cloud computing*) lahendusi (Chen *et al*., 2014), mis lubavad ühendada omavahel asutuste arvutusvõimsused ja pakkuda paindlikumat ligipääsu (Farley *et al*., 2018). Uuringud on näidanud, et andmebaasidega seonduvad probleemid kerkivad sageli esile andmete ebakorrektsetest jagamise ja kureerimise praktikatest (Augustine *et al*., 2024). Selleks, et need andmebaasid oleksid ka edaspidi kõrge kvaliteediga, tasub panustada aega ja raha palkamaks spetsialiste, kelle ülesandeks on teostada korrektset andmehaldust (Augustine *et al*., 2024).

Uuringute ja andmete hoiustamiseks on neile vaja külge objekti digitaalidentifikataorit ehk DOI-d (ingl *Digital Object Identifier*; Farley *et al*., 2018). Cybernetica AS (2023d) järgi on DOI "standardne märgistring füüsiliste, digitaalsete või abstraktsete objektide identifitseerimiseks ja püsivaks eristuseks", mis ühtse identifitseerimise standardse taristuna moodustab DOI-süsteemi (Cybernetica AS, 2023b). Erinevalt URL-ist on DOI standardne hüperlink (Chandrakar, 2006). DOI-koodi saamiseks tuleb eelnevalt täita kindlad tingimused, näiteks on identifikaatorile vaja lisada metaandmed (DOI Foundation, 2023). DOI kasutamisel on mitmeid eeliseid, muuhulgas on see ülemaailmselt aktsepteeritav ja püsiv (Chandrakar, 2006).

5.6. Andmete kättesaadavuse tagamine

Praegu, kus andmeid tekib meeletu kiirusega, tasub lisaks uute andmete kogumisele mõelda ka sellele, kuidas kasutada maksimaalselt ära olemasolevaid andmeid (Hampton *et al*., 2013; Harris *et al*., 2023). Andmete ja ka analüüsiskriptide kättesaadavusel on oluline aspekt, kas need on kellegi personaalarvutis või on leitavad avalikust andmebaasist (Beck *et al*., 2012; Michener & Jones, 2012; Reichman *et al*., 2011). Lisaks on oluline, kas andmed avalikustatakse kohe peale nende kogumist või alles peale teadusartikli avaldamist. Kui andmete kogumise ja avalikustamise vahele tekib mõneaastane viivis, siis pole neid andmeid võimalik teiste spetsialistide poolt reaalajas uuesti kasutada, mis võib viia uute teadmiste kujunemise aeglustumiseni (Hampton *et al*., 2013; Harris *et al*., 2023).

Mõni valdkond on pika andmete jagamise praktikaga, kuna neil on suured jagatud infrastruktuurid (astronoomia) ja/või nende uuritavad andmed on pigem ühetaolised (geneetika ja füüsika; Marx, 2013; Reichman *et al*., 2011). Eriti tasub esile tuua geneetikat, kus aastakümneid väldanud geenijärjestuste jagamise tava on juba iseenesest mõistetav (Reichman *et al*., 2011). Andmete kättesaadavuse tagamine on toonud kaasa andmemahtude kasvu ja sillutanud makroökoloogide tee suurandmeteni (Augustine *et al*., 2024). Oluline on muuta suhtumist – andmete ja analüüsiskriptide avalikustamine peaks olema iseenesest mõistetav, mitte tüütu kohustus (Hampton *et al*., 2013). Aina avalikuks muutuvas maailmas võtab andmete nõuded kokku FAIR-andmete printsiip, mille juures on oluline andmete leitavus (ingl ***f****indability*), kättesaadavus (ingl ***a****ccessibility*), koostalitlusvõime (ingl ***i****nteroperability*) ja korduvkasutus (ingl ***r****euse*; Wilkinson *et al*., 2016).

Leevendamaks andmete avalikustamata jätmist, nõuavad teadusajakirjad koos teadusartikliga ka kogutud, analüüsitud ja kasutatud andmete ning skriptide avalikustamist koos DOI-koodiga (Farley *et al*., 2018; Reichman *et al*., 2011). Sellised teadusartiklid saavad tihti ka rohkem viitamisi kui need, kes mainitud praktikat ei rakenda (Hampton *et al*., 2013; Harris *et al*., 2023). Seejuures võib öelda, et suhtumine andmete ja analüüsiskriptide jagamisse on aina paranenud ja muutunud tavapäraseks (Augustine *et al*., 2024). Koodihaldust koos üksikasjalike protsesside dokumenteerimisega (detailid andmete, analüüsi ja tulemuste kohta) saab teostada erinevates tarkvarades (Farley *et al*., 2018; Pichler & Hartig, 2023; Reichman *et al*., 2011).

5.7. Andmete lõimimine

Makroökoloogilistes uuringutes on keeruline piirduda vaid ühest allikast pärinevate andmetega, kui on seda võrdlemisi ühetaoliste andmete ja/või ajaliselt-ruumilisest piiritletud analüüsidega valdkondades nagu geneetika ja füüsika (Marx, 2013; Reichman *et al*., 2011; Zipkin *et al*., 2021). Erinevate andmete koos uurimist leevendavad andmete lõimimise tehnikad, mille käigus tuuakse ühte analüüsiraamistikku kokku erinevatest allikatest pärinevad andmed (Zipkin *et al*., 2021).

Andmete lõimimine võib siiski osutuda parajaks proovikiviks, mille kurja juureks võib pidada andmete mitmelaadilisust (Farley *et al*., 2018; Zipkin *et al*., 2021). Lisaks on makroökoloogias hakanud tekkima ebakõlad erinevate ruumi- või ajaskaaladega andmete kombineerimisel, kus andmeid on sageli ühtlustamise nimel kas suurendatud või vähendatud (Nguyen *et al*., 2014; Zipkin *et al*., 2021). Võib esineda ka probleeme asjaoluga, et lõimitavad andmed pole tasakaalustatud (ingl *unbalanced data*), mis viitab sellele, et erinevatest allikatest pärit andmepunktide hulk on erinev (Zipkin *et al*., 2021). Levinud subjektiivne võte selle lahendamiseks on võtta olemasolevast andmestikust omakorda alamvalim või vähendada üldist andmete mahtu, võides viia siiski kallutatud mudelite ja järeldusteni (Zipkin *et al*., 2021).

Lõimimise probleemide seljatamiseks on tehtud mitmeid ambitsioonikaid ponnistusi universaalsete meetodite arendamiseks (Fer *et al*., 2018; Johnston *et al*., 2018; Pacifici *et al*., 2019), kuid hetkel tuleb neile probleemidele läheneda pigem uuringu kontekstist lähtuvalt (Zipkin *et al*., 2021). Andmete lõimimine pole lihtsate killast, ent laiendab makroökoloogiliste uuringute võimalusi (Pearse *et al*., 2018), parendades uuringute ning järelduste ajalis-ruumilist ulatust ja parameetrite hinnangute täpsust (Zipkin *et al*., 2021).

5.8. Andmete analüüsimine

Analüüs on protsess, mille käigus omandavad andmed oma tõelise väärtuse – analüüsi käigus saadav informatsioon on see, mis rajab tee uute teadmisteni (Beck *et al*., 2012; Wüest *et al*., 2020). Aina enam muutub suurandmetega töötamisel oluliseks adekvaatne statistiline analüüs, kus on oluline arendada pidevalt uusi analüüsimeetodeid, et teha üha täpsemaid prognoose ja paremaid ülevaateid tunnuste vahelistest seostest (Fan *et al*., 2014). Ka makroökoloogias esineb palju keerulisi ja mittelineaarseid süsteeme, mille modelleerimiseks on tarvis võimekamaid analüüsimeetodeid (Yu *et al*., 2021). Masinõpe on järjest enam hakanud kanda kinnitama pea kõikides teadusvaldkondades, puutumata ei ole jäänud ka

makroökoloogia (Anderson *et al*., 2021; Pichler & Hartig, 2023). Suurandmete tulekuga makroökoloogiasse on aina enam kasvanud nende haldamise ja analüüsimise keerukus, mistõttu on selles vallas töötavatel teadlastel võimalus pidevalt uusi meetodeid õppida (Farrell *et al*., 2021; Robeva *et al*., 2020).

Statistika (ingl *statistics*) on Tiidu ja Toodingu (2019, lk 236) järgi "matemaatika osa, mis tegeleb andmete kogumise, organiseerimise, töötlemise, analüüsimise, tõlgendamise ja esitamise ning selleks vajaliku teooria ja metoodika arendamisega". Statistikat võib laias laastus jaotada kaheks: kirjeldav statistika (ingl *descriptive statistics*) ja järeldusstatistika (ingl *inferential statistics*; Kaliyadan & Kulkarni, 2019). Kirjeldavas statistikas ei tehta uuritava valimi põhjal tõenäosusteoorial põhinevaid järeldusi, vaid antakse valimist kokkuvõtlik arvuline ja/või graafiline ülevaade (Cybernetica AS, 2023c; Kaliyadan & Kulkarni, 2019; Mishra *et al*., 2019). Seevastu järeldusstatistikas tegeletakse tõenäosusteooria põhjal hüpoteeside kontrollimise ja eri tüüpi analüüsidega (Kaliyadan & Kulkarni, 2019; Mishra *et al*., 2019).

Traditsiooniliste statistiliste analüüsimeetodite kõrvalt on viimastel aastakümnetel hakanud pead tõstma masinõpe (Farley *et al*., 2018; Jordan & Mitchell, 2015; Pichler & Hartig, 2023). Masinõpe on valdkond, kus arvuti saab andmete või kogemuste põhjal ise õppida, kasutades selleks arvutustehnilisi meetodeid, olles samal ajal selgesõnalise programmeerimiseta (Alzubi *et al*., 2018; Samuel, 1959). Masinõppe meetodid annavad võimaluse teha suurte andmemahtude pealt võimekamaid prognoosimudeleid ja viia läbi suuremahulisemaid analüüse (Farley *et al*., 2018; Jordan & Mitchell, 2015; Pichler & Hartig, 2023). Masinõppe meetodid võib jaotada kaheks: juhendatud masinõpe (ingl *supervised machine learning*) ja juhendamata masinõpe (ingl *unsupervised machine learning*; Alzubi *et al*., 2018). Juhendatud masinõppe puhul tuletab algoritm etteantud treeningandmete põhjal funktsiooni, juhendamata masinõppe puhul ei anta ette treeningandmeid ja olemasolevate mustrite tuvastamine toimub etteseatud reegliteta (Alzubi *et al*., 2018; Pichler & Hartig, 2023).

Paljud masinõppe meetoditest pakuvad lahendusi juba varem mainitud suurandmetega seotud probleemidele, ent ka analüüsi käigus tekkivatele probleemidele (Fan *et al*., 2014). Makroökoloogias annavad masinõppe meetodid võimaluse paremini teostada audio- ja videoanalüüsiga liigituvastust (Christin *et al*., 2019; Pichler & Hartig, 2023), prognoosida tunnuste muutumisi (Schiller *et al*., 2021; Viskari *et al*., 2015); hinnata liikide ohustatust looduskaitse perspektiivist (Bachman *et al*., 2024), loomade käitumist ja bioloogilist

mitmekesisust (Christin *et al*., 2019); modelleerida liikide levikut (Benkendorf & Hawkins, 2020; Pichler & Hartig, 2023; Prasad *et al*., 2006) ning teostada liigikaitset ja erinevate ökosüsteemide majandamist (Pichler & Hartig, 2023).

Aina kasvavate andmemahtudega muutub üha keerulisemaks ka nende statistiline analüüs (Beck *et al*., 2012; Marx, 2013). Praegu kasutatavad statistilised meetodid pole sageli ülekantavad ega võimelised analüüsima suuremahulisi ja mitmemõõtmelisi andmeid (Fan *et al*., 2014; Farley *et al*., 2018), kuna suurandmete analüüs nõuab meetodeid, mis oskavad arvestada suurandmete iseärasustega (Wüest *et al*., 2020). Makroökoloogia perspektiivist teeb ülekantavuse keeruliseks see, et uuritavad looduslikud protsessid toimuvad erinevatel ruumiskaaladel (näiteks lokaalne ja globaalne) ja on sageli skaalast sõltuvad (Beck *et al*., 2012; Farley *et al*., 2018; Willig *et al*., 2003). Suurandmete analüüsiga seotud katsumused motiveerivad arendama aina uusi statistilisi meetodeid (Fan *et al*., 2014). Oluline on leida algoritmid, mis on kiired, skaleeritavad suurele andmemahule ja mitmemõõtmelisusele (Farley *et al*., 2018; Wüest *et al*., 2020). See nõuab erinevate valdkondade (statistika, rakendusmatemaatika, masinõpe jms) kombineerimist ning erinevate statistiliste meetodite ja nende kasutamise teadlikkuse tõstmist (Fan *et al*., 2014; Michener & Jones, 2012).

5.9. Koostöö

On olnud aegu, kus (makro)ökoloogiat ei peetud valdkonnaks, mis osaleb suurte investeeringutega rahvusvahelistes uuringutes (ingl *big science*; Hampton *et al*., 2013). Leidub näiteid ka üsna ekstreemsetest juhtumistest, kus poldud nõus läbi viima rahvusvahelist uuringut, kuna see oleks nõudnud tol ajal (makro)ökoloogias tavatuid lähenemisi nagu uurimismeetodite ja lähenemiste ühtlustamine ning tsentraliseerimine (Michener *et al*., 2007). Aastakümnete möödumisega on õnneks näha märgatavat suhtumise muutumist (Michener & Jones, 2012). Makroökoloogia nõuab sageli koostööd külgnevate (bioinformaatika, evolutsioon, geneetika, klimatoloogia jms) kui ka lahknevate valdkondadega (andmeteadus, majandus, meditsiin, statistika jms), sillutades tee avatud teaduseni (ingl *open science*; Beck *et al*., 2012; Farley *et al*., 2018; McGill, 2019; Reichman *et al*., 2011; Voor *et al*., 2023; Wüest *et al*., 2020). Tänu koostöö parenemisele on võimalus luua etemaid andmete, analüüside ja standardite hoiustamise ning haldamise infrastruktuure (Farley *et al*., 2018; Hampton *et al*., 2013; Reichman *et al*., 2011).

Koostöös teiste oma ala ekspertidega on kvantitatiivsete lähenemistega võimalik kontrollida hüpoteese, mis on ka üldistusvõimelised (Fraser *et al*., 2013), ent neile lasub teatav kriitika

(Harrison, 2011). Üheks võimalikuks lahenduseks on erinevad rahvusvahelisel tasandil tehtavad koostöö uurimisvõrgustikud (ingl *coordinated distributed experiments*), kus samu katseid korraldavad erinevad uurimisrühmad eri maailma paigus samaaegselt kontrollitud standardite järgi (Fraser *et al*., 2013). Sellised koostöö uurimisvõrgustikud seljatavad logistika ja kulutustega seotud probleemid (Fraser *et al*., 2013). Näiteid edukatest koostöö uurimisvõrgustikest on International Tundra Experiment (ITEX), Nutrient Network (NUTNET), TreeDivNet ja Eestis arendatud DarkDivNet (Fraser *et al*., 2013; Pärtel *et al*., 2019).

Ambitsioonikatele koostöö ponnistustele vaatamata võib esineda olukordi, kus makroökoloogid pole aldis tegema koostööd oma või teiste valdkondadega (Farley *et al*., 2018; Gallagher *et al*., 2020; Jones *et al*., 2006). Tuleb praegugi veel ette olukordi, kus makroökoloogid nokitsevad üksinda mõne tehnilise lahenduse kallal, mida pole võimalik universaalse tööriistana kasutada ja mis on sobilik vaid konkreetse ülesande kontekstis (Farley *et al*., 2018). Makroökoloogidel tasuks olla veelgi julgem andmeteadlaste poole pöördumisega juhtudel, kui on vaja leida tehnilisi ja metodoloogilisi lahendusi makroökoloogiliste suurandmetega seotud probleemidele (Wüest *et al*., 2020). Samuti tasub olla avatum erinevate töövoo platvormide kasutamisele – nende tarbeks ei pea omama liialt kõrgeid tehnilisi teadmisi (Farley *et al*., 2018; Lee, 2017). Paratamatult võib koostööga tekkida juurde uusi probleeme, näiteks kipub valdkonnal olema oma terminoloogia ja standardid (Leonelli, 2019; Reichman *et al*., 2011), ent lõppude lõpuks on kõik kinni tahtmises (Palmer *et al*., 2005).

5.10. Oskused

Makroökoloogid peavad sageli omandama andmeteaduse, modelleerimise, andmebaaside haldamisega jms seonduvaid oskusi iseseisvalt, kuna need pole integreeritud eriala õppekavadesse või õpetatakse neid õppeaineid loodusteadustega seostamata (Hampton *et al*., 2017; Jones *et al*., 2006; Michener & Jones, 2012; Robeva *et al*., 2020). Hetkel kiputakse bioloogidele õpetama matemaatikat, milles omandatavad teadmised ei ole sobilikud praktiliseks rakendamiseks tänapäeva andmeteaduses ja loodusteadustes (Robeva *et al*., 2020). Üheks põhjuseks võib olla asjaolu, et vastutavad instituudi ei ole omavahel tihedas koostöös, ent samuti ei saa mööda vaadata asjaolust, et muutuste teadvustamine ja seejärel uuenduste sisseviimine võtab lihtsalt aega (Robeva *et al*., 2020).

Tuleb tõsta ka teadlikkust erinevate statistiliste meetodite ja nende kasutamise osas (Michener & Jones, 2012). Selleks, et kõigi akadeemilise bioloogiahariduse omandanute tööriistapagas oleks võimalikult ajaga kaasaskäiv, tasub õppekavasid pidevalt arendada ja teha koostööd ka teiste valdkondade instituutidega (Farley *et al*., 2018; Robeva *et al*., 2020). Kindlasti tasub kitsaskohti arutada läbi nii kraadiõppurite kui ka eriala spetsialistidega (Hampton *et al*., 2013). Samas ei pea tingimata alati vaid uusi õppeaineid lisama, vaid võib ka olemasolevaid täiendada, näiteks lisada väikeseid grupitöid, andmeanalüüsi projekte uutes programmeerimiskeeltes või õpetada olemasolevaid loodusteadustele suunitletult (Hampton *et al*., 2017; Robeva *et al*., 2020). Selline lähenemine annab tudengitele võimaluse näha andmeteaduse, matemaatika, statistika jms meetodite rakendamise võimalusi bioloogias, olla kursis uusimate valdkonna arengute ja põhilisemate andmetöötluse etappidega (Robeva *et al*., 2020). Õppekavade, eriti esimese õppeastme, uuendamise teeb keeruliseks asjaolu, et alusmaterjale, mida õpetada, on niivõrd palju, et selle ümber on keeruline lisada juurde veel uusi õppeaineid (Robeva *et al*., 2020). Peale kõrghariduse omandamist on võimalus hoida end uusimate meetoditega kursis, näidates üles initsiatiivi täiendada end koolitustel, töötubades ja kursustel (Farley *et al*., 2018; Hampton *et al*., 2013).

KOKKUVÕTE

Praeguseks on laialt levinud tõdemus, et makroökoloogia on jõudnud suurandmete ajastusse. Suurandmete käsitlus on makroökoloogias laiahaardeline ning potentsiaalikas ökoloogiliste protsesside ja mustrite uurimiseks. Makroökoloogia ja suurandmetega seotud terminoloogia areneb pidevalt, kuna nendega seotud arvutusvõimekuse ja analüüsi edusammud nõuavad pidevalt ajaga kaasaskäimist. Suurandmete laialdane kasutus on tekitanud nende määratluses ebakindlust, viies kahe erineva raamistiku kujunemiseni. Üks raamistik iseloomustab suurandmeid ühesõnaliste omaduste kaudu, teine aga kontseptualiseerib suurandmeid kui väga keerulist andmekogu oma teatavate iseärasustega. Vaatamata lahknevatele lähenemisviisidele, katavad mõlemad raamistikud suurandmete põhilisema olemuse ära.

Erinevad – sageli reaalajas ja automatiseeritud – vaatlused, seired, harrastusteaduse projektid ja võrgustikud aitavad kaasa primaarsete ja sekundaarsete makroökoloogiliste suurandmete tekkele. Laienenud digiteerimise võimalused on oluliselt tõstnud muuseumide, raamatukogude ja erakogude panust suurandmetesse. Nende andmete koondamine erineva otstarbega andmebaasidesse annab võimaluse tagada andmete kättesaadavust.

Andmete erinevaid funktsioone ja kasutamise viise on võimalik kujutada ette eluringina, mille igas etapis toimuvad sellele iseloomulikud tööprotsessid, toimides lõpuks spiraalina. Seejuures võib makroökoloogiliste suurandmete juures esile kerkida nii mõningaid probleeme nende kogumisest analüüsini välja. Need probleemid tulenevad peamiselt suurandmete keerulisest olemusest: maht, mitmelaadilisus, kiirus, tõepärasus ja väärtus. Standardite järgimine ja avatud teaduse põhimõtete omaksvõtt on nende proovikivide leevendamiseks hädavajalikud. Sellele vaatamata nõuavad mõned kitsaskohad ikka pigem kontekstipõhist lähenemist kui universaalseid lahendusi. Samuti saavad makroökoloogid kasutada üha enam masinõppes kasutatavaid meetodeid, mis nõuab tihedamat koostööd teiste valdkondade spetsialistidega. Seejuures tuleb olla pidevalt kursis kaasaegseimate meetodite kui ka eriala arengutega, jagades neid ka järeltuleva teadlaste põlvkonnaga.

Aina enam muutuvad makroökoloogias aktuaalseks suurandmetega seotud arutelud ja selle pakutavad võimalused, lahendamaks makroökoloogidele huvipakkuvaid teadusküsimusi. Selle kõrvalt ei tohiks unustada nende suurandmete pealt tuletatavat informatsiooni – see, mis rajabki tee uute teadmisteni.

SUMMARY

It is now widely accepted that macroecology has entered the era of big data. The big data approach in macroecology is broad and has the potential to study ecological processes and patterns. The terminology associated with macroecology and big data is constantly evolving, as advances in the computational and analytical power associated with big data constantly requires to keep up with the times. The widespread use of big data has created uncertainty in its definition, giving rise to two distinct frameworks. One framework characterises big data through one-word adjectives, while the other conceptualises it as a highly complex dataset with its own specific characteristics. Despite these divergent approaches, both frameworks capture the most fundamental nature of big data.

Various real-time automated observations, monitorings, citizen science projects, and networks contribute to primary and secondary macroecological data. The expanded possibilities of digitisation have significantly increased the contributions of museums, libraries, and private collections to big data. The aggregation of data into databases for different purposes provides an opportunity to ensure data availability.

The different functions and ways of using data can be represented as a life cycle, with specific work processes at each stage, ultimately acting as a spiral. At the same time, with macroecological big data, some problems may arise from their collection to analysis. These problems stem mainly from the complex nature of big data: volume, variety, velocity, veracity, and value. Adhering to standards and embracing the principles of open science are essential to mitigate these challenges. Nevertheless, some bottlenecks still require a context-specific approach rather than one-size-fits-all solutions. Macroecologists are also increasingly able to make use of machine-learning methods, which requires closer collaboration with specialists from other fields. In doing so, it is important to keep up to date with the most modern methods and developments in the field, sharing them with the next generation of scientists.

Conversations surrounding big data and its potential to address scientific inquiries are gaining prominence in macroecology. However, emphasis should not solely be on the data; equal attention should be given to their insights – insights that pave the way for new knowledge.

TÄNUAVALDUSED

Neid, keda soovin ära tänada, on palju. Soovin tänada oma juhendajat Meelis Pärtelit, kelle näpunäited, kannatlikkus ja lausa ennast unustavad vestlused muutsid lõputöö kirjutamise vägagi meeldivaks kogemuseks. Võiks isegi julgelt öelda, et leidsin taas üles selle sädeme teaduse vastu, mis mind kunagi ülikooli õppima tõi. Lisaks soovin tänada Ene Kooki, kelle julgustavad vestlused taimede süstemaatika ja fülogeneesi praktikumides tõid mind jäädavalt käesoleva lõputöö teemani. Kindlasti ei jaksa ma ära tänada Vete-Marit ja Minnat, kel oli minusse vahepeal isegi et rohkem usku kui mul endil. Tahan tänada ka Kellyt, kelle abiga sai käesolev töö pilbasteni läbi arutatud ja laused ümber sõnastatud. Samuti ei saa tänamata jätta kursusekaaslasi ja teisi kaasbiolooge, kellega koos veedetud aeg on loonud nii mõnedki kustumatud mälestused. Viimaks tahan tänada oma perekonda, tänu kelle kasvatustele ja õpetustele olen ma täna kujunenud selliseks inimeseks, nagu ma olen. Aitäh!

KASUTATUD KIRJANDUS

Aavik, T., Carmona, C. P., Träger, S., Kaldra, M., Reinula, I., Conti, E., Keller, B., Helm, A., Hiiesalu, I., Hool, K., Kaisel, M., Oja, T., Lotman, S., & Pärtel, M. (2020). Landscape context and plant population size affect morph frequencies in heterostylous Primula veris—Results of a nationwide citizen-science campaign. *Journal of Ecology*, *108*(6), 2169–2183. https://doi.org/10.1111/1365-2745.13488

Abarenkov, K., Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., May, T. W., Frøslev, T. G., Pawlowska, J., Lindahl, B., Põldmaa, K., Truong, C., Vu, D., Hosoya, T., Niskanen, T., Piirmann, T., Ivanov, F., Zirk, A., Peterson, M., Cheeke, T. E., Ishigami, Y., … Kõljalg, U. (2024). The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: Sequences, taxa and classifications reconsidered. *Nucleic Acids Research*, *52*(D1), D791–D797. https://doi.org/10.1093/nar/gkad1039

Abarenkov, K., Tedersoo, L., Nilsson, R. H., Vellak, K., Saar, I., Veldre, V., Parmasto, E., Prous, M., Aan, A., Ots, M., Kurina, O., Ostonen, I., Jõgeva, J., Halapuu, S., Põldmaa, K., Toots, M., Truu, J., Larsson, K.-H., & Kõljalg, U. (2010). PlutoF—a Web Based Workbench for Ecological and Taxonomic Research, with an Online Implementation for Fungal ITS Sequences. *Evolutionary Bioinformatics Online*, *6*, 189–196. https://doi.org/10.4137/EBO.S6271

Al-Mekhlal, M., & Ali Khwaja, A. (2019). A Synthesis of Big Data Definition and Characteristics. *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 314–322. https://doi.org/10.1109/CSE/EUC.2019.00067

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*(1), 012012. https://doi.org/10.1088/1742-6596/1142/1/012012

Alves, C., Castro, J. A., Ribeiro, C., Honrado, J. P., Lomba, Â., Alves, C., Castro, J. A., Ribeiro, C., Honrado, J. P., & Lomba, Â. (2018, 5. november). Research data management in the field of Ecology: An overview. *Proceedings of the International Conference on Dublin Core and Metadata Applications*. International Conference on Dublin Core and Metadata Applications. https://doi.org/10.23106/dcmi.952138986

Anderson, S. C., Elsen, P. R., Hughes, B. B., Tonietto, R. K., Bletz, M. C., Gill, D. A., Holgerson, M. A., Kuebbing, S. E., McDonough MacKenzie, C., Meek, M. H., & Veríssimo, D. (2021). Trends in ecology and conservation over eight decades. *Frontiers in Ecology and the Environment*, *19*(5), 274–282. https://doi.org/10.1002/fee.2320

Andrew, C., Heegaard, E., Kirk, P. M., Bässler, C., Heilmann-Clausen, J., Krisai-Greilhuber, I., Kuyper, T. W., Senn-Irlet, B., Büntgen, U., Diez, J., Egli, S., Gange, A. C., Halvorsen, R., Høiland, K., Nordén, J., Rustøen, F., Boddy, L., & Kauserud, H. (2017). Big data integration: Pan-European fungal species observations' assembly for addressing contemporary questions in ecology and global change biology. *Fungal Biology Reviews*, *31*(2), 88–98. https://doi.org/10.1016/j.fbr.2017.01.001

Andrews, L., Higgins, A., Andrews, M. W., & Lalor, J. G. (2012). *Classic Grounded Theory to Analyse Secondary Data: 11*(1).

Augustine, S. P., Bailey-Marren, I., Charton, K. T., Kiel, N. G., & Peyton, M. S. (2024). Improper data practices erode the quality of global ecological databases and impede the progress of ecological research. *Global Change Biology*, *30*(1), e17116. https://doi.org/10.1111/gcb.17116

Bachman, S. P., Brown, M. J. M., Leão, T. C. C., Nic Lughadha, E., & Walker, B. E. (2024). Extinction risk predictions for the world's flowering plants to support their conservation. *New Phytologist*, *242*(2), 797–808. https://doi.org/10.1111/nph.19592

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., … Wofsy, S. (2001). FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities. *Bulletin of the American Meteorological Society*, *82*(11), 2415–2434. https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2

Ball-Damerow, J. E., Brenskelle, L., Barve, N., Soltis, P. S., Sierwald, P., Bieler, R., LaFrance, R., Ariño, A. H., & Guralnick, R. P. (2019). Research applications of primary biodiversity databases in the digital age. *PLoS ONE*, *14*(9), e0215794. https://doi.org/10.1371/journal.pone.0215794

Beck, J., Ballesteros-Mejia, L., Buchmann, C. M., Dengler, J., Fritz, S. A., Gruber, B., Hof, C., Jansen, F., Knapp, S., Kreft, H., Schneider, A.-K., Winter, M., & Dormann, C. F. (2012). What's on the horizon for macroecology? *Ecography*, *35*(8), Article 8. https://doi.org/10.1111/j.1600-0587.2012.07364.x

Benkendorf, D. J., & Hawkins, C. P. (2020). Effects of sample size and network depth on a deep learning approach to species distribution modeling. *Ecological Informatics*, *60*, 101137. https://doi.org/10.1016/j.ecoinf.2020.101137

Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J. A., Mozzherin, D., Rees, T., Matasci, N., Narro, M. L., Piel, W. H., Mckay, S. J., Lowry, S., Freeland, C., Peet, R. K., & Enquist, B. J. (2013). The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinformatics*, *14*(1), 16. https://doi.org/10.1186/1471-2105-14-16

Brown, J. H., & Maurer, B. A. (1989). Macroecology: The Division of Food and Space Among Species on Continents. *Science*, *243*(4895), Article 4895. https://doi.org/10.1126/science.243.4895.1145

Bruelheide, H., Dengler, J., Purschke, O., Lenoir, J., Jiménez-Alfaro, B., Hennekens, S. M., Botta-Dukát, Z., Chytrý, M., Field, R., Jansen, F., Kattge, J., Pillar, V. D., Schrodt, F., Mahecha, M. D., Peet, R. K., Sandel, B., van Bodegom, P., Altman, J., Alvarez-Dávila, E., … Jandt, U. (2018). Global trait–environment relationships of plant communities. *Nature Ecology & Evolution*, *2*(12), Article 12. https://doi.org/10.1038/s41559-018-0699-8

Camerini, J. R. (2024, 15. aprill). Alfred Russel Wallace. *Encyclopedia Britannica*. Vaadatud 6.05.2024 https://www.britannica.com/biography/Alfred-Russel-Wallace

Cayuela, L., Granzow-de la Cerda, Í., Albuquerque, F. S., & Golicher, D. J. (2012). taxonstand: An r package for species names standardisation in vegetation databases. *Methods in Ecology and Evolution*, *3*(6), 1078–1083. https://doi.org/10.1111/j.2041-210X.2012.00232.x

Chandrakar, R. (2006). Digital object identifier system: An overview. *The Electronic Library*, *24*(4), 445–452. https://doi.org/10.1108/02640470610689151

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, *19*(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

Cheruvelil, K. S., Yuan, S., Webster, K. E., Tan, P.-N., Lapierre, J.-F., Collins, S. M., Fergus, C. E., Scott, C. E., Henry, E. N., Soranno, P. A., Filstrup, C. T., & Wagner, T. (2017). Creating multithemed ecological regions for macroscale ecology: Testing a flexible, repeatable, and accessible clustering method. *Ecology and Evolution*, *7*(9), 3046–3058. https://doi.org/10.1002/ece3.2884

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, *10*(10), Article 10. https://doi.org/10.1111/2041-210X.13256

Cornwell, W. K., Pearse, W. D., Dalrymple, R. L., & Zanne, A. E. (2019). What we (don't) know about global plant diversity. *Ecography*, *42*(11), 1819–1831. https://doi.org/10.1111/ecog.04481

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z.-Q., & Bourne, P. E. (2013). Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, *28*(8), 454–461. https://doi.org/10.1016/j.tree.2013.05.002

Cox, M., & Ellsworth, D. (1997). *Managing Big Data for Scientific Visualization*. 1–17.

Cybernetica AS. (2023a). Andmete kvaliteet. *Andmekaitse ja infoturbe portaal (AKIT)*. Vaadatud 9.02.2024 https://akit.cyber.ee/term/6121

Cybernetica AS. (2023b). DOI-süsteem. *Andmekaitse ja infoturbe portaal (AKIT)*. Vaadatud 21.03.2024 https://akit.cyber.ee/term/13244

Cybernetica AS. (2023c). Kirjeldav statistika. *Andmekaitse ja infoturbe portaal (AKIT)*. Vaadatud 30.01.2024 https://akit.cyber.ee/term/5814

Cybernetica AS. (2023d). Objekti digitaalidentifikaator. *Andmekaitse ja infoturbe portaal (AKIT)*. Vaadatud 21.03.2024 https://akit.cyber.ee/term/2404

Daru, B. H., Park, D. S., Primack, R. B., Willis, C. G., Barrington, D. S., Whitfeld, T. J. S., Seidler, T. G., Sweeney, P. W., Foster, D. R., Ellison, A. M., & Davis, C. C. (2018). Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist*, *217*(2), 939–955. https://doi.org/10.1111/nph.14855

Deakin University Library. (2023, 17. juuli). *Primary versus secondary data*. Vaadatud 3.03.2024 https://www.deakin.edu.au/library/research/manage-data/plan/primary-versus-secondary-data

Desmond, A. J. (2024, 24. aprill). Charles Darwin. *Encyclopedia Britannica*. Vaadatud 6.05.2024 https://www.britannica.com/biography/Charles-Darwin

Dimson, M., Berio Fortini, L., Tingley, M. W., & Gillespie, T. W. (2023). Citizen science can complement professional invasive plant surveys and improve estimates of suitable habitat. *Diversity and Distributions*, *29*(9), 1141–1156. https://doi.org/10.1111/ddi.13749

Diniz-Filho, J. A. F., Loyola, R. D., Raia, P., Mooers, A. O., & Bini, L. M. (2013). Darwinian shortfalls in biodiversity conservation. *Trends in Ecology & Evolution*, *28*(12), 689–695. https://doi.org/10.1016/j.tree.2013.09.003

DOI Foundation (2023, aprill). *DOI Handbook*. Vaadatud 21.03.2024 https://doi.org/10.1000/182

Dormann, C. F., Purschke, O., Márquez, J. R. G., Lautenbach, S., & Schröder, B. (2008). Components of Uncertainty in Species Distribution Analysis: A Case Study of the Great Grey Shrike. *Ecology*, *89*(12), 3371–3386. https://doi.org/10.1890/07-1772.1

Drew, J. A., Moreau, C. S., & Stiassny, M. L. J. (2017). Digitization of museum collections holds the potential to enhance researcher diversity. *Nature Ecology & Evolution*, *1*(12), Article 12. https://doi.org/10.1038/s41559-017-0401-6

Eck, N. van, & Waltman, L. (2009). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Edwards, J. L., Lane, M. A., & Nielsen, E. S. (2000). Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. *Science*, *289*(5488), 2312–2314. https://doi.org/10.1126/science.289.5488.2312

Ewing, E. T., Kimmerly, V., & Ewing-Nelson, S. (2016). Look Out for 'La Grippe': Using Digital Humanities Tools to Interpret Information Dissemination during the Russian Flu, 1889–90. *Medical History*, *60*(1), 129–131. https://doi.org/10.1017/mdh.2015.84

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293–314. https://doi.org/10.1093/nsr/nwt032

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience*, *68*(8), 563–576. https://doi.org/10.1093/biosci/biy068

Farrell, K. J., Weathers, K. C., Sparks, S. H., Brentrup, J. A., Carey, C. C., Dietze, M. C., Foster, J. R., Grayson, K. L., Matthes, J. H., & SanClements, M. D. (2021). Training macrosystems scientists requires both interpersonal and technical skills. *Frontiers in Ecology and the Environment*, *19*(1), 39–46. https://doi.org/10.1002/fee.2287

Favaretto, M., Clercq, E. D., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLOS ONE*, *15*(2), e0228987. https://doi.org/10.1371/journal.pone.0228987

Fer, I., Kelly, R., Moorcroft, P. R., Richardson, A. D., Cowdery, E. M., & Dietze, M. C. (2018). Linking big models to big data: Efficient ecosystem model calibration through Bayesian model emulation. *Biogeosciences*, *15*(19), 5801–5830. https://doi.org/10.5194/bg-15-5801-2018

Flantua, S. G. A., Mottl, O., Felde, V. A., Bhatta, K. P., Birks, H. H., Grytnes, J.-A., Seddon, A. W. R., & Birks, H. J. B. (2023). A guide to the processing and standardization of global palaeoecological data for large-scale syntheses using fossil pollen. *Global Ecology and Biogeography*, *32*(8), 1377–1394. https://doi.org/10.1111/geb.13693

Franklin, J., Serra-Diaz, J. M., Syphard, A. D., & Regan, H. M. (2017). Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeography*, *26*(1), 6–17. https://doi.org/10.1111/geb.12501

Fraser, L. H., Henry, H. A., Carlyle, C. N., White, S. R., Beierkuhnlein, C., Cahill Jr, J. F., Casper, B. B., Cleland, E., Collins, S. L., Dukes, J. S., Knapp, A. K., Lind, E., Long, R., Luo, Y., Reich, P. B., Smith, M. D., Sternberg, M., & Turkington, R. (2013). Coordinated distributed experiments: An emerging tool for testing global hypotheses in ecology and environmental science. *Frontiers in Ecology and the Environment*, *11*(3), 147–155. https://doi.org/10.1890/110279

Gallagher, R. V., Falster, D. S., Maitner, B. S., Salguero-Gómez, R., Vandvik, V., Pearse, W. D., Schneider, F. D., Kattge, J., Poelen, J. H., Madin, J. S., Ankenbrand, M. J., Penone, C., Feng, X., Adams, V. M., Alroy, J., Andrew, S. C., Balk, M. A., Bland, L. M., Boyle, B. L., … Enquist, B. J. (2020). Open Science principles for accelerating trait-based science across the Tree of Life. *Nature Ecology & Evolution*, *4*(3), Article 3. https://doi.org/10.1038/s41559-020-1109-6

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gonçalves-Souza, T., Chaves, L. S., Boldorini, G. X., Ferreira, N., Gusmão, R. A. F., Perônico, P. B., Sanders, N. J., & Teresa, F. B. (2023). Bringing light onto the Raunkiæran shortfall: A comprehensive review of traits used in functional animal ecology. *Ecology and Evolution*, *13*(4), e10016. https://doi.org/10.1002/ece3.10016

Hampton, S. E., Jones, M. B., Wasser, L. A., Schildhauer, M. P., Supp, S. R., Brun, J., Hernandez, R. R., Boettiger, C., Collins, S. L., Gross, L. J., Fernández, D. S., Budden, A., White, E. P., Teal, T. K., Labou, S. G., & Aukema, J. E. (2017). Skills and Knowledge for Data-Intensive Environmental Research. *BioScience*, *67*(6), Article 6. https://doi.org/10.1093/biosci/bix025

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), Article 3. https://doi.org/10.1890/120103

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*(4), 447–457. https://doi.org/10.1037/met0000120

Harris, M. A., Slippers, B., Kemler, M., & Greve, M. (2023). Opportunities for diversified usage of metabarcoding data for fungal biogeography through increased metadata quality. *Fungal Biology Reviews*, *46*, 100329. https://doi.org/10.1016/j.fbr.2023.100329

Harrison, F. (2011). Getting started with meta-analysis. *Methods in Ecology and Evolution*, *2*(1), 1–10. https://doi.org/10.1111/j.2041-210X.2010.00056.x

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, *118*(6), Article 6. https://doi.org/10.1073/pnas.2018093118

Higgins, D., Berkley, C., & Jones, M. B. (2002). Managing heterogeneous ecological data using Morpho. *Proceedings 14th International Conference on Scientific and Statistical Database Management*, 69–76. https://doi.org/10.1109/SSDM.2002.1029707

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W.-K., & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution*, *27*(2), Article 2. https://doi.org/10.1016/j.tree.2011.11.006

Hoekendijk, J. P. A., Kellenberger, B., Aarts, G., Brasseur, S., Poiesz, S. S. H., & Tuia, D. (2021). Counting using deep learning regression gives value to ecological surveys. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-02387-9

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, *46*(1), 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M., & Baselga, A. (2008). Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, *117*(6), 847–858. https://doi.org/10.1111/j.0030-1299.2008.16434.x

Huisman, J. M., & Millar, A. J. K. (2013). Australian seaweed collections: Use and misuse. *Phycologia*, *52*(1), 2–5. https://doi.org/10.2216/12-089.1

Johnston, A., Fink, D., Hochachka, W. M., & Kelling, S. (2018). Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution*, *9*(1), 88–97. https://doi.org/10.1111/2041-210X.12838

Johnston, A., Matechou, E., & Dennis, E. B. (2023). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, *14*(1), 103–116. https://doi.org/10.1111/2041-210X.13834

Johnston, M. P. (2014). Secondary Data Analysis: A Method of which the Time Has Come. *Qualitative and Quantitative Methods in Libraries*, *3*(3), Article 3.

Jones, M. B., Schildhauer, M. P., Reichman, O. J., & Bowers, S. (2006). The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere. *Annual Review of Ecology, Evolution, and Systematics*, *37*(1), 519–544. https://doi.org/10.1146/annurev.ecolsys.37.091305.110031

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

JSTOR Global Plants. (2013, 19. april). *Raunkiaer, Christen Christiansen (1860-1938)*. Vaadatud 15.05.2024 https://plants.jstor.org/stable/10.5555/al.ap.person.bm000006869

Kaliyadan, F., & Kulkarni, V. (2019). Types of Variables, Descriptive Statistics, and Sample Size. *Indian Dermatology Online Journal*, *10*(1), 82–86. https://doi.org/10.4103/idoj.IDOJ_468_18

Keppel, G., Craven, D., Weigelt, P., Smith, S. A., Van Der Sande, M. T., Sandel, B., Levin, S. C., Kreft, H., & Knight, T. M. (2021). Synthesizing tree biodiversity data to understand global patterns and processes of vegetation. *Journal of Vegetation Science*, *32*(3), e13021. https://doi.org/10.1111/jvs.13021

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, *3*(3), 262–267. https://doi.org/10.1177/2043820613513388

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 2053951716631130. https://doi.org/10.1177/2053951716631130

König, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J., & Kreft, H. (2019). Biodiversity data integration—The significance of data resolution and domain. *PLOS Biology*, *17*(3), e3000183. https://doi.org/10.1371/journal.pbio.3000183

Laanisto, L., & Pärtel, M. (2019). Makroökoloogia – mis see on, kust ta tuleb ja kuhu läheb? *Schola Biotheoretica*, *45*, 129–142.

LaDeau, S. L., Han, B. A., Rosi-Marshall, E. J., & Weathers, K. C. (2017). The Next Decade of Big Data in Ecosystem Science. *Ecosystems*, *20*(2), Article 2. https://doi.org/10.1007/s10021-016-0075-y

Laney, D. (2001, 6. veebruar). 3-D Data Management: Controlling Data Volume, Velocity and Variety. *Application Delivery Strategies by META Group Inc.*, *949*. https://pdfcoffee.com/ad949-3d-data-management-controlling-data-volume-velocity-and-variety pdf-pdf-free.html

Lang, P. L. M., Willems, F. M., Scheepens, J. F., Burbano, H. A., & Bossdorf, O. (2019). Using herbaria to study global environmental change. *New Phytologist*, *221*(1), 110–122. https://doi.org/10.1111/nph.15401

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational Social Science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons*, *60*(3), 293–303. https://doi.org/10.1016/j.bushor.2017.01.004

Leonelli, S. (2019). The challenges of big data biology. *eLife*, *8*, e47381. https://doi.org/10.7554/eLife.47381

L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*, *5*, 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365

Lobo, J. M. (2001). Decline of roller dung beetle (Scarabaeinae) populations in the Iberian peninsula during the 20th century. *Biological Conservation*, *97*(1), 43–50. https://doi.org/10.1016/S0006-3207(00)00093-8

Lobo, J. M., Baselga, A., Hortal, J., Jiménez-Valverde, A., & Gómez, J. F. (2007). How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? *Diversity and Distributions*, *13*(6), 772–780. https://doi.org/10.1111/j.1472-4642.2007.00383.x

Mahecha, M. D., Rzanny, M., Kraemer, G., Mäder, P., Seeland, M., & Wäldchen, J. (2021). Crowd‐sourced plant occurrence data provide a reliable description of macroecological gradients. *Ecography*, *44*(8), Article 8. https://doi.org/10.1111/ecog.05492

Marx, V. (2013). The big challenges of big data. *Nature*, *498*(7453), Article 7453. https://doi.org/10.1038/498255a

Mascarenhas, R., Ruziska, F. M., Moreira, E. F., Campos, A. B., Loiola, M., Reis, K., Trindade-Silva, A. E., Barbosa, F. A. S., Salles, L., Menezes, R., Veiga, R., Coutinho, F. H., Dutilh, B. E., Guimarães, P. R., Assis, A. P. A., Ara, A., Miranda, J. G. V., Andrade, R. F. S., Vilela, B., & Meirelles, P. M. (2020). Integrating Computational Methods to Investigate the Macroecology of Microbiomes. *Frontiers in Genetics*, *10*, 1344. https://doi.org/10.3389/fgene.2019.01344

McCleery, R., Guralnick, R., Beatty, M., Belitz, M., Campbell, C. J., Idec, J., Jones, M., Kang, Y., Potash, A., & Fletcher, R. J. (2023). Uniting Experiments and Big Data to advance ecology and conservation. *Trends in Ecology & Evolution*, *38*(10), 970–979. https://doi.org/10.1016/j.tree.2023.05.010

McGill, B. J. (2019). The what, how and why of doing macroecology. *Global Ecology and Biogeography*, *28*(1), Article 1. https://doi.org/10.1111/geb.12855

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, *19*(8), 992–1006. https://doi.org/10.1111/ele.12624

Michener, W. K., Beach, J. H., Jones, M. B., Ludäscher, B., Pennington, D. D., Pereira, R. S., Rajasekar, A., & Schildhauer, M. (2007). A knowledge environment for the biodiversity and ecological sciences. *Journal of Intelligent Information Systems*, *29*(1), 111–126. https://doi.org/10.1007/s10844-006-0034-8

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*(2), Article 2. https://doi.org/10.1016/j.tree.2011.11.016

Mishra, P., Pandey, C. M., Singh, U., Keshri, A., & Sabaretnam, M. (2019). Selection of Appropriate Statistical Methods for Data Analysis. *Annals of Cardiac Anaesthesia*, *22*(3), 297–301. https://doi.org/10.4103/aca.ACA_248_18

Moore, A. L., & McCarthy, M. A. (2016). Optimizing ecological survey effort over space and time. *Methods in Ecology and Evolution*, *7*(8), 891–899. https://doi.org/10.1111/2041-210X.12564

Moritz, C., & Agudo, R. (2013). The Future of Species Under Climate Change: Resilience or Decline? *Science*, *341*(6145), 504–508. https://doi.org/10.1126/science.1237190

Morueta-Holme, N., & Svenning, J.-C. (2018). Geography of Plants in the New World: Humboldt's Relevance in the Age of Big Data 1. *Annals of the Missouri Botanical Garden*, *103*(3), 315–329. https://doi.org/10.3417/2018110

Müller-Wille, S. (2024, 22. märts). Carolus Linnaeus. *Encyclopedia Britannica*. Vaadatud 6.05.2024 https://www.britannica.com/biography/Carolus-Linnaeus

Nelson, G., & Ellis, S. (2018). The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1763), 20170391. https://doi.org/10.1098/rstb.2017.0391

Nguyen, H., Katzfuss, M., Cressie, N., & Braverman, A. (2014). Spatio-Temporal Data Fusion for Very Large Remote Sensing Datasets. *Technometrics*, *56*(2), 174–185. https://doi.org/10.1080/00401706.2013.831774

Pacifici, K., Reich, B. J., Miller, D. A. W., & Pease, B. S. (2019). Resolving misaligned spatial data with integrated species distribution models. *Ecology*, *100*(6), e02709. https://doi.org/10.1002/ecy.2709

Palmer, M. A., Bernhardt, E. S., Chornesky, E. A., Collins, S. L., Dobson, A. P., Duke, C. S., Gold, B. D., Jacobson, R. B., Kingsland, S. E., Kranz, R. H., Mappin, M. J., Martinez, M. L., Micheli, F., Morse, J. L., Pace, M. L., Pascual, M., Palumbi, S. S., Reichman, O., Townsend, A. R., & Turner, M. G. (2005). Ecological science and sustainability for the 21st century. *Frontiers in Ecology and the Environment*, *3*(1), 4–11. https://doi.org/10.1890/1540-9295(2005)003[0004:ESASFT]2.0.CO;2

Parsons, M. A., Godøy, Ø., LeDrew, E., de Bruin, T. F., Danis, B., Tomlinson, S., & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science*, *37*(6), 555–569. https://doi.org/10.1177/0165551511412705

Pearse, W. D., Barbosa, A. M., Fritz, S. A., Keith, S. A., Harmon, L. J., Harte, J., Silvestro, D., Xiao, X., & Davies, T. J. (2018). Building up biogeography: Pattern to process. *Journal of Biogeography*, *45*(6), Article 6. https://doi.org/10.1111/jbi.13242

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*, 2041-210X.14061. https://doi.org/10.1111/2041-210X.14061

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218. https://doi.org/10.1145/505248.506010

Plesser, H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, *11*. https://doi.org/10.3389/fninf.2017.00076

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, *9*(2), 181–199. https://doi.org/10.1007/s10021-005-0054-1

Prendergast, J. R., Wood, S. N., Lawton, J. H., & Eversham, B. C. (1993). Correcting for Variation in Recording Effort in Analyses of Diversity Hotspots. *Biodiversity Letters*, *1*(2), 39–53. https://doi.org/10.2307/2999649

Pärtel, M., Carmona, C. P., Zobel, M., Moora, M., Riibak, K., & Tamme, R. (2019). DarkDivNet – A global research collaboration to explore the dark diversity of plant communities. *Journal of Vegetation Science*, *30*(5), 1039–1043. https://doi.org/10.1111/jvs.12798

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and Opportunities of Open Data in Ecology. *Science*, *331*(6018), Article 6018. https://doi.org/10.1126/science.1197962

Robeva, R. S., Jungck, J. R., & Gross, L. J. (2020). Changing the Nature of Quantitative Biology Education: Data Science as a Driver. *Bulletin of Mathematical Biology*, *82*(10), 127. https://doi.org/10.1007/s11538-020-00785-0

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Schiller, C., Schmidtlein, S., Boonman, C., Moreno-Martínez, A., & Kattenborn, T. (2021). Deep learning and citizen science enable automated plant trait predictions from photographs. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-95616-0

Schnase, J. L., Duffy, D. Q., Tamkin, G. S., Nadeau, D., Thompson, J. H., Grieg, C. M., McInerney, M. A., & Webster, W. P. (2017). MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service. *Computers, Environment and Urban Systems*, *61*, 198–211. https://doi.org/10.1016/j.compenvurbsys.2013.12.003

Shipley, J. R., Kelly, J. F., & Frick, W. F. (2018). Toward integrating citizen science and radar data for migrant bird conservation. *Remote Sensing in Ecology and Conservation*, *4*(2), 127–136. https://doi.org/10.1002/rse2.62

Smith, F. A., Lyons, S. K., Morgan Ernest, S. K., & Brown, J. H. (2008). Macroecology: More than the division of food and space among species on continents. *Progress in Physical Geography: Earth and Environment*, *32*(2), Article 2. https://doi.org/10.1177/0309133308094425

Snijders, C. C. P., Matzat, U., & Reips, U. D. (2012). „Big Data“: Big gaps of knowledge in the field of internet science. *International Journal of Internet Science*, *7*(1), 1–5.

Zipkin, E. F., Zylstra, E. R., Wright, A. D., Saunders, S. P., Finley, A. O., Dietze, M. C., Itter, M. S., & Tingley, M. W. (2021). Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment*, *19*(1), 30–38. https://doi.org/10.1002/fee.2290

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261–1277. https://doi.org/10.1111/ecog.04960

Tiit, E-M. & Tooding, L-M. (2019). Statistikaleksikon. Tartu: Tartu Ülikooli Kirjastus

Troia, M. J., & McManamay, R. A. (2016). Filling in the GAPS: Evaluating completeness and coverage of open-access biodiversity databases in the United States. *Ecology and Evolution*, *6*(14), 4654–4669. https://doi.org/10.1002/ece3.2225

Vilela, B., & Villalobos, F. (2015). letsR: A new R package for data handling and analysis in macroecology. *Methods in Ecology and Evolution*, *6*(10), 1229–1234. https://doi.org/10.1111/2041-210X.12401

Viskari, T., Hardiman, B., Desai, A. R., & Dietze, M. C. (2015). Model-data assimilation of multiple phenological observations to constrain and predict leaf area index. *Ecological Applications*, *25*(2), 546–558. https://doi.org/10.1890/14-0497.1

Voor, T., Pärtel, M., Peet, A., Saare, L., Hyöty, H., Knip, M., Davison, J., Zobel, M., & Tillmann, V. (2023). Atopic sensitization in childhood depends on the type of green area around the home in infancy. *Clinical & Experimental Allergy*, *53*(8), 850–853. https://doi.org/10.1111/cea.14317

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Willig, M. R., Kaufman, D. M., & Stevens, R. D. (2003). Latitudinal Gradients of Biodiversity: Pattern, Process, Scale, and Synthesis. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 273–309. https://doi.org/10.1146/annurev.ecolsys.34.012103.144032

Wolf, S., Mahecha, M. D., Sabatini, F. M., Wirth, C., Bruelheide, H., Kattge, J., Moreno Martínez, Á., Mora, K., & Kattenborn, T. (2022). Citizen science plant observations encode global trait patterns. *Nature Ecology & Evolution*, *6*(12), Article 12. https://doi.org/10.1038/s41559-022-01904-x

Wüest, R. O., Zimmermann, N. E., Zurell, D., Alexander, J. M., Fritz, S. A., Hof, C., Kreft, H., Normand, S., Cabral, J. S., Szekely, E., Thuiller, W., Wikelski, M., & Karger, D. N. (2020). Macroecology in the age of Big Data – Where to go from here? *Journal of Biogeography*, *47*(1), Article 1. https://doi.org/10.1111/jbi.13633

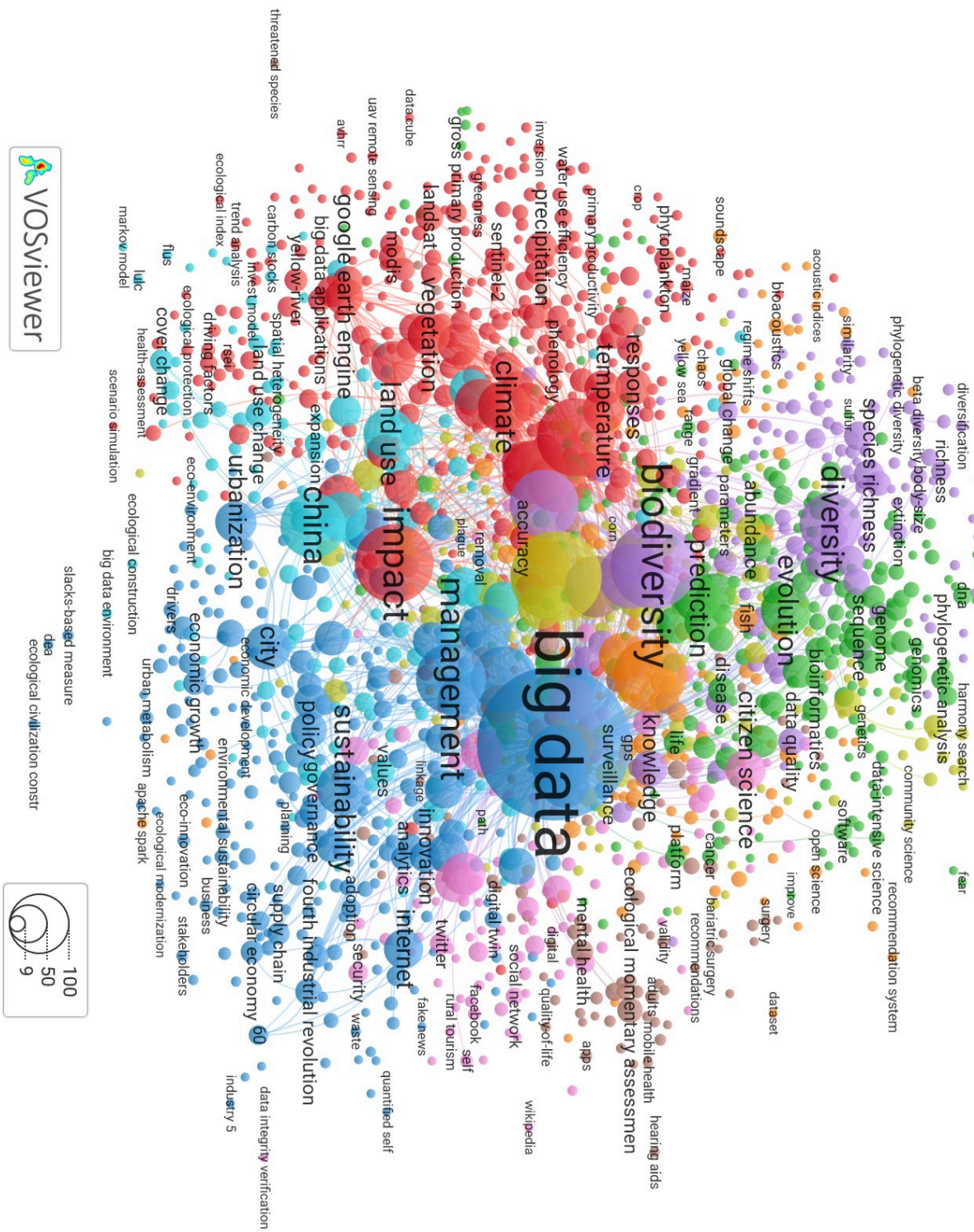
Öpik, M., Vanatoa, A., Vanatoa, E., Moora, M., Davison, J., Kalwij, J. M., Reier, Ü., & Zobel, M. (2010). The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist*, *188*(1), 223–241. https://doi.org/10.1111/j.1469-8137.2010.03334.x

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A., & Culham, A. (2007). How Global Is the Global Biodiversity Information Facility? *PLOS ONE*, *2*(11), e1124. https://doi.org/10.1371/journal.pone.0001124

Yu, Q., Ji, W., Prihodko, L., Ross, C. W., Anchang, J. Y., & Hanan, N. P. (2021). Study becomes insight: Ecological learning from machine learning. *Methods in Ecology and Evolution*, *12*(11), Article 11. https://doi.org/10.1111/2041-210X.13686

LISAD

Lisa 1. Rühmitatud märksõnad, mis esinevad suurandmeid käsitlevates makroökoloogilistes uuringutes kirjanduse andmebaasis Web of Science.



Lisa 2. Valik makroökoloogilisi andmebaase

Siinkohal toon oma parima teadmise juures välja mõned olulisemad makroökoloogilised andmebaasid üle erinevate elusorganismide rühmade, olles lahterdatud temaatiliselt tähestikulises järjekorras. Kirjelduste tegemisel on kasutatud refereerivalt andmebaaside tutvustavat infot.

Elurikkus ja levik

- **Botanical Information Network and Ecology Network** (BIEN, https://bien.nceas.ucsb.edu/bien/) – läbi R-i paketi kättesaadav andmebaas, mis kajastab taimeliikide levikut, arvukust ja tunnuseid;
- **eBird** (https://ebird.org/) – andmebaas linnuvaatlustega soetud andmete kohta;
- **eElurikkus** (https://elurikkus.ee/) – andmebaas, mis koondab kokku Eesti liikide vaatlused ja sellega seotud andmed, mis on omakorda ühildatud PlutoF-i liidesega;
- **FishBase** (https://www.fishbase.se/search.php) – andmebaas, kus on kaladeliikide nimistu koos tunnuseid ja elupaiku puudutavad andmed;
- **Global Index of Vegetation-Plot Databases** (GIVD, https://www.givd.info/databases.xhtml) – keskne andmebaas, mis koondab mitmest andmebaasist kokku andmed maailma taimkatte kohta;
- **Global Inventory of Floras and Traits** (GIFT, https://gift.uni-goettingen.de/home) – andmebaas, mis koondab kokku taimeliikide nimistud ja funktsionaalsed tunnused, olles suunitletud makroökoloogiale ja biogeograafiale;
- **IUCN Spatial Data** (https://www.iucnredlist.org/resources/spatial-data-download) – andmebaas, mis pakub andmeid IUCN-i punasesse nimestikku kuuluvate liikide leviku kohta;
- **Map of Life** (MOL, https://mol.org/) – keskne andmebaas, mis koondab kokku erinevaid elurikkusega seotud andmeid: vaatlused, liikide levikukaardid, liigikaitse, elurikkuse indeksid ja ökoloogilised mustrid;
- **Movebank** (https://www.movebank.org/) – andmebaas loomade liikumiste ja rännete kohta;
- **Ocean Biogeographic Information System** (OBIS, https://obis.org/) – merede ja ookeanide elurikkust koondav andmebaas (liiginimistud, tunnused, sekventsid jms);
- **Plants of the World Online** (https://powo.science.kew.org/) – andmebaas, mis pakub taimede elurikkust puudutavaid andmeid: nomenklatuur, tunnused, molekulaarsed andmed, ohustatuse tase jms;

- **sPlot** (https://www.idiv.de/en/splot.html) – andmebaas, mis koondab kokku taimkatet ja liigilist koosseisu puudutavad andmed;
- **VegBank** (http://vegbank.org/vegbank/index.jsp) – keskne andmebaas, mis koondab kokku taimkatet puudutavad graafilised andmed.

Funktsionaalsed tunnused

- **AnimalTraits** (https://animaltraits.org/) – andmebaas mis koondab kokku andmed maismaaloomade funktsionaalsete tunnuste kohta: kehamass, metabolism ja aju suurus;
- **DataOne** (https://www.dataone.org/) – keskne andmebaas, mis koondab üle mitme repositooriumi kokku erinevad keskkonna- ja morfoloogiliste tunnuste andmed, mis on teadustööde käigus kogutud ja avaldatud;
- **Encyclopedia of Life** (https://eol.org/) – keskne andmebaas, mis koondab üle mitme koostööpartneri (nii andmebaasid, laboratooriumid kui muuseumid) kokku elusloodu rühmade tunnused ning pakub ka toiduvõrgustike diagramme;
- **Global Biodiversity Information Facility** (GBIF, https://www.gbif.org/) – väga rangete standarditega keskne infrastruktuur, mis koondab üle erinevate andmebaaside ja nimistute kokku andmeid elusloodu vaatluste ja eksemplaride kohta;
- **LEDA** (https://uol.de/en/landeco/research/leda) – andmebaas, mis pakub andmeid Loode-Euroopa taimede tunnuste kohta;
- **MammalBase** (https://www.mammalbase.net/index_about) – andmebaas imetajate funktsionaalsete tunnuste ja toitumise kohta;
- **TRY** (https://www.try-db.org/TryWeb/Home.php) – andmebaas taimede tunnuste kohta.

Fülogenees

- **BirdTree** (http://birdtree.org/) – andmebaas lindude fülogeneesi kohta;
- **Open Tree of Life** (https://opentreeoflife.github.io/) – andmebaas arhede, bakterite ja eukarüootide fülogeneesipuude ning ka teadustööde raames kogutud fülogeneesiandmete kohta;
- **PHYLACINE_1.2** (https://megapast2future.github.io/PHYLACINE_1.2/) – andmebaas imetajate makroökoloogiliste fülogeneesiandmete kohta;
- **TreeBase** (https://www.treebase.org/) – keskne andmebaas, mis koondab kokku elusloodu fülogeneetilised andmed üle erinevate repositooriumide ja muuseumide;

- **TreeFam**, http://www.treefam.org/ – andmebaas loomade/eukarüootide fülogeneesipuude kohta;
- **VertLife**, https://vertlife.org/ – andmebaas maismaa selgroogsete fülogeneesi kohta.

Geneetika

- **BOLD** (https://www.boldsystems.org/index.php) – andmebaas, mis hõlmab DNA-triipkoodide andmeid;
- **European Nucleotide Archive** (https://www.ebi.ac.uk/ena/browser/home) – andmebaas nukleotiidide järjestuste kohta, millega sarnast andmebaasi pakub Jaapanis poolt kureeritav **DNA DataBank of Japan** (https://www.ddbj.nig.ac.jp/index-e.html);
- **Genome Taxonomy Database**, https://gtdb.ecogenomic.org/ – andmebaas bakterite ja arhede geneetika ja taksonoomia kohta;
- **MaarjAM** (https://maarjam.ut.ee/) – andmebaas, mis hõlmab endas spetsiifiliselt arbuskulaarse mükoriisa DNA-järjestuste andmeid;
- **National Library of Medicine, National Center for Biotechnology Information** (NCBI, https://www.ncbi.nlm.nih.gov/) – pakub keskset juurdepääsu mitmetele biomeditsiini ja geneetika andmebaasidele, millest üks esinduslikuim on **GenBank** (https://www.ncbi.nlm.nih.gov/genbank/), mis hoiustab DNA-järjestuste andmeid;
- **Phytozome**, https://phytozome-next.jgi.doe.gov/ – andmebaas taimede sekveneeritud DNA andmete kohta;
- **Plant GARDEN**, https://plantgarden.jp/en/index – andmebaas taimede genoomide ja markerite kohta;
- **UNITE** (https://unite.ut.ee/) – andmebaas, mis hõlmab endas eukarüootide tuuma ribosoomide ITS-regiooni andmeid..

Keskkonna monitoorimine

- **Terrestrial Ecosystem Research Network** (TERN, https://www.tern.org.au/) – andmebaas, mis koondab seirete ja välitööde käigus kogutud keskkonnaandmeid.

Kogud ja kirjandus

- **Australasian Virtual Herbarium** (AVH, https://avh.chah.org.au/) – virtuaalne herbaarium, mis koondab kokku Austraalia ja Uus-Meremaa taimestiku koos keskkonna- ja elupaiga andmetega;
- **Botanicus** (http://www.botanicus.org/) – andmdebaas ajaloolise botaanikaalase kirjanduse kohta;

- **Harvard University Herbaria & Libraries Databases** (https://kiki.huh.harvard.edu/databases/) – mitmele andmebaasile keskset juurdepääsu pakkuv repositoorium seoses botaaniliste kogude, nende kogujate, kirjanduse ja piltidega;
- **JACQ** (https://www.jacq.org/) – virtuaalne herbaarium, mis koondab kokku eksemplare Kesk-Ameerika, Euroopa ja Kesk-Aasia taimestiku kohta;
- **Royal Botanic Gardens Edinburgh online herbarium** (RBGE, https://websites.rbge.org.uk/multisite/multisite3.php) – virtuaalne herbaarium, mis koondab kokku eksemplare Euroopa ja Põhja-Ameerika kogudest.

Paleontoloogia

- **Neotoma Paleoecology Database** (https://www.neotomadb.org/) – andmebaas paleoökoloogiliste andmete kohta;
- **The Paleobiology Database** (https://paleobiodb.org/#/) – andmebaas paleontoloogiliste andmete kohta.

Taksonoomia ja nomenklatuur

- **AlgaeBase** (https://www.algaebase.org/) – andmebaas vetikate taksonoomia, nomenklatuuri ja leviku kohta;
- **ASM Mammal Diversity Database** (https://www.mammaldiversity.org/) – andmebaas elavate ja hiljuti väljasurnud imetajate taksonoomia kohta;
- **Catalogue of Life** (https://www.catalogueoflife.org/) – keskne nimestik, mis koondab üle mitme allika kokku kõikide elusloodu rühmade liiginimed ja nende sünonüümid;
- **International Plant Names Index** (IPNI, https://www.ipni.org/) – andmebaas soontaimede taksonoomia kohta;
- **MYCOBANK** (https://www.mycobank.org/) – andmebaas mükoloogia nomenklatuuri kohta;
- **The World Flora Online** (https://www.worldfloraonline.org/) – andmebaas taimede taksonoomia kohta.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Merilin Radvilavicius,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

   "Suurandmete kasutus makroökoloogias", mille juhendaja on prof Meelis Pärtel,

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Merilin Radvilavicius

22.05.2024