MILJØ-
DIREKTORATET

RAPPORT

M-1638 | 2020

Kriterier for lagring av miljø-DNA prøver og data, herunder henvisning til referansemateriale

KOLOFON

Utførende institusjon

NTNU Vitenskapsmuseet

Oppdragstakers prosjektansvarlig

Anders Finnstad

Kontaktperson i Miljødirektoratet

Sunniva Aagaard

M-nummer

M-1638|2020

År

2020

Sidetall

40

Miljødirektoratets kontraktnummer

nr. 19087461

Utgiver

Miljødirektoratet

Prosjektet er finansiert av

Miljødirektoratet

Forfatter(e)

Finstad et. al.

Tittel – norsk og engelsk

Kriterier for lagring av miljø-DNA prøver og data, herunder henvisning til referansemateriale
Criteria for depositing eDNA samples and data, including vouchered specimens

Kort sammendrag – short summary

Denne rapporten er bestilt av Miljødirektoratet fra NorBOL (Norwegian Barcode of Life), et nasjonalt nettverk av forskningsinstitusjoner som koordineres av NTNU Vitenskapsmuseet.
NTNU Vitenskapsmuseet har hatt prosjektledelsen.

4 emneord

Miljø-DNA, referansesekvenser, referansebibliotek, datahåndtering, prøvelagring

4 subject words

eDNA, reference sequences, sequence database, data mangament, storage of samples

Forsidefoto

Torbjørn Ekrem, NTNU Vitenskapsmuseet (CC-BY 4.0)

Kriterier for lagring av miljø-DNA prøver og data, herunder henvisning til referansemateriale

Forfattere
Anders Gravbrøt Finstad[1], Hugo de Boer[13], May Bente Brurberg[2], Geir Dahle[3], Kristin Skarsfjord Edgar[4], Alexander Eiler[5,6],Torbjørn Ekrem[1], Dag Endresen[7,13], Frode Fossøy[8], Haakon Hansen[9], Anders Hobæk[10], Stein Arild Hoem[11], Aino Hosia[12], Knut Anders Hovstad[11], Thomas Stjernegaard Jeppesen[7], Arild Johnsen[13], Eveliina Kallioiniemi[11], Aud Larsen[14], Jan T. Lifjeld[13], Iva Pitelkova[15], Maria Prager[16,17], Jessica Louise Ray[14], Ingrid Salvesen[1] , Trude Vrålstad[9], Endre Willassen[12]

[1]NTNU Vitenskapsmuseet, 7491-Trondheim
[2]Norsk institutt for bioøkonomi, Postboks 115, 1431 Ås
[3]Havforskningsinstituttet, Postboks 1870 Nordnes, 5817 Bergen
[4]Folkehelseinstituttet, Postboks 222 Skøyen, 0213 Oslo
[5]Institutt for Biovitenskap, Universitetet i Oslo, Postboks 1172 Blindern, 0318 Oslo
[6]eDNA Solutions AB, Gøteborg, Sverige
[7]Global Biodiversity Information Facility (GBIF), Universitetsparken 15, 2100 København Ø, Danmark
[8]Norsk institutt for naturforskning, Postboks 5685 Torgarden, 7485 Trondheim
[9]Veterinærinstituttet, Postboks 750 Sentrum, 0106 Oslo
[10]Norsk institutt for vannforskning, Gaustadallén 21, 0349 Oslo
[11]Artsdatabanken, Havnegata 9, 7010 Trondheim
[12]Universitetsmuseet i Bergen, Universitetet i Bergen, Postboks 7800, 5020 Bergen
[13]Naturhistorisk museum, Universitetet i Oslo, Postboks 1172 Blindern, 0318 Oslo
[14]NORCE, Postboks 22 Nygårdstangen, 5838 Bergen
[15]Norges arktiske universitetsmuseum, UiT Norge arktiske universitet, Postboks 6050 Langnes, 9037 Tromsø
[16]Science for Life Laboratory, Institutionen för ekologi, miljö och botanik, Stockholms Universitet, SE-106 91 Stockholm
[17]Gothenburg Global Biodiversity Centre, Göteborgs Universitet, Carl Skottsbergs gata 22B , SE-41319 Göteborg

Innhold

Sammendrag

Denne rapporten er bestilt av Miljødirektoratet fra NorBOL (Norwegian Barcode of Life), et nasjonalt nettverk av forskningsinstitusjoner som koordineres av NTNU Vitenskapsmuseet. NTNU Vitenskapsmuseet har hatt prosjektledelsen.

Miljødirektoratet ønsker å sikre en god fundering og etterprøvbarhet av offentlige oppdrag som inkluderer bruk av DNA-strekkoding og miljø-DNA. Det innebærer å sette krav til:

1. En omforent minstestandard på henvisning til og bruk av offentlig tilgjengelig referansemateriale.
2. En felles plattform for innrapportering og offentlig tilgjengeliggjøring av DNA- sekvenser og tilhørende metadata fra oppdrag.
3. Oppbevaring av innsamlede prøver som ikke analyseres umiddelbart over lengre tid for å sikre at disse er tilgjengelig for framtidig forskning og forvaltning, uavhengig av opprinnelig oppdragsinstitusjon.

På bakgrunn av gjeldende praksis i norske fagmiljøer og nasjonal og internasjonal tilgjengelig infrastruktur, gir dette dokumentet en gjennomgang og vurdering av hvilke kriterier forvaltningen kan vurdere å kreve av tilbydere i håndtering av og referering til referansemateriale, DNA-sekvenser fra miljø-DNA og DNA-strekkoding studier. Dokumentet gir også en vurdering av hvilke krav som eventuelt kan settes til lagring av prøver for senere bruk, samt tilhørende dokumentasjon og metadata. Dokumentet går imidlertid ikke inn på tekniske krav til design av studier og tekniske detaljer rundt laboratorie-protokoller og analyser. Dette er viktige tema, men utenfor rammene til denne teksten.

Rapporten er basert på informasjon fra arbeidsmøte avholdt på Miljødirektoratet 24.09.2019 med deltagere fra sentrale norske Institusjoner innen fagområdet; NTNU, NINA, NIVA, UiO, UiB, UiT, Folkehelseinstituttet, NIBIO, NORCE, Havforskningsinstituttet, Veterinærinstituttet, Artsdatabanken og Miljødirektoratet. I tillegg deltok representanter fra relevante internasjonale institusjoner, firmaer og infrastrukturer (GBIF, Science for Life Laboratory, eDNA Solutions).

Representanter fra ulike institusjoner redegjorde i en serie presentasjoner for dagens praksis innen bruk av referansedatabaser, data og materiale fra studier som benytter DNA-strekkoding og andre molekylære verktøy til analyse av biologisk mangfold (inkludert miljø-DNA) i egen institusjon. I tillegg er informasjon innsamlet i ettertid.

Redaktører av rapporten har vært; Anders G. Finstad (kapittel 1.4, 2.2 samt 3), Torbjørn Ekrem (kapittel 1.3 og 2.1), Frode Fossøy (kapittel 1.1 og 1.2) og Arild Johnsen (kapittel 1.5 og 2.3). Sunniva Aagaard har vært Miljødirektoratets kontaktperson i arbeidet, og Dag Rosland, Siv Grethe Aarnes og Ragnvald Larsen fra Miljødirektoratet har blitt konsultert.

1. Innledning

Dokumentet er ment som en veileder og ikke som et teknisk-standard dokument. Dette gjenspeiler seg i fremstillingen, og i avveiningen mellom teknisk presisjon og lesbarhet er det siste gitt prioritet. Eksempler på teknisk beskrivelse av miljø-DNA data er gitt som elektronisk vedlegg (Vedlegg 1: https://github.com/NTNU-VM/Darwinize-eDNA).

Bruk av molekylære metoder til biodiversitetsinventeringer er ikke nytt, men forholdsvis lite brukt innen forvaltning. Dokumentet vil derfor starte med en generell innføring til ulike metoder.

1.1 Hva er miljø-DNA?

Miljø-DNA er spor av arvestoff (DNA) hentet fra miljøprøver uten noen åpenbare tegn på biologisk kildemateriale (Thomsen and Willerslev 2015). Miljø-DNA er et relativt nytt begrep slik det forstås i dag. Uttrykket har imidlertid vært i bruk siden 1987, men da knyttet til DNA fra mikrober i sedimentprøver (Ogram, Sayler, and Barkay 1987). Nå definerer vi miljø-DNA bredt som en kompleks miks av DNA fra ulike organismer (Taberlet et al. 2018, 2012). Miljø-DNA er altså alt DNA vi kan finne i en spesifikk miljøprøve, uavhengig av hvilket substrat miljøprøven kommer fra og hvilke arter den inneholder. DNA i miljøet kan stamme fra hud- og hårceller, spytt og avføring med mer fra levende eller nylig døde organismer (Pietramellara et al. 2009). Miljø-DNA er derfor i utgangspunktet uspesifikt og representer ideelt sett alle arter i et gitt økosystem. I praksis er imidlertid tilstedeværelse av DNA i miljøprøven avhengig av kroppsstørrelsen til organismen, morfologi, aktivitetsnivå og habitatvalg for de ulike artene (Taberlet et al. 2018).

Flere studier viser at enkle vannprøver og analyser basert på miljø-DNA kan ha en høyere sannsynlighet for å finne sjeldne arter enn konvensjonelle metoder (Thomsen et al. 2012; Biggs et al. 2015; Valentini et al. 2016). Miljø-DNA kan derfor være velegnet for overvåking av sjeldne rødlistearter og uønskede fremmede arter som ofte har lave tettheter og som er vanskelige å oppdage med konvensjonelle metoder. I tillegg kan miljø-DNA også benyttes til observasjon av mange arter samtidig, og beskrive hele eller deler av et artssamfunn (Ekrem and Majaneva 2019). Identifikasjon av sekvenser avhenger av at referansebiblioteket er dekkende, noe som ofte ikke er tilfelle. Taksonomisk kompetanse og forsiktighet i vurdering av resultater er derfor påkrevet.

Enkelte studier viser en sammenheng mellom mengden DNA i en miljøprøve og biomassen av arten i miljøet. Man kan derfor potensielt også tenke seg miljø-DNA brukt til å lage et såkalt semikvantitativt estimat (indirekte mål) for biomasse av en art, både fra miljøprøver og bulkprøver (Takahara et al. 2012; Thomsen et al. 2012; Andersen et al. 2012; Lacoursière-Roussel, Rosabal, and Bernatchez 2016; Thomsen et al. 2016; Valentini et al. 2016; Fossøy et al. 2019; Yates, Fraser, and Derry 2019; Doi et al. 2017). Andre studier viser imidlertid liten sammenheng mellom miljø-DNA mengde og estimert populasjonstetthet (f.eks. S. W. Knudsen et al. 2019). For eksempel kan skallskifte, reproduksjon og massedød bidra til økte nivåer av miljø-DNA i vannet hos krepsdyr, mens turbiditet og dårlig vannkvalitet reduserer mengden

påviselig miljø-DNA (Strand, Johnsen, et al. 2019). Kvantitative estimater av bestandsstørrelser fra miljø-DNA vil derfor kreve mer uttesting før eventuell operasjonalisering anses mulig.

Miljø-DNA er altså en prøvetype, og ikke en metode. Utgangspunktet for miljø-DNA undersøkelser kan være både vannprøver, jordprøver og luftprøver. Ofte inkluderes også DNA fra avføringsprøver og bulkprøver (f.eks. planktonprøver eller malaisefelleprøver bestående av flere individer fra mange arter) som miljø-DNA (Taberlet et al. 2018). For å studere miljø-DNA finnes det en rekke analysemetoder (Tekstboks 1.1). Man kan dele disse inn i to hovedtyper der man 1) ønsker å påvise en spesifikk art eller 2) ønsker å beskrive et samfunn av flere arter. Ulike analysemetoder vil generere ulike typer og mengder av data.

Vi vil videre se litt nærmere på hva man kan finne ut ved hjelp av de forskjellige metodene og hvilke typer data som blir produsert. Det finnes mange feilkilder ved analyser av miljø-DNA som kan både føre til at man ikke finner en art (falsk negativ) og til at man påviser en art som ikke finnes i prøven (falsk positiv). Det presiseres derfor at resultater fra miljø-DNA studier kan være krevende å tolke, og at alle resultater må evalueres av økologiske og taksonomiske eksperter.

Tekstboks 1.1 : Miljø-DNA begreper

Generelt

Miljø-DNA (environmental DNA - eDNA)
- Løst eller bundet DNA i en miljøprøve, f.eks. jord, vann eller luft. En vanlig brukt definisjon er at miljø-DNA er arvestoff (DNA) hentet fra miljøprøver uten noen åpenbare tegn på biologisk kildemateriale (Thomsen and Willerslev 2015).

Markør
- Et navngitt DNA-fragment. Kan, men må ikke, være en del av et gen.

Primer
- Et kort, syntetisk enkelttrådet DNA-fragment som binder seg til utvalgt område på mål-DNA (markør) under PCR. Nødvendig for at enzymet polymerase skal kopiere utvalgt markør.

Probe
- Et kort, syntetisk enkelttrådet DNA-fragment med fluoriserende merking som binder seg til utvalgt område på mål-DNA (markør) under PCR. Øker spesifisiteten og kan brukes i tillegg til primere i qPCR og ddPCR for å påvise og kvantifisere en genetisk markør.

Påvisning av enkeltarter (species-specific detection)

qPCR (quantitative Polymerase Chain Reaction
- Kvantitativ PCR. Metode som måler relativ DNA-mengde av en markør i en prøve.

ddPCR (droplet digital Polymerase Chain Reaction
- Dråpe digital PCR. Metode som måler absolutt DNA-mengde (antall kopier) av en markør i en prøve.

Beskrivelse av artssamfunn

DNA-strekkoding (DNA-barcoding)
- Bruk av et kort, standardisert DNA-fragment til identifisering av enkeltarter.

DNA-metastrekkoding (DNA-metabarcoding)
- Database som inneholder DNA-sekvenser (DNA-strekkoder) fra identifiserte individer av kjente arter eller høyere taksonomisk nivå (f. eks. slekt, familie).

Organelle metagenomikk (organellar metagenomics)
- PCR-fri sekvensering av mitokondrielle genomer i en blandet prøve.

Målrettet DNA-fragment fangst og sekvensering (target-capture sequencing)
- Sekvensering av DNA-fragmenter isolert med hybridiseringsprober.

1.2 Bruk av qPCR og ddPCR for påvisning av enkeltarter

For påvisning av enkeltarter bruker man som oftest qPCR (Quantitative Polymerase Chain Reaction) eller ddPCR (Droplet Digital Polymerase Chain Reaction). Dette er metoder der man benytter arts-spesifikke primere, og ofte også prober, som kun vil feste seg på DNA-molekyler fra én enkelt art. Dette betyr at vi kun får et positivt signal dersom det finnes DNA fra arten vi ønsker å påvise i miljøprøven.

Utvikling av slike artsspesifikke primere og prober krever en god del uttesting. Har man først laget slike primere og validert en artsspesifikk markør er det relativt raskt og kostnadseffektivt å analysere store mengder prøver. Se imidlertid Harper m.fl. (2018) for en sammenligning mellom qPCR og DNA-metastrekkoding i kostnadseffektivitet.

Metoden vil kunne kjøres av de aller fleste DNA-laboratorier med adgang til en qPCR eller ddPCR maskin. Positive kontroller kan imidlertid være vanskelig tilgjengelig for sjeldne/rødlistede arter, eller for kryptiske arter og taksomomiske artskomplekser, men syntetiske positive kontroller kan nå bestilles fra flere produsenter. Både qPCR og ddPCR gir et estimat på kvantitet av DNA i prøven, som for eksempel DNA-mengden per liter vann eller per gram jord. Dette gjør at man teoretisk sett kan sammenligne DNA-mengden fra enkeltarter mellom ulike prøver, mellom ulike lokaliteter og mellom ulike år.

Selv om oppdagelses-sannsynligheten øker med økende mengde av en art i miljøet (Dougherty et al. 2016), er det imidlertid mange ulike biotiske og abiotiske faktorer som påvirker mengde miljø-DNA som observeres (Laurendz 2017; Strand, Rusch, et al. 2019). Bruk av DNA-mengde i miljøprøver for å si noe om mengde (abundans) eller biomasse av arten i miljøet, er derfor et område under utvikling hvor det gjenstår forskning og uttesting før det kan anses som operasjonelle alternativer i forvaltningssammenheng.

Den første miljø-DNA-studien på vannprøver ble gjort ved hjelp av qPCR, der man benyttet en arts-spesifikk genetisk markør for å påvise amerikansk oksefrosk Lithobates catesbeianus (Gentile Francesco Ficetola et al. 2008). I Norge bruker vi disse metodene blant annet til å påvise fremmede ferskvannsfisk (Fossøy et al. 2017, 2018), lakseparasitten Gyrodactylus salaris (Rusch et al. 2018; Fossøy et al. 2019), kreps Astacus astacus og krepsepest Aphanomyces astaci (Strand, Johnsen, et al. 2019; Strand et al. 2011), storsalamander Triturus cristatus og småsalamander Lissotrition vulgaris (Taugbøl et al. 2018), den patogene soppen Batrachochytrium dendrobatidis (Bd) (Wacker et al. 2019), elvemusling Margaritifera margaritifera (d'Auriac et al. 2019; Wacker et al. 2019) og vasspest Elodea canadensis (d'Auriac et al. 2019)

Når det gjelder datalagring fra slike analyser, er det relativt enkle data som produseres og som ikke krever mye lagringsplass. For qPCR og ddPCR er det dessuten allerede utviklet retningslinjer (MIQE - Minimum Information for publication of Quantitative PCR Experiments) for hva som bør rapporteres i forbindelse med slike analyser (Bustin et al. 2009; Huggett et al. 2013). Ifølge disse retningslinjene bør man ha minst 20 replikater der 19 er positive for å estimere et qPCR-resultat med 95% konfidens. Men dette er som regel for omfattende og kostbart, og de fleste studier benytter seg av minimum 3 replikater der enten alle 3 eller

minst 2 av 3 må være positive for å karakterisere en prøve som positiv. Noen laber bruker en standard på minst 8 eller 12 replikater. I tillegg peker retningslinjene på viktigheten av at data kan reproduseres.

Estimering av mengde DNA i qPCR-analyser er basert på fortynningsrekker av ekstrahert genomisk DNA fra arten det analyseres for. Informasjon om tekniske replikater og fortynningsrekker må følge resultatene fra qPCR-analyser for at disse skal være etterprøvbare og reproduserbare.

1.3 Bruk av DNA-metastrekkoding for å beskrive artssamfunn

DNA-metastrekkoding kan benyttes til å påvise flere arter samtidig, og beskrive hele eller deler av artssamfunn (Taberlet et al. 2018; Ekrem and Majaneva 2019). Metoden bygger på to store fremskritt, nemlig muligheten til å identifisere arter ved bruk av standardiserte biter av arvestoffet (DNA-strekkoding) (Hebert et al. 2003) med arts-generelle primere.

1.3.1 DNA-strekkoding og referansebibliotek

En forutsetning for å bruke DNA til å identifisere arter er at det finnes referansedatabaser med DNA-sekvenser ("strekkoder") fra kjente arter. Det vil si at individer identifisert til art på bakgrunn av utseende har vært analysert og fått sin DNA-strekkode beskrevet. Dette kan sammenlignes med bruk av DNA for å finne forbryteren i en rettssak. Har man ikke DNA fra forbryteren i et referansearkiv, kan man ikke bruke DNA til identifisering.

Oppbyggingen av et offentlig tilgjengelig internasjonalt bibliotek med DNA-strekkoder er et storstilt internasjonalt prosjekt ledet av iBOL (The International Barcode of Life Consortium) der målet er å registrere strekkodene for alle levende organismer. Prosjektet legger til rette for bruk av DNA-strekkoding i forskning og forvaltning, og genererer svært verdifulle data til bruk i taksonomisk, evolusjonær og økologisk forskning.

Ettersom det er viktig at DNA-sekvensen som benyttes til identifisering viser tilstrekkelig variasjon mellom arter, standardiseres markører for ulike organismegrupper innen prosjektet. Disse er cytochrome c oxidase underenhet I (COI) for dyr, internal transcribed spacer (ITS) for sopp, Ribulose bisphosphate carboxylase large chain (rbcL) + maturase K (matK) for planter, 18S ribosomalt RNA-gen (18S) for protister, 16S ribosomalt RNA-gen for prokaryoter, og rbcL og COI for alger.

I tillegg er det utviklet protokoller for en del alternative markører, som gir betydelig tilleggsinformasjon, og som kan være nødvendig for å skille nært beslektede arter. Noen eksempler her er ITS for karplanter og insekter og 12S ribosomalt RNA-gen (12S) for fisk. DNA-strekkodebiblioteket over norske arter er en del av den nasjonale forskningsinfrastrukturen for DNA-strekkoding, Norwegian Barcode of Life (NorBOL). NorBOL benytter Barcode of Life Data Systems (BOLD, http://www.boldsystems.org/) for tilgjengeliggjøring og vedlikehold av referansebiblioteket. BOLD er en fritt tilgjengelig database for strekkoder og informasjon om assosierte belegg (Ratnasingham and Hebert 2007).

Tekstboks 1.3.1 Sentrale begrep tilknyttet referansebibliotek

Referansebibliotek / referansedatabase
- Database som inneholder DNA-sekvenser (DNA-strekkoder) fra identifiserte individer av kjente arter eller høyere taksonomisk nivå (f. eks. slekt, familie).

Referansesekvens
- DNA-sekvens fra et individ identifisert til art eller høyere taksonomisk nivå. En DNA-referanse til arten / taksonet.

Referansemateriale / beleggsmateriale
- Fysisk individ som DNA-sekvensen, observasjonen og identifikasjonen stammer fra. Må oppbevares sikkert og være tilgjengelig for eksaminering om resultat skal være etterprøvbare.

Artshypotese
- Henviser her til artshypoteser i databasen UNITE. Genetiske grupperinger basert på genetisk likhet etter en gitt, selvvalgt, terskelgrense. Hver genetisk gruppering får et unikt alfanumerisk nummer (SH) som kan benyttes i kommunikasjon.

Barcode Index Number (BIN)
- Et online rammeverk for gruppering av DNA-strekkoder fra dyr basert på genetisk likhet og genetisk avstand til andre genetiske grupper i databasen BOLD. Hver genetisk gruppering får et unikt alfanumerisk nummer (BIN) som kan benyttes i kommunikasjon og annotasjon.

1.3.2 DNA-metastrekkoding

Med "neste generasjons sekvensering" (også kalt "high-throughput sekvensering") har forskersamfunnet fått en helt ny måte å lese DNA-koder på. Disse instrumentene kan i løpet av kort tid lese fra noen millioner til flere hundre millioner DNA-sekvenser. Ved hjelp av avanserte bioinformatiske analyser kan man så sammenligne hver av disse millionene av DNA-sekvenser med referansestrekkoder i databaser, for å avgjøre hvilken art de kommer fra. Det er denne koblingen mellom DNA-sekvenseringsteknologi og DNA-referansebibliotek som til sammen muliggjør DNA-metastrekkoding (Taberlet et al. 2018).

DNA-metastrekkoding er en veldig anvendelig metode som kan brukes til å beskrive hele artssamfunn i ulike miljøprøver. Vann- og jordprøver har vært mye brukt som miljø-DNA (Andersen et al. 2012; Epp et al. 2012; Minamoto et al. 2012; Schnell et al. 2012; Thomsen et al. 2012; Majaneva, Diserud, Eagle, Boström, et al. 2018), men DNA-metastrekkoding blir også brukt til å bestemme artssammensetning i bulkprøver av insekter og andre invertebrater (A. R. Mahon et al. 2013; Elbrecht and Leese 2017; Majaneva, Diserud, Eagle, Hajibabaei, et al. 2018), samt diett hos både planteetere og rovdyr fra avføringsprøver (Pompanon et al. 2012; Shehzad et al. 2012; Clare 2014; Kartzinel et al. 2015). I tillegg har man også brukt denne teknologien til å påvise sjeldne arter av pattedyr gjennom analyser av blodigler og samfunn av vertebrater gjennom analyser av spyfluer (Calvignac-Spencer et al. 2013).

DNA-metastrekkoding genererer store mengder data, og rådata krever mye lagringsplass. Resultatene fra slike analyser vil være avhengige av metodene man benytter til innsamling av en miljøprøve, isolering av DNA og videre analyser av prøven, samt valg av bioinformatisk plattform og analyseflyt i etterkant (Callahan et al. 2016; Alberdi et al. 2018; Majaneva, Diserud, Eagle, Boström, et al. 2018; Doi et al. 2019; Hajibabaei et al. 2019). Det er derfor viktig at tilstrekkelig informasjon om hele analyseforløpet (metadata) følger resultatene fra DNA-metastrekkoding for at disse skal være etterprøvbare og reproduserbare. Selv om ulike internasjonale initiativ (f. eks. EU COST-aksjonen DNAqua-Net) jobber med metodeutvikling og forslag til analysestandarder for DNA-metastrekkoding og bruk av miljø-DNA, er det per i dag ingen standard for hvilke metadata som bør følge analyseresultatene.

1.3.3 Metagenomikk og andre metoder

DNA-metastrekkoding er avhengig av mangfoldiggjøring av korte DNA-fragmenter med generelle *primere* (PCR) før sekvensering. Ideelt sett, vil valgte primere binde seg like lett til alle arter i en prøve slik at antallet sekvenser kan relateres til biomasse og abundans. Slik er det sjelden i virkeligheten. Derfor har såkalte PCR-frie metoder som mitokondriell metagenomikk med shotgun sekvensering, og målrettet DNA-fragment sekvensering vært foreslått som alternativer til DNA-metastrekkoding. Her vil antallet DNA-sekvenser fra en art korrelere bedre med biomasse (Bista et al. 2018; Crampton-Platt et al. 2016; Gauthier et al. 2019; Gómez-Rodríguez et al. 2015). Imidlertid vil fortsatt usikkerheter knyttet til raten ulike arter skiller ut DNA og fysisk transport av DNA i miljøet være tilstede.

Mitokondriell eller kloroplast metagenomikk krever imidlertid et betydelig referansebibliotek av hele mitokondrielle eller kloroplast genomer for å fungere effektivt, eller alternativt må prøvene sekvenseres dypere enn ellers nødvendig (Braukmann et al. 2019). Utviklingen av målrettet DNA-fragment sekvensering har kommet langt i planter, og standard probe kits er nå til salg som har blitt utviklet av det Royal Botanic Gardens Kew ledede prosjekt PAFTOL (Johnson et al. 2019).

For andre organismegrupper trengs det en god del forskning og utvikling av prober for å kunne fungere på et bredt spekter av taksa. Begge metodene er mer kostbare, og også mer krevende enn DNA-metastrekkoding rent bioinformatisk. DNA-metastrekkoding fremstår derfor som den foreløpig beste løsningen for bruk av DNA-basert identifisering av miljøprøver i stor skala.

1.4 Rapportering av data og etterprøvbarhet i miljø-DNA baserte biodiversitets-inventeringer

1.4.1 Innledning

Data fra molekylære biodiversitetsundersøkelser vokser raskt i omfang. Det er sannsynlig at et stort volum av overvåkningsdata i løpet de nærmeste årene vil komme inn i form av resultater fra biodiversitetsinventeringer med molekylære metoder, enten fra vann eller jordprøver, eller bulkprøver av organismer. En utfordring er hvordan slike resultater og data skal gjøres tilgjengelig på en måte som sikrer etterprøvbarhet av undersøkelser og interoperabilitet med eksisterende overvåkningsprogrammer utført med tradisjonell metodikk. Det presiseres at når vi i det følgende snakker om interoperabilitet av data, tenker vi på tekniske egenskaper ved dataformater slik at disse kan settes sammen med andre datakilder, ikke om det rent biologisk gir mening med sammenligninger. Det siste er et viktig og stort tema, men utenfor rammene av denne teksten. Vi vil i dette kapittelet gjøre rede for dagens praksis på området. Siden dette er et nytt felt og det ikke eksisterer gode norske eksempler, vil vi nedenfor gjøre rede for temaet i et internasjonalt perspektiv.

Det er i de siste årene blitt en mye større oppmerksomhet på at data fra vitenskapelige studier skal være åpent tilgjengelige og dokumentert på en slik måte at studier kan reproduseres og etterprøves, og at data skal være gjenbrukbare i andre studier. Prinsipper for dette er blant annet nedfelt i FAIR paradigmet (Wilkinson et al. 2016) som slår fast at data skal være "Findable, Accessible, Interoperable and Reusable". Disse prinsippene er allment anerkjent innen vitenskapelig publisering av data og legges til grunn i de fleste policydokumenter som omhandler forvaltning av forskningsdata (Pilat and Fukasaku 2007; Kunnskapsdepartementet 2017).

Det legges vekt på at både data og metadata skal være mulig å finne (findable), være nedlastbare (accessible), være på tekniske og semantiske formater som gjør at de kan settes sammen med andre datakilder (interoperable), og være utstyrt med lisenser som tillater gjenbruk og muligheter for sitering av originalkilder (reusable) - både for mennesker og maskiner. Med maskinlesbar forstår vi i denne sammenhengen data eller metadata i et format som er egnet for å leses, bearbeides og forstås av en datamaskin og dermed lett kan deles på tvers av IT-systemer. Med et økende volum av data som gjør manuell sammenstilling av data svært ressurskrevende eller rett og slett umulig, er løsninger som gjør data maskinlesbare helt nødvendig i dagens forskningslandskap.

Disse prinsippene poengteres også i nasjonale strategier som fremhever at kunnskapsproduksjon som skjer ved bruk av offentlige midler skal komme fellesskapet til gode[1]. Det er derfor viktig at data er tilgjengelig for flest mulig, både andre forskere og forvaltning og næringsliv. Det er også andre og rent vitenskapelige grunner til at

[1] Nasjonal strategi for tilgjengeliggjøring og deling av forskningsdata. Kunnskapsdepartementet, 2017. https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-tilgjengeliggjoring-og-deling-av-forskningsdata/id2582412/

forskningsdata bør tilgjengeliggjøres etter FAIR prinsippene. Dette inkluderer muligheten til å analysere resultater på nytt etter hvert som data-analysemetoder utvikler seg eller muligheten for å sammenstille kunnskap på tvers av studier (meta-analyser / datasynteser).

De fleste av argumentene for tilgjengeliggjøring av data fra forskning vil også være gyldige for data generert for overvåking eller konsekvensutredninger av offentlig forvaltning eller på oppdrag av denne. Viktigheten av å kunne sammenstille data fra ulike kilder er da også poengtert i ulike styringsdokumenter for naturforvaltningen (f.eks. Meld. St. 14 2015–2016). Argumenter mot tilgjengeliggjøring av forskningsdata etter FAIR prinsippene går i hovedsak langs to linjer. Først de praktiske og ressursmessige aspektene, hvor mangel på kunnskap i fagmiljøene, relevant infrastruktur og finansiering til denne delen av forskningsprosessen oppleves som reelle hindre. Det andre, men ikke gjensidig utelukkende argumentet, går på manglende akademisk kreditering for forskningsdata i et system hvor vitenskapelige publikasjoner er valutaen som både forskere og institusjoner blir belønnet for. Eksklusiv tilgang til forskningsdata kan således være et konkurransefortrinn både i kampen om forskningsmidler og posisjonering av karriere.

Mens det første poenget i høyeste grad er relevant også for data generert via overvåkning og konsekvensutredninger bestilt av miljøforvaltningen, kan sistnevnte argument vanskelig brukes i denne sammenhengen, i alle fall sett fra oppdragsgivers side. På siden av dette er det selvfølgelig legitime argumenter for at sensitive data skal unntas offentlighet. For miljødata vil dette i praksis gjelde presis stadfesting av lokaliteter for rødlistearter eller arealrepresentativ overvåkning. I enkelte tilfeller kan imidlertid også personvernhensyn gjøre seg gjeldende.

I forhold til data fra miljø-DNA undersøkelser kan krav om tilgjengeliggjøring av primærdata i maskinlesbar form, samt sikker oppbevaring av DNA-ekstrakter, være spesielt viktig. Dette er et felt hvor utvikling av verktøy for bioinformatisk arbeidsflyt går svært raskt og referansebiblioteker stadig utvides. Resultater i form av artsidentifikasjoner er ferskvare, og det vil derfor ofte være ønskelig, eller nødvendig, å jevnlig oppdatere disse med basis i rådata. Dette vil være lite hensiktsmessig å gjøre gjennom manuelle prosesser. Fra forvaltningens side kan det også antas at det vil være uhensiktsmessig å være bundet til opprinnelig dataleverandør ved slike oppdrag. Dette gjør at krav om lisensiering som muliggjør gjenbruk er nødvendig. Entydig lisensiering er en forutsetning for å ikke måtte inngå individuelle avtaler med dataleverandører fra gang til gang, noe som kan medføre at forvaltningen må betale flere ganger for samme datasett.

Det er i dag flere gode og standardiserte lisenser som både sikrer gjenbruk av data og kreditering av opphavs-person/institusjon. Det vanligste i bruk innenfor datadeling, både nasjonalt og internasjonalt, er CC-BY lisens (https://creativecommons.org/licenses/by/4.0/) fra "Creative Commons" familien (https://creativecommons.org/) som sikrer gjenbruk og kreditering av opphavs-person/institusjoner. Et godt eksempel og en kort innføring i bruk av lisenser med spesielt blikk på norske miljødata er gitt i Artsdatabankens datapolicy (https://www.artsdatabanken.no/Files/14213/Artsdatabankens_policy_for__pen_datadeling.pdf). Krav til god dokumentasjon av data (diskutert videre i 1.6.2-3 og 2.2) vil, i tillegg til prosedyrer og gjenbruk av data, sannsynligvis også være vesentlig i forhold til regelverksoppfølging og eventuell juridisk prøving av resultater.

En årsak til at kommunikasjon rundt dataflyt kan være forvirrende er at det ofte mangler en felles forståelse for hva data er. Det kan i denne sammenhengen være klargjørende og ha en begrepsbruk som skiller mellom primærdata og deriverte data (inkludert analyseresultater). Tekstboks 1.4.1 definerer noen av de mest sentrale begrepene nærmere.

Det understrekes at forsøket på definering av begreper innen dataforvaltning er ikke er ment som en akademisk øvelse. Det vil også være ulike oppfatninger om bruken av de ulike begrepene. Skille mellom de ulike begrepene (slik vi forstår disse her) kan ha store praktiske konsekvenser for etterprøvbarhet og gjennomsiktighet, samt muligheter for gjenbruk. For eksempel: Det har de siste 10 til 20 årene skjedd kvantesprang i muligheter for statistiske analyser og maskinell bearbeiding av data. Dette gjør at vi i dag ofte har helt andre standarder for hvor skillet mellom primærdata og deriverte data går enn bare for noen få år siden. Tidligere var det vanlig å kun forholde seg til aggregerte data (for eksempel gjennomsnitt av replikater) i statistiske analyser, mens vi i dag innen økologi og biodiversitetsforskning rutinemessig bruker variasjon mellom replikater inn i statistiske modeller.

Tekstboks 1.4.1 Sentrale begreper knyttet til data og dataforvaltning

Data

- Vi forstår data som: tall, tekst, bilde og lyd i analog eller digital form som genereres eller oppstår underveis i prosjekter, undersøkelser eller utredninger. Dette kan for eksempel være data som er generert gjennom ny analyse, sammenstilling av eksisterende data, eller helt nye data generert gjennom ny datainnsamling.

Primærdata

- Med primærdata mener vi data som kommer direkte fra kilden og som ikke har blitt manipulert eller forandret. Dette vil i sammenheng med miljø-DNA undersøkelser inkludere for eksempel; beskrivelser av innsamling av materiale (lokalitet, tidspunkt, habitat m.m.), laboratorieprotokoller og utdata (rå-data) fra sekvenserings eller qPCR/ddPCR maskiner. Primærdata brukes ofte synonymt med rådata (se under).

Kildedata

- Data som ikke genereres av det enkelte prosjekt eller undersøkelse, men som brukes som grunnlag for denne. I en miljø-DNA sammenheng vil kildedata være for eksempel et eksisterende referansebibliotek eller utviklede primere. Kildedata skal ikke være nødvendig å rapportere direkte, men bør refereres til og være tilgjengelig.

Deriverte data / avledede data

- Deriverte data (eller avledede data) omfatter både analyseresultater og visualisering (for eksempel kart), aggregeringer (sammenslåing av originalobservasjoner) og tolkninger. Artslister som genereres på bakgrunn av et sett med sekvenser fra en miljø-DNA undersøkelse vil således være deriverte data. Et skille mellom deriverte data og primærdata vil ofte ikke være krystallklart. For eksempel vil begrepet primærdata vanligvis bli brukt både om eventuelle analoge (for eksempel håndskrevne laboratorie-protokoller, feltnotater) og den digitale representasjonen av disse (for eksempel feltnotater som er lagt inn i regneark) selv om det siste ofte inkluderer kvalitetskontroll, transformeringer til standard vokabular og eventuelle tolkninger av skrift. I forbindelse med miljø-DNA-data vil ofte deriverte data innebære henvisning og tilgjengeliggjøring av kode brukt til framstilling av deriverte data fra primærdata.

Standardiserte (evt. harmonisierte) data

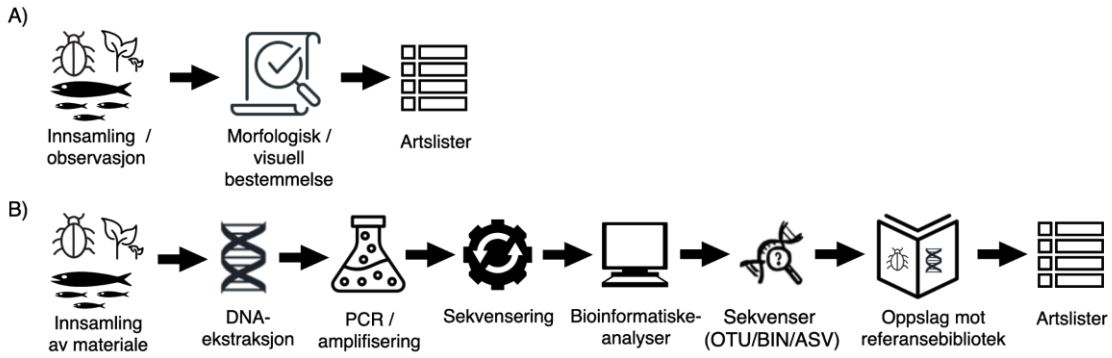
- Begrepene standardiserte / harmoniserte data kan brukes for å skille mellom de originale datakildene og data som har undergått kvalitetskontroll og transformering. Med standardiserte / harmoniserte data forstår vi primærdata (eller rådata) som har blitt kodet om eller transformert til en felles standard uten at betydningen av innholdet er endret (for eksempel felles kodeverk og felles navngivning av variabler).

Metadata

- Begrepet metadata (data om data) forstås i det følgende som en beskrivelse eller dokumentasjon av en samling (aggregering) av data på et gitt nivå (mest vanlig som et datasett). Dette inkluderer blant annet informasjon om personer og institusjoner som skal krediteres for datasettet, taksonomisk, geografisk og tidsmessig avgrensning av undersøkelsen samt en tekstuell beskrivelse av datasettet. Dataposter som koder for f.eks innsamlingsprotokoll og laboratorieprosedyrer kan i mange sammenhenger bli beskrevet som metadata, mens de i andre vil bli sett på som en integrert del av data. Det siste gjelder for eksempel når innsamlingsprotokoll blir en faktor i resultatanalyser.

1.4.2 Data fra miljø-DNA undersøkelser sammenlignet med data fra annen overvåkning

Data fra miljø-DNA studier har både ulemper og fordeler sammenlignet med "tradisjonelle" overvåkningsdata når det kommer til muligheten for tilgjengeliggjøring slik vi forstår det i en FAIR kontekst. Som utledet i kapittel 1.1-1.4, involverer artsobservasjoner via miljø-DNA mange trinn med til dels avanserte metoder, både i forhold til prøvebehandling og databehandling, før en miljøprøve resulterer i en liste over arter (Figur 1.4.2).



Figur 1.4.2: Karikert sammenligning av ulike deler av innsamlingsprosessen hvor data samles inn via A) tradisjonelt økologiske feltstudie og B) eDNA basert studie eksemplifisert ved metastrekkoding. Det presiseres at dette er en forenklet framstilling. For eDNA vil de fleste av trinnene fram til sekvensering innebære tekniske eller biologiske replikasjoner, noe som gjør strukturen på data og metadata hierarkisk. Ofte vil imidlertid studier inkludere begge typer innsamling. For eksempel: Hvis det viser seg at man i B) ved 'Oppslag mot referansebibliotek' ikke har alle artene innenfor en gitt gruppe organismer er det nødvendig å gå tilbake til A). Det kan også være at 'Oppslag mot referansebibliotek' gir overraskende resultater og nærmere undersøkelser med tradisjonell metodikk er nødvendig for og be- eller avkrefte om den resulterende 'art' identifisert vha bioinformatisk analyse kan være riktig.

Det vil her også være nødvendig med både biologiske og tekniske replikater av ulike deler av prosessen for å sikre resultater av en viss soliditet. Siden tolkning av resultater fra et miljø-DNA studie må skje i lys av hvilke metoder som er brukt på de ulike delene av prosessen, er det essensielt at de ulike metode-trinnene er godt dokumentert i data og metadata.

Dokumentasjon av både tekniske og biologiske replikater er også helt nødvendig for å kunne analysere miljø-DNA data med hensiktsmessige statistiske metoder og å få en pålitelig tolkning av resultater fra slike undersøkelser (Strickland and Roberts 2019; Schmidt et al. 2013). Biologiske replikater betyr i denne sammenhengen flere prøvetakinger av samme substrat eller organisme (for eksempel replikasjoner av vannprøver). Med tekniske replikater mener vi i replikering av ulike trinn i laboratorie-protokollen. Biologiske replikater skiller ikke miljø-DNA data prinsipielt fra data samlet med tradisjonelle metoder. På grunn av at man så godt som alltid baserer resultater på et lite utvalg fra den virkelige verden, og innsamling eller observasjoner er gjenstand for tilfeldig eller systematisk variasjon, er gjentatt prøvetaking nødvendig for å sikre representative resultater.

I miljø-DNA studier er det også tilfeldig variasjon i forhold til PCR prosesser og delprøver av materiale. Tekniske replikater og dokumentasjon av disse blir derfor spesielt viktig her (Alberdi et al. 2018; Gentile F. Ficetola et al. 2015; Taberlet et al. 2018). De mange ulike

trinnene i datahøstingen og viktigheten av dokumentasjon av disse gjør nødvendigvis datastrukturer i miljø-DNA data noe mer komplekse enn mange andre miljødatatyper.

1.4.3 Eksisterende infrastruktur og datastandarder

Data fra miljø-DNA undersøkelser har også fordeler sett i forhold til mange andre typer overvåkningsdata når det kommer til formidling og gjenbruk. Resultater fra sekvenseringsmaskiner er det vi kan kalle maskingenererte data ("digitalt fødte data"). Dette gjør at de er godt egnet for standardisering. Molekylærbiologi er også et relativt moderne fagfelt som har informasjonsteknologi sentralt i arbeidsflyten. Faggrupper som anvender denne metodikken er derfor nødt til å inneha sterk digital kompetanse, eller samarbeide tett med miljøer som har dette. Innenfor fagfeltet er det også tradisjoner for deling og offentliggjøring av primærdata i en helt annen grad enn for eksempel innen de tradisjonelle økologiske disipliner.

Det eksisterer veletablerte internasjonale strukturer for deling av primærdata / rådata fra sekvensering (INSDC, http://www.insdc.org/). Dette inkluderer åpne databaser som European Nucleotide Archive (ENA), Genbank (NCBI) og DNA Data Bank of Japan (DDBJ). Arkivering av primærdata fra sekvensering i internasjonale arkiver er også standard ved vitenskapelig publisering av molekylærbiologiske forskningsresultater og kreves i dag av de aller fleste vitenskapelige tidsskrifter. Det eksisterer etablerte metadata-standarder for å beskrive denne type data. Infrastruktur for lagring og tilgjengeliggjøring av bearbeidede sekvenser (deriverte data) er imidlertid langt dårligere utbygd.

Selv om det er helt avgjørende å tilgjengeliggjøre rådata fra sekvensering for å kunne gjenta den bioinformatiske arbeidsflyten, enten formålet er replikering eller reanalyse med oppdatert metodikk, er det først etter at sekvenser har blitt knyttet til kjente taksa at data blir forståelige i en økologisk/miljømessig sammenheng. Sekvensdata som gjøres offentlig tilgjengelig gjennom førnevnte kanaler (INSDC databaser), har imidlertid vanligvis store mangler i forhold til metadata og miljødata som er nødvendig for å tolke resultater og gjenbruke data i en biodiversitetskontekst.

Selv om det i dag ikke er noen omforent internasjonal standard for tilgjengeliggjøring av molekylære biodiversitetsdata, er det et stort internasjonalt fokus på området. Flere initiativer har til dels operasjonelle systemer for tilgjengeliggjøring av miljø-DNA data som kobler nødvendig metadata og miljø-data sammen. Dette inkluderer for eksempel mARS (Microbial Antarctic Resource System, http://mars.biodiversity.aq/) som er en belgisk initiert infrastruktur (Sweetlove et al. 2019) og Ecobiomics (Environmental metagenomic biomonitoring) som er et kanadisk prosjekt (Macklin, Baird, and Newton 2019; Edge et al. 2019). Begge disse jobber med databaser og nettplattformer basert på åpne internasjonale standarder. GLOMICON (The Global Omics Observatory Network, https://glomicon.org/) nettverket (Buttigieg et al. 2019) jobber for global implementering og operasjonalisering av standarder på området. Fellesnevnere for disse systemene og organisasjonene er et fokus på gjenbruk av standarder fra tilgrensende felter (biodiversitets- og sekvensdata). Generelt sett dominerer standarder brukt av GBIF (Global Biodiversity Information Facility, https://www.gbif.org/) (Schigel et al. 2019), sammen med standarder fra Genomic Standards Consortium, GSC, (Field et al. 2011) som plattform.

Som et eksempel deler mARS informasjon om arktisk mikrobiell biodiversitet gjennom programvare utviklet av GBIF, kalt Integrated Publishing Toolkit, IPT (Robertson et al. 2014) og tilleggsmoduler basert på GSC standarder[2]. Her blir biodiversitetsrelaterte data og metadata lenket opp mot sekvensdata lastet opp til databaser knyttet til INSDC.

Et annet og nærliggende eksempel er Swedish Biodiversity Data Infrastructure (SBDI[3]). SBDI er en sammenslåing av to infrastrukturer for svensk biodiversitetsdata, Swedish LifeWatch og Biodiversity Atlas Sweden (https://bioatlas.se/). SBDI jobber med å bygge opp en dataflyt for molekylære biodiversitetsdata og tilgjengeliggjøre informasjonen gjennom en nasjonalt samlende portal for biodiversitetsinformasjon, basert på åpen kildekode-teknologi fra Atlas of Living Australia (ALA, https://www.ala.org.au/). SBDI er en del av et globalt nettverk av land og institusjoner som benytter denne plattformen for tematisk, regional eller nasjonal tilgjengeliggjøring av biodiversitetsdata (Living Atlases, https://living-atlases.gbif.org/) og som baserer seg på standarder og data utviklet i tett samarbeid med GBIF.

Det er også initiativer som arbeider med å implementere denne plattformen for norske biodiversitetsdata gjennom Living Norway Ecological data network (LivingNorway, https://livingnorway.no/). ALA plattformen støtter i dag data fra miljø-DNA undersøkelser i sin dataflyt og tilgjengeliggjør dette sammen med data fra andre typer innsamlingsprotokoller.

GBIF er en global, mellomstatlig organisasjon og infrastruktur for biomangfoldsinformasjon etablert etter råd fra OECD's megascience forum (OECD 1999). Norge har deltatt i og finansiert infrastrukturen siden 2004 via Kunnskapsdepartementet. Den norske GBIF-noden er per i dag lokalisert ved Naturhistorisk Museum, UiO. GBIF's oppdrag er å gjøre globale biodiversitetsdata tilgjengelig gjennom en felles plattform for søk og nedlastning via både online/web-baserte brukergrensesnitt og maskin-til-maskin systemer (APIer). Systemet tilgjengeliggjør i dag data fra over 1,3 milliard naturhistoriske objekter og observasjoner av arter fra hele verden, inkludert biodiversitetsobservasjoner basert på miljø-DNA. Dette omfatter både storskala strømming av data og oversetting av data fra molekylærbiologiske infrastrukturer som ENA tilknyttede MGnify[4], eller datasett publisert fra enkeltstående prosjekter eller forskningsgrupper (Telfer 2019; Frøslev and Ejrnæs 2018).
GBIF har inkludert artshypoteser og BINs fra både UNITE og BOLD (PlutoF 2018; The International Barcode of Life Consortium 2016) i den taksonomiske indeksen[5] (Schigel et al. 2019). Dette betyr at molekylært definerte arter er søkbare og gjenfinnbare sammen med observasjoner basert på tradisjonell Linneisk taksonomi. Alternativt kan sekvenser som det (enda) ikke er mulig å knytte til kjent taksonomi tilgjengeliggjøres som artshypoteser dersom disse finnes i referansebibliotekene. Som en global dataintegreringsplattform, tilgjengeliggjør GBIF bare en liten del av informasjonen fra miljø-DNA datasett gjennom sine søkefunksjoner per i dag. Fokus er på identifiserte taksa, sted og tid for innsamling og annen ekstra informasjon er tilgjengelig i kildedata (se kapittel 2.2).

Det er forøvrig ofte mulig å hente flere detaljer enn GBIF-portalen indekserer og leverer per i dag, direkte fra de transformerte/standardiserte original-datasettene som ligger åpent

[2] https://press3.mcs.anl.gov/gensc/mixs/
[3] https://bioatlas.se/about/
[4] https://www.gbif.org/publisher/ab733144-7043-4e88-bd4f-fca7bf858880, https://www.ebi.ac.uk/metagenomics/
[5] https://www.gbif.org/news/2LrgV5t3ZuGeU2WIymSEuk/adding-sequence-based-identifiers-to-backbone-taxonomy-reveals-dark-taxa-fungi

tilgjengelige i IPT-programvaren. GBIF tilbyr dessuten allerede noen enkle verktøy for å søke med sekvensdata som søkekriterie, etter artshypoteser og eventuelle koblede Linneiske navn, for sopp (ITS) og dyr (COI)[6].

1.5 Om biobanker og langtidsoppbevaring av prøver

Tradisjonelle biobanker sørger for lagring og tilgjengeliggjøring av DNA- og vevsprøver fra enkeltorganismer. I tillegg befinner det seg mye frosset vev og DNA-prøver i frysere hos de biologiske forskningsmiljøene ved universitetene og i instituttsektoren. I instituttsektoren er det flere biobanker med delvis tilgjengelig materiale, f.eks. Havforskningsinstituttets biobank for marine prøver, Marbank. Miljøprøvebanken, som eies av Klima- og miljødepartementet og styres av Miljødirektoratet, oppbevarer hovedsakelig tidsserier av vevsprøver fra et utvalg dyr og planter, men også noen abiotiske prøver (slam og luft). Universitetsmuseene har i dag ansvar for varig lagring, kuratering og tilgjengeliggjøring av biodiversitetmateriale gjennom sitt samfunnsoppdrag.

Samlinger av DNA- og vevsprøver fra norsk biodiversitet finnes per i dag ved de naturhistoriske museene. For eksempel inneholder DNA-banken ved Naturhistorisk museum i Oslo i dag mer enn 300 000 prøver. Her lagres DNA-ekstrakter og vevsprøver fra alle organismetyper i -80°C frysere, mens metadata og informasjon om prøvenes plassering i fryserne databaseføres og tilgjengeliggjøres gjennom egne søkbare internettportaler (Collection Explorer for fugl og karplanter), GBIF og dataportalen til Global Genome Biodiversity Network (GGBN). DNA- og vevsprøver ved andre forskningsinstitusjoner i Norge er også betydelige i omfang, men ikke like godt organisert eller tilgjengeliggjort. Det ble nylig (2018) sendt en søknad til Norges Forskningsråds infrastrukturprogram, med formål å bygge opp en nasjonal, distribuert infrastruktur for biodiversitetsgenomikk (Norwegian Infrastructure for Biodiversity Genomics, NIBIGEN) (ikke tilslag). Denne inkluderte blant annet utvikling av en omforent måte å lagre og tilgjengeliggjøre DNA- og vevsprøver for biodiversitetsforskning på i Norge.

Den finnes ingen europeisk standard (CEN) for oppbevaring av vevsprøver og/eller miljø-DNA prøver, men International Society for Biological and Environmental Repositories (ISBER) publiserer jevnlig oppdaterte og relativt detaljerte protokoller for “best practices” for alle aspekter relatert til drift av biobanker. I denne gis det også anbefalinger til f. eks. drift, oppbevaringstemperaturer og prosedyrer for opptining/nedfrysing (Campbell et al. 2018). Generelt kan en si at langtidslagring av DNA og vevsprøver er best ved så lave temperaturer som mulig (-196°C), men at det av økonomiske og praktiske hensyn benyttes -80°C frysere for langtidslagring av DNA- og vevsprøver.

[6] https://www.gbif.org/tools/sequence-id

2. Løsninger for lagring av miljø-DNA prøver, tilgjengeliggjøring av data og henvisning til referansemateriale

2.1 Krav til referansedatabaser

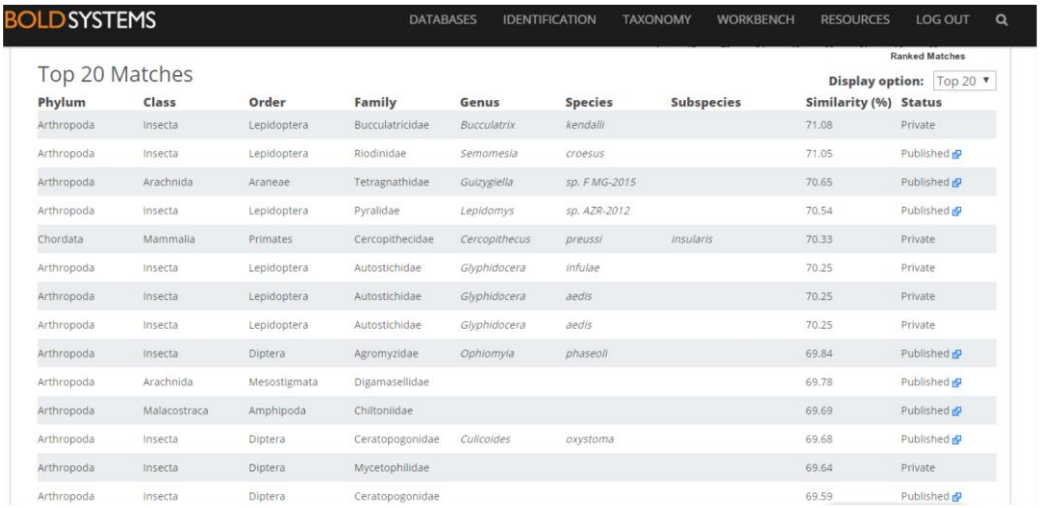
Referansedatabasenes kvalitet er avgjørende for korrekt identifisering med DNA-strekkoding og DNA-metastrekkoding. For bruk av miljø-DNA i biologisk overvåkning er det derfor avgjørende at benyttede referansebibliotek, eller deler av disse, blir tilstrekkelig beskrevet og gjort offentlig tilgjengelig gjennom permanente lenker (f. eks. DOI). På denne måten vil resultatene fra analysene være reproduserbare og etterprøvbare. Det finnes ingen formelle europeiske standarder for referansedatabaser i dag, men en teknisk rapport utviklet for kiselalger (CEN 2018) beskriver en rekke krav og anbefalinger som også gjelder andre organismegrupper.

Gode referansedatabaser kjennetegnes ved en rekke viktige kvalitetskriterier, både tilknyttet organismen som ga opphav til referansesekvensen og til den molekylærbiologiske analysen av organismens DNA. For eksempel er det viktig at en referansesekvens er tilknyttet et fysisk eksemplar - også kalt belegg (Pleijel et al. 2008; Puillandre et al. 2012; Chakrabarty et al. 2013; DiBattista et al. 2017; Ward, Hanner, and Hebert 2009). Uten belegg og tilstrekkelig dokumentasjon om dette, er det umulig å etterprøve eller revidere identifiseringer, og dermed navnet, som følger en referansesekvens i databasen. Mulighet for revisjoner er vesentlig siden endringer i en referansedatabase forventes over tid; både på grunn av feilidentifiseringer og på grunn av at taksonomiske revisjoner endrer både klassifikasjon og arters avgrensninger.

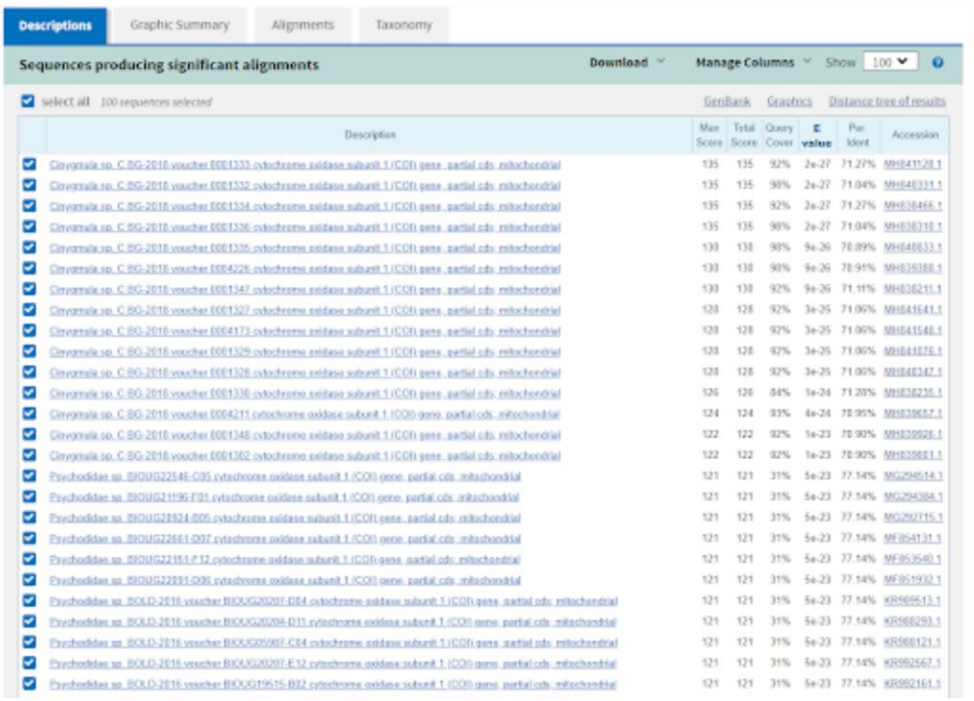
Belegget bør derfor være tilgjengelig for taksonomiske eksperter slik at revisjon av klassifisering er mulig og framtidig bevaring bør være sikret mot forringelse og uforutsette hendelser. Også identifikatorens navn, dato for identifisering og alle innsamlingsdata (sted, dato, innsamler) må være tilgjengeliggjort og følge eksemplaret. I praksis er det i dag de naturhistoriske samlingene ved universitetsmuseene i Bergen, Oslo, Trondheim og Tromsø som tilfredsstiller krav om sikring, kuratering, tilgjengeliggjøring og databaseregistrering av fysiske samlingsobjekter i Norge (St.meld.nr. 15, 2007-2008, Tingenes tale).

Svært mange av organismegruppene er dårlig representert i referansedatabaser (Kvist 2013; Weigand et al. 2019). Identifikasjoner til et høyere takson-nivå kan derfor sjeldent per i dag erstatte manglende treff på artsnivå ved søk i en database (se avsnitt 2.2.5). I mange grupper av invertebrater er genetisk avstand mellom nært beslektede arter like stor som avstanden mellom en art og et fylum (Figur 2.1).

A)



B)



Figur 2.1: A) Skjermdump fra søk med BOLDs identifiseringsmaskin med en sekvens av en isopodart som ikke er representert I database. B) Skjermdump fra blast-søk i Genbank med samme isopodsekvens som i (A). Nærmeste treff er døgnfluer og sommefuglfluer. Til forskjell fra BOLD, ser vi her også dekningsgrad for sekvensen, som kan være et viktig poeng å referere I dokumentasjon av identifikasjoner

Ved ukritisk bruk av databaser forekommer derfor at praktiserende i feltet miljø-DNA mener de har påvist tilstedeværelse av arter i et miljø der de ellers ikke har kunnet observere arten på annen måte (- sågar som «ny art for Norge»!). Det er derfor svært viktig å tydelig dokumentere DNA-identifikasjoner med presis referanse til referansesekvensen, angi likhet mellom sekvens og referansesekvens, samt å stille krav til brukere av referansedatabaser i forhold til organismekunnskap. I tillegg er det for mange arter med redusert spredningsevne viktig at regionale populasjoner er representert i databasen ettersom det kan være betydelige genetiske forskjeller mellom geografisk adskilte populasjoner. Uten en slik

representasjon, kan det være vanskelig å identifisere korrekt art med DNA-strekkoding (Bergsten et al. 2012).

Referansebibliotek er altså i kontinuerlig utvikling, og noen (slik som BOLD) er samtidig en arbeidsbenk for utvikling av strekkode-prosjekter på ulike organismegrupper. Samtidig som også mange invertebratgrupper er under taksonomisk revisjon (kryptiske arter, synonymer etc.). Dette gjør at tidspunktet for sammenligning vil være avgjørende for tolkning av analyseresultatet, og det er derfor en viktig del av dokumentasjonen som følger resultatene. I tillegg vil dokumentasjon av artsavgrensninger, taksonomiske konsepter og bestemmelseslitteratur benyttet også være viktig, både for *de novo* sekvensering av DNA fra vevsprøver, men også for kryptiske arter i miljø-DNA-prøver (Meier 2017).

I mange sammenhenger vil det også være nødvendig å kjøre jevnlig oppdatering av sammenligningen med referansebiblioteket, spesielt om prøver innsamlet og analysert ved ulike tidspunkt skal sammenlignes, for eksempel ved tidsserieanalyser eller "før og etter tiltak"-undersøkelser. DNA-baserte identifikasjoner i miljøundersøkelser bør derfor alltid utføres i tett samarbeid med økologisk / taksonomisk ekspertise på gjeldende system / artsgrupper.

2.2 Tilgjengeliggjøring og lagring av data og interoperabilitet med eksisterende forvaltningsløsninger

2.2.1 Eksisterende forvaltningssystemer

De fleste portaler, databaser eller kartløsninger miljøforvaltningen bruker som redskaper (fagsystemer / forvaltningssystemer) for forvaltning av norske biodiversitet, hverken lagrer eller formidler primærdata i den originale betydningen. Dette er uproblematisk så lenge funksjonen til systemet er definert tydelig til å være aggregering og/eller tilgjengeliggjøring av standardiserte eller deriverte data, og ikke lagring av rådata / primærdata. Det kan imidlertid være problematisk at rådata / primærdata ofte ikke blir koblet til data formidlet via fagsystemer og tilgjengeliggjort. Dette er spesielt viktig i forbindelse med miljø-DNA data, hvor det som før nevnt, sannsynligvis vil være nødvendig med tilgang til rådata for reanalyse og oppdateringer i forbindelse med bruk av data i senere sammenhenger.

To eksempler på fagsystemer for biodiversitetsinformasjon hvor data fra undersøkelser basert på miljø-DNA umiddelbart vil være viktige, er Artskart og Vannmiljø. Artskart leverer artsforekomster til norsk offentlighet og forvaltning. Tjenesten er organisert slik at dataeiere gjør sine data tilgjengelig for alle gjennom Darwin Core standarden på programvaren IPT(Integrated Publishing Toolkit; se kapittel 1.4). Dette kan enten være gjennom installasjoner av IPT på egne servere eller gjennom vertstjenester driftet av Artsdatabanken eller GBIF-Norge. For institusjoner uten ønske eller mulighet for å drifte egne systemer for databehandling, og for privatpersoner, har Artsdatabanken utviklet Artsobservasjoner

(https://www.artsobservasjoner.no) som er en tjeneste for høsting, lagring og formidling av primærdata. Data fra Artskart gjøres tilgjengelig både gjennom et grafisk brukergrensesnitt (nettportal og tilhørende nedlastningsløsninger) og et programerbart grensesnitt (API). Det er verdt å merke seg at denne løsningen sikrer at alle data parallelt blir gjort tilgjengelig og søkbar i internasjonale systemer for overnasjonale synteser eller forskning (GBIF).

Vannmiljø er miljøforvaltningens fagsystem for lagring og analyse av data om miljøtilstand i vann. Dette omfatter store mengder biodiversitetsrelaterte data, men også data på fysisk og kjemisk tilstand. Kodeverket for å beskrive data i Vannmiljø er for biologiske parametre i stor grad basert på Darwin Core. Data lastes inn i Vannmiljø av brukere (offentlige institusjoner eller oppdragstakere) ved at data legges inn i en importmal (regneark). Vannmiljø følger et kontrollert vokabular for registrering av informasjon som prøvetakingsmetode, analysemetode, prøvetakningsmedium etc. Dataposter i Vannmiljø kodes opp mot vannlokalitetsregister og stasjonsregister for innsamlingspunkt. Standarder for beskrivelse av data i vannmiljø ligger tilgjengelig på https://vannmiljokoder.miljodirektoratet.no/.

Vannmiljø har i dag en dobbeltrolle både som datalagringssenter, dataformidlingsportal og portal for analyse. Data fra Vannmiljø blir videre levert til Artskart. Imidlertid blir de samme datapunktene (artsfunn) ofte registrert direkte i Artskart og GBIF via institusjoners egne datastrømmer. Dette siden Vannmiljø har en begrenset beskrivelse av dataposter, og enkelte institusjoner ønsker å tilgjengeliggjøre en større dybde av informasjon. Siden Vannmiljø setter en egen ID på hver enkelt datapost blir dobbeltregistreringer i videre datastrøm noe utfordrende å nøste opp i.

2.2.2 Eksempel på mulig dataflyt for miljø-DNA data

I tillegg til de før nevnte utfordringene som ligger iboende i dokumentasjon av komplekse datastrukturer som miljø-DNA data, er det en utfordring at resultater skal brukes av ulike fagsystemer. De ulike fagsystemene bruker igjen til dels ulike standarder og kodeverk. Siden de samme dataene ofte vil brukes av flere ulike eksisterende eller framtidige fagsystem for miljøinformasjon, kan en mulighet være å sette en felles minstestandard for lagring og tilgjengeliggjøring for miljø-DNA data bestilt av miljøforvaltningen generelt. Men det bør da settes spesielle krav til data man vet vil bli brukt av spesifikke systemer. Et slikt eksempel er illustrert i figur 2.2.3., side 28. Merk at det i sammenheng med miljø-DNA også er spesielt viktig med tilgjengeliggjøring av kode for arbeidsflyt.

Prinsippet ligger i en felles standard for datautveksling hvor standardiserte maskinlesbare data leses av ulike fagsystemer, enten med eller uten transformering. Dette brukes allerede i dag av Artskart, hvor data og metadata tilgjengeliggjøres i et maskinlesbart utvekslingsformat (DwC-A, Darwin Core Archive (Wieczorek et al. 2012)) gjennom en dedikert programvare (IPT (Robertson et al. 2014)). Den ligger enten installert på den enkelte institusjons tjenere eller via vertstjenester, og leses inn på Artskart. Siden Artskart baserer seg på GBIFs dataflyt, blir tjenesten fullt ut interoperabel med data levert av ca. 2000 institusjoner verden over. Tjenesten kan da også presentere data om norsk biodiversitet publisert av institusjoner i andre land. I Artskart sitt tilfelle foregår kvalitetskontroll ved innlesing av standardiserte data, men det er også støttetjenester som dataeiere kan bruke for kvalitetskontroll av data ved opplastning til standardisert format.

En datautvekslingspakke med standardiserte data vil inneholde en blanding av primærdata og deriverte data, hvor det vanligvis vil være deriverte data (f.eks. artslister) som er av umiddelbar interesse for forvaltningen i forhold til formålet med kunnskapsinnhenting. Deriverte data, som artslister, vil imidlertid ha blitt generert på bakgrunn av en rekke steg som inneholder subjektive valg og filtreringer den enkelte oppdragstaker må gjøre i forbindelse med for eksempel ved valg av primere og innstillinger i bioinformatisk arbeidsflyt. For metastrekkodingsdata betyr dette at artslister vanligvis kun vil være gyldige innenfor et definert taksonomisk og geografisk fokusområde. Det vil derfor sannsynligvis være mest hensiktsmessig å begrense innholdet i en datautvekslingspakke til treff i referansedatabasen innenfor de relevante taksonomiske grupper oppdragstaker kan gå god for.

Det er imidlertid viktig at dette fokusområde blir klart definert (kodet inn i data) og at valgene er etterprøvbare. Det vil være helt vesentlig for gjenbruk og gjennomsiktighet til resultater at rådata (protokoller, bioinformatiske arbeidsflyt, resultater fra sekvensering m.m.) blir tilgjengeliggjort. Da godt etablert infrastruktur for slike data eksisterer (e.g. GenBank, ENA, eventuelt generalistarkiver for ustrukturerte data), og disse i stor grad er kjent og i bruk av fagmiljøene i dag, er sannsynligvis bruk av slike det mest hensiktsmessige sett fra et kostnadsperspektiv. Rådata må lenkes opp til standardiserte data gjennom permanente maskinlesbare nettlenker (f.eks. DOI) for sikre maskin-til-maskin kommunikasjon i framtidige systemer. Dette prinsippet, eller variasjoner av det, er allerede benyttet i drift av ulike infrastrukturer som håndterer molekylære biodiversitetsdata fra miljø-DNA prøver (se kapittel 1.4.3).

I forbindelse med kravspesifikasjoner stilt til leverandører av tjenester bestilt av Miljødirektoratet, må imidlertid kostnadseffektivitet veies opp mot forvaltningens eventuelle behov eller krav om å lagre data hos nasjonale vertstjenester.

2.2.3 Forslag til felles plattform for deling av miljø-DNA data basert på eksisterende løsninger, datastandarder og infrastrukturer

Følgende prinsipper bør legges til grunn for utvikling av minstekrav til innrapportering og offentlig tilgjengeliggjøring av DNA-sekvenser og tilhørende *metadata* i miljøovervåkningen:

1. Data og metadata skal kunne leveres til en eller flere forvaltningssystemer som miljøforvaltningen i dag administrerer. Miljødirektoratet har i dag en rekke databaser og karttjenester, og resultater fra miljø-DNA vil sannsynligvis framover inngå i flere av disse. I mange sammenhenger vil samme datapost presenteres via flere systemer. Det skal tas høyde for framtidige utviklingsmuligheter.
2. Data og metadata skal tilgjengeliggjøres på en måte som muliggjør: i) etterprøvbarhet av resultater fra undersøkelsen; ii) gjenbruk i forhold til re-analyse ved for eksempel oppdatering av referansebibliotek eller oppfølgingsstudier. Gjenbruk og etterprøvbarhet skal være mulig uavhengig av oppdragsinstitusjon.
3. Standarder som legges til grunn for beskrivelse og struktur av data og metadata, og løsninger for formidling av data, skal i størst mulig grad gjenbrukes fra eksisterende åpne standarder og løsninger.

Vi beskriver her et forslag basert på gjenbruk av eksisterende infrastruktur. Med hensyn på kostnadseffektivitet og interoperabilitet med andre datakilder, nasjonalt og internasjonalt,

skal det svært tungtveiende grunner til før man kan anbefale bygging av nye systemer fra grunnen av. Så lenge systemet baserer seg på åpne standarder og programvare med åpen kildekode, vil det heller ikke normalt sett være problemer forbundet med leverandøravhengighet. Vi vil i det følgende gi en oversikt over eksisterende standarder og systemer dette kan basere seg på. Det må presiseres at disse forslagene ikke nødvendigvis på alle punkter innebærer en optimal løsning på sikt, og at dagens standarder med fordel kan utvikles videre. Dette blir diskutert videre i kapittel 2.2.5. De viktigste datastandarder foreslått benyttet er beskrevet nærmere i Tekstboks 2.2.3.

Tekstboks 2.2.3 Relevante eksisterende standarder for miljø-DNA data og bruksområder

Generelle metadata

Ecological metadata language (EML)

- EML er en metadata-standard utviklet spesielt for økologiske disipliner (Jones et al. 2006). Metadata beskriver typisk informasjon om et datasett, dette inkluderer en tekstuell beskrivelse av for eksempel; tittel, metoder, opphavspersoner og kontaktpersoner, taksonomisk, geografisk og temporær utstrekning, metode etc. EML-standardisert metadata gjøres maskinlesbart gjennom bruk av datadeskriptorer som er definert av EML-standarden. EML-standaren forvaltes av The Knowledge Network for Biocomplexity (KNB) finansiert av National Science Foundation, USA. Standarden blir brukt av en lang rekke infrastrukturer og organisasjoner som håndterer biodiversitetsdata, blant annet GBIF og Artsdatabanken.

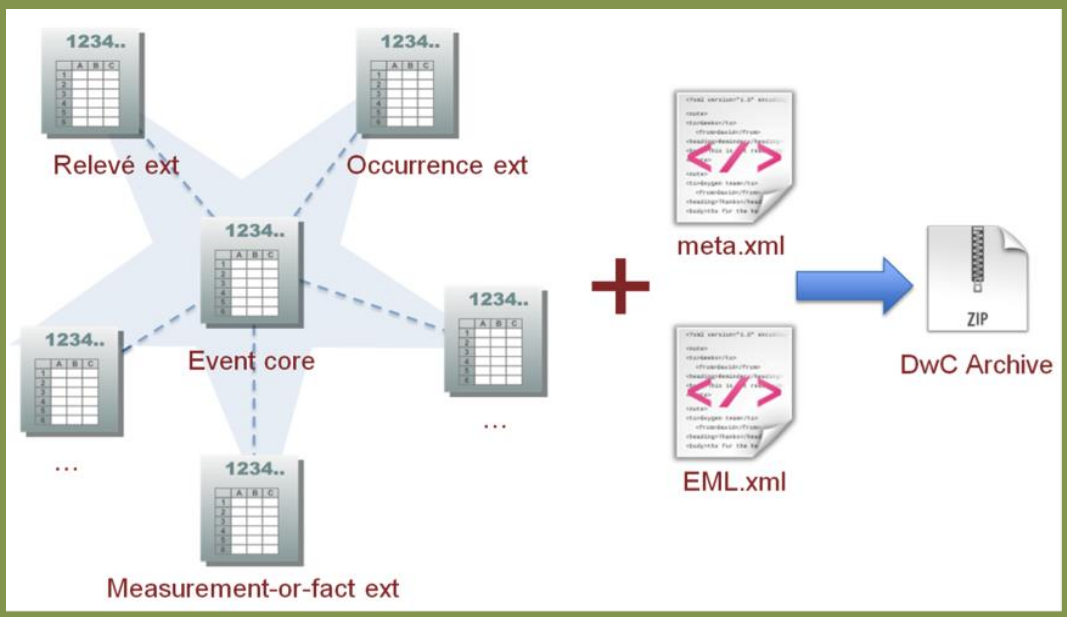
Feltinnsamling og biodiversitet

Darwin Core

- Darwin Core (DwC) er et vokabular eller en samling av termer (også i ulike sammenhenger kalt elementer, felter, attributter, konsepter) til bruk for å beskrive informasjon om biodiversitet (Wieczorek et al. 2012). DwC er primært brukt for organismer og deres tilstedeværelse i naturen og inkluderer blant annet termer for beskrivelse av geografisk lokasjon, innsamling-hendelser, kvantitet av organismer, fysiske eksemplarer og andre belegg. DwC gir ingen nærmere definisjon av datamodell eller teknisk standard for utveksling av data. De ulike termene er imidlertid organisert i klasser som korresponderer til ulike grupper med elementer av informasjon assosiert med en biodiversitets observasjon (f.eks. innsamlings-hendelse, funn, taksonomisk beskrivelse, lokasjon).
- DwC er i utgangspunktet en utvidelse av Dublin Core (en mye brukt standard for å assosiere metadata med, i hovedsak, digitale, ressurser) og er per i dag en stabil standard med bred anvendelse for beskrivelse av biodiversitetsdata. Darwin Core er grunnlaget for datautveksling gjennom Artskart og GBIF systemet.
- Standarden er kompatibel med INSPIRE. Standarden forvaltes av et internasjonalt nettverk, Biodiversity Information Standards (TDWG), med formell status som "non-profit organisation" i EU og USA.
- En oversikt over termer i DwC finnes på https://dwc.tdwg.org/terms/.

Darwin Core archive

- Vanligvis brukes DwC til datadeling i sammenheng med et standardisert filformat (Darwin core archive, DwC-A). Dette er en kompakt datautvekslingspakke i form av en .zip fil som inneholder koblede filer med data i tabular form (.txt format). I tillegg til en kjerne-fil som inneholder basisinformasjon kan det legges på en eller flere tilleggstabeller (med datadeskriptorer fra DwC-standarden, eller andre terminologi standarder) som beskriver andre deler av data. I tillegg inneholder et DwC-A to elementer i form av .xml skjemaer for å gjøre innholdet lesbart for både mennesker og maskiner; en metadatafil (EML.xml) og en fil som beskriver sammenhengen mellom filene i arkivet (meta.xml) (Figur tekstboks_2.2.3).



Figur tekstboks_2.2.3: Et DwC-A er en .zip fil hvor data er kodet i et tabulert format og lenket sammen med en kjerne (core)-fil, og ulike utvidelser, en metadata-fil som beskriver datasettet i EML (EML.xlm) og en beskrivelse av sammenhengen mellom de ulike filene (meta.xlm). Figur: Global Biodiversity Information Facility, CC-BY.

Darwin Core utvidelser (extensions)

- Utvidelser (“extentions”) er en benevnelse som brukes om data-tabeller som lenkes opp til kjernefilen i et DwC-A. Disse kan basere seg på termer som enten er en del av DwC-vokabularet, eller på andre eksterne standarder. GBIF forvalter i dag et sett med tillegg til DwC, inkludert tillegg for å beskrive genetiske ressurser (se nedenfor).

Standarder for ekstraksjon, amplifisering, sekvensering og fysiske prøver

MIxS (Yilmaz et al. 2011)

- *Minimum Information about any (x) Sequence (MIxS)* er en standard utviklet av Genomic Standards Consortium (GSC, http://gensc.org/) med formål å utvikle en felles standard for informasjon om alle typer sekvens data fra alle taksonomiske domener. Dette inkluderer også termer for miljøbeskrivelser.
- Oversikt over termer finnes på http://gensc.org/gc_wiki/index.php/MIxS
- MIxS er kompatibel med *DwC* og tilgjengelig som tillegg (http://rs.gbif.org/sandbox/extension/mixs_sample_2019_10_04.xml).

GGBN *datastandard (Droege et al. 2016)*

- GGBN-datastandarden er utviklet av Global Genome Biodiversity Network (GGBN, http://www.ggbn.org/), et globalt nettverk av vevsprøvebanker, med formål å fasilitere utveksling av informasjon som lenker fysiske prøver og resulterende molekylære sekvensdata. *GGBN*-datastandaren komplementerer eksisterende standarder som DwC, MIxS, og EML, og er motivert ut fra et behov om utveksling av informasjon om genetiske ressurser for og oppfylle blant annet Nagoya protokollen. GGBN er kompatibel med alle termer i MIxS, samt datastandararer relatert til humane vevsprøver og vevsprøvebanker som SPREC og BRISQ (Droege et al. 2016; Moore et al. 2013; Lehmann et al. 2012).
- Oversikt over ulike termer finnes på https://terms.tdwg.org/wiki/GGBN_Data_Standard
- *GGBN*-vokabularet er fullt ut kompatibelt som et supplement til DwC og er tilgjengelig som tillegg (http://rs.gbif.org/extension/ggbn/).

MIQE (Bustin et al. 2009)

- *M*inimum *I*nformation for publication of *Q*uantitative PCR *E*xperiments (MIQE) er et sett med veiledninger for hvilken informasjon som bør følge tilgjengeliggjøring av resultater fra qPCR og ddPCR for å sikre transparens og reproduserbarhet. Det er teknisk sett ikke en datastandard i seg selv, i og med at det ikke inkluderer anbefalte termer eller vokabular. Det gir imidlertid en sjekkliste for informasjon som bør eller skal være gitt i forbindelse med publisering av resultater og data. Sjekklisten er kompatibel med GGBN og MIxS datastandardene

Vårt primære forslag er å adoptere prinsippene bak dataflyt som allerede i dag eksisterer for Artskart, men med tilleggskrav for rapportering av data fra miljø-DNA undersøkelser (Figur 2.2.3, side 28). Forslaget baserer seg på gjenbruk prinsippene bak dataflyt for Artskart hvor et Darwin-Core arkiv blir benyttet som datautvekslingspakke, distribuert gjennom programvare fra en åpen kildekode løsning (Integrated Publishing Toolkit, IPT) som ligger installert på den enkelte institusjons tjenere eller via en vertstjener. Fordelene med dette er umiddelbart at data fra miljø-DNA undersøkelser kan flyte inn i Artskart mer eller mindre uten endringer av eksisterende system. Tilsvarende vil alle datastrømmer som allerede er kompatibel med Artskart og GBIF, umiddelbart kunne strømme inn til andre informasjonssystemer som etablerer støtte for denne standarden.

For andre forvaltningssystemer (som Vannmiljø) vil det være nødvendig med transformering av data, men så lenge det stilles krav om spesifikke tekniske standarder for innhold og koding

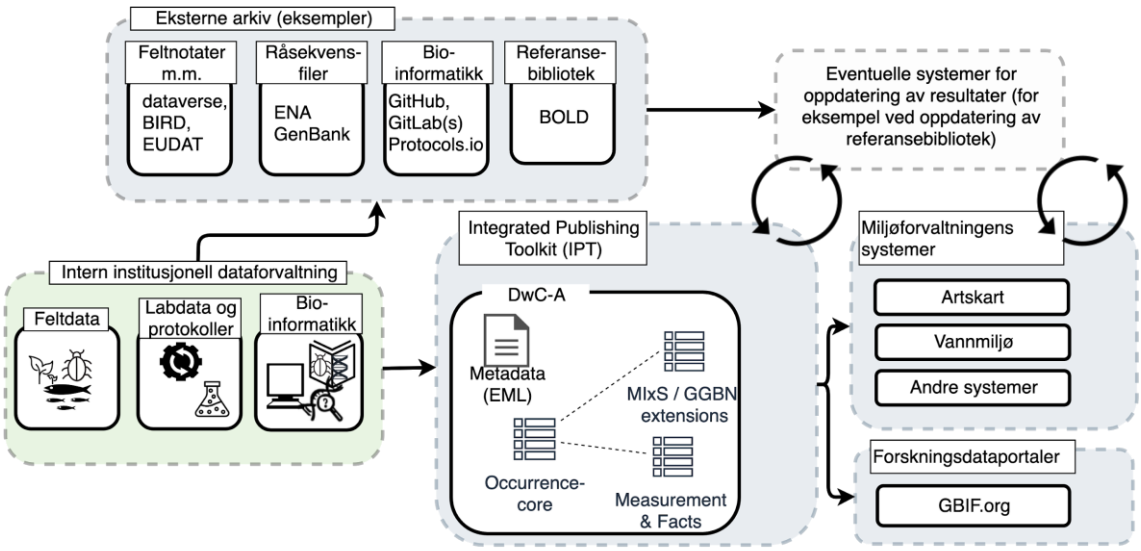
av et DwC arkiv, vurderes det at tekniske løsninger for dette være relativt enkelt å implementere. Alternativt kan det defineres egne DwC-utvidelser for spesifikke standarder (e.g. Vannmiljø). Grunnleggende termer fra DwC -standarden som i dag kreves av Artskart vil være naturlige å inkludere i eventuelle krav til dataleveranser basert på miljø-DNA, men med tillegg.

Minstestandard for slike tilleggskrav vil innebære:

- Lagring av relevante rådata i eksterne arkiver og permanent maskinlesbar referanse til disse.
- Kildedata og andre underliggende ressurser som brukes for å fremskaffe resultater må være offentlig tilgjengelig og korrekt referert. Dette inkluderer, men begrenser seg ikke, til; referansebibliotek, bioinformatisk arbeidsflyt, laboratorieprotokoller.

- Bruk av DwC-utvidelser for å fange relevante detaljer ved miljø-DNA data, samt inkludering av flere termer fra DwC som obligatoriske felter for rapportering og bruk av kontrollerte vokabular for disse (se Vedlegg 1: https://github.com/NTNU-VM/Darwinize-eDNA).

Det vil i tillegg sannsynligvis være nødvendig å stille flere særlige krav til koding og/eller innhold i data fra undersøkelser som er bestilt for bruk i spesifikke forvaltningssystemer (et eksempel for Vannmiljø diskutert under). I figur 2.2.3 gis en oversikt over et mulig system basert på dagens standarder.



Figur 2.2.3: Skisse av mulige plattform for innrapportering og offentlig tilgjengeliggjøring av DNA- sekvenser og tilhørende metadata fra oppdrag (grønn boks) basert på eksisterende systemer og datastandarder (grå bokser). Et tenkt system for jevnlig (basert på maskin-til-maskin lesing av data) oppdatering av resultater (hvit boks) kan enten lese og oppdatere Darwin Core Arkivet eller ulike forvaltningsystemer. Forflytning av data mellom de ulike elementene (svarte piler) vil i forskjellig grad kreve transformering og harmonisering av data og kan inneholde enten maskinell eller menneskelige kvalitetssikring.

Figur 2.2.3 inneholder altså som diskutert over i kapittel 2.2.2 ikke ett system, men heller et modularisert konsept bestående av systemer som "snakker sammen". Fordelen med dette er at de ulike elementene kan videreutvikles i ulik hastighet så lenge interoperabilitet ivaretas.

Det er her valgt å skissere lagring av rådata inn i eksisterende eksterne arkiver. I forhold til enkelte deler av datastrømmen (primærdata fra sekvensering), er dette både et spørsmål om kostnader knyttet til lagring og tilgjengeliggjøring av til dels store datamengder, samt hvilke systemer som er kjent og brukt i fagmiljøene. For andre datatyper (e.g. ustrukturert rådata og maskinkode) er det relativt mange ulike alternativer i forhold til arkiv-plattform og forholdsvis små datamengder. Man kan her enten stille minstekrav til plattformer, for eksempel godkjennelse gjennom internasjonale sertifiseringsordninger som CoreTrustSeal, https://www.coretrustseal.org/, eller peke på enkeltløsninger. Ansvar for datakvalitet og konformitet med data-standarder er en sentral del av integritet til resultater og bør således legges til dataleverandør (oppdragstaker). Systemer for kvalitetssjekk av data og ansvar for oppfølging overfor oppdragstaker legges til oppdragsgiver (miljøforvaltningen).

En fordel med denne løsningen er, i tillegg til at data kan gjenbrukes i ulike forvaltningsplattformer, tilgjengeliggjøring for miljøforskning. Etter modell fra samarbeidet mellom NorBOL og miljøforvaltningens institusjoner anbefales derfor på alle punktene at det etableres samarbeid og kontakt med relevante forskningsinfrastrukturer og nettverk som håndterer de ulike datatypene både i Norge (ELIXIR, GBIF, LivingNorway) og i naboland (Biodiversity Atlas Sweden, Sverige). I Norge eksisterer det allerede et tett samarbeid mellom Artsdatabanken og GBIF.

Eksempel på hvordan konkret innhold i et DwC-arkiv er kodet er gitt i vedlegg 1. Her gis en kort oppsummering over de viktigste punktene og hvordan dette skiller seg fra dagens minstekrav for levering av data til Artskart. Ikke all informasjon vil bli brukt inn i alle forvaltningssystemer, men henvisning til data i DwC arkiv vil gi mulighet for etterprøvbarhet og gjennomsiktighet i resultater.

2.2.4 Interoperabilitet med eksisterende forvaltningssystemer - Vannmiljø som eksempel

I Vannmiljø benyttes artsdata/artslister (eller taksa identifisert til gitt taksonomisk nivå) til å beregne biologiske indekser, som gir grunnlag for å klassifisere tilstand i en vannforekomst. Det betyr at prøvematerialet må være representativt for vannforekomsten i tid og rom. Enkelte indekser baseres kun på tilstedeværelse av taksa/arter, hvor hvert takson/art kan ha indeksverdier basert på toleranse/sensitivitet overfor en bestemt type forurensning (eks. forsuring). Andre indekser krever også informasjon om kvantitet av organismer, f.eks. individantall, dekningsgrad (makrofytter eller makroalger) eller størrelsesfraksjoner. Bruk av miljø-DNA inn i kvantitative indekser ligger evt. noe fram i tid og avhengig av uttesting / kalibrering (se kapittel 1.2), og vil heller ikke være mulig i mange tilfeller. En vurdering av interkalibrering med eksisterende kvalitetselementer for bruk av miljø-DNA inn i vannovervåkning ligger imidlertid utenfor omfanget av denne rapporten.

Per i dag virker det mest aktuelt og bruke miljø-DNA til generering av artslister og da for leveranse inn i indekser som baserer seg på tilstedeværelse (forekomst) data. Det vurderes at dataflyt som skissert ovenfor i kapittel 2.2.3 vil være kompatibelt med Vannmiljø, men det vil være behov for å modifisering av importregimet til Vannmiljø for å sikre at vi ivaretar krav til informasjon som skal følge med transformerte/aggregerte data fra miljø-DNA. Det er også et behov for å klargjøre hvilke felter som må gjøres obligatoriske for data som skal leveres til Vannmiljø. Innholdet i feltene må også kreves kodet etter vannmiljøstandard, eller,

kompatible standarder som transformeres i et eventuelt modifisert importregime. Dette er et eksempel på at data som kreves levert inn i spesifikke fagsystemer vil kunne ha spesifikke krav i tillegg til det som er grunnleggende nødvendig for miljø-DNA prøver generelt. Et eksempel på harmonisering av sentrale termer er gitt i vedlegg 1. Det er også felter som i dag benyttes av Vannmiljø som sannsynligvis ikke dekkes av DwC eller eksisterende utvidelser av denne. Dette gjelder for eksempel "Aktivitet_ID". Slike termer kan på kort sikt integreres i dataflyten ved å bruke generelle utvidelser som MeasurmentOrFact (https://rs.gbif.org/extension/dwc/measurements_or_facts.xml), ExtendedMeasurmentOrFacts (https://rs.gbif.org/extension/obis/extended_measurement_or_fact.xml), eller kodet inn i JSON streng og delt i DwC-feltet dynamicProperties (https://dwc.tdwg.org/terms/#dwc:dynamicProperties).

Alternativt kan det utvikles utvidelser skreddersydd for vannforvaltningsformål.

2.2.5 Utfordringer og utviklingsbehov på kort og mellomlang sikt

Det er på kort sikt noen tekniske utfordringer i forhold til beskrivelse av data og bruk av gjeldende terminologi for DwC og utvidelser som bør løses. Dette er i stor grad snakk om klargjøring og definering som bør inngå i en veileder / kravspesifikasjon til dokumentasjon av data fra miljø-DNA undersøkelser. På noe lengre sikt, og på basis av dette, er det videre muligheter for etablering av egne strukturer for fremvisning og tilgjengeliggjøring av data fra miljø-DNA spesielt. Det bør også i forhold til forvaltningens behov vurderes om ikke en offisiell norsk oversettelse av termer brukt til å beskrive data (Darwin Core felter og utvidelser) er ønskelig. Deler av dette er allerede gjort i forbindelse med Artskart og Vannmiljø, men disse oversettelsene mangler en autorativ kilde. På grunn av stadig utvikling innenfor feltet kan det være utfordrende med tekniske beskrivelser av hvordan data samles og analyseres. Se f.eks Knudsen m.fl. (2018).

En utfordring er at det i mange tilfeller ikke er en 1:1 match mellom sekvensdata og vitenskapelig navn, eller at sekvensdata ikke refererer direkte til artsnivået. En mulig løsning er en rapportering av nærmeste definerte høyere taksa i artslisten, sammen med en liste over mulige taksa som sekvensen kan gjelde i kommentarfelt o.l. Men, det må da komme frem hvor lik den ukjente er på nærmeste referansesekvens, og også foreligge dokumentasjon på hvor godt referansebiblioteket er for akkurat denne organismegruppen. Det vil også være nødvendig å rapportere DNA-sekvens direkte i datautvekslingspakke. Det kan i tillegg tenkes å være uoverensstemmelser mellom navngivning i referansebibliotek og anbefalt offisiell norsk navngivning av arter (Artsnavnebasen). Artsdatabankens Artsnavnebase, koblet mot riktig ID må ligge til grunn for navneverket som brukes i rapportering mot Norsk forvaltning. Generelt er tolkninger av resultater avhengig av god taksonomisk kompetanse og organismekunnskap. For noen organismegrupper (prokaryota og protister) mangler det dessuten vitenskapelige navn for organismer som sekvenseres, og eksisterende plattformer har også mangelfull taksonomi for disse gruppene der vitenskapelig navn eksisterer.

Det bør komme klare definisjoner på hvilke felter som skal brukes for de ulike delene av informasjonsinnholdet i et datasett, samt til innhold i feltene. Eksempel på dette er vedlagt (Vedlegg 1). Darwin Core er definert noe løst, noe som også er hensiktsmessig i forhold til å fungere som en global standard på tvers av fagområder. Dette betyr ikke at

enkeltorganisasjoner (som Miljødirektoratet) kan sette mer spesifikke krav til bruk av f.eks. kodeverk innenfor rammene av standarden. Relatert til dette må det også defineres et undergruppe av termer som skal gjøres obligatoriske, samt krav til metadata-innhold.

Stabile globale unike identifikatorer (ID) på dataposter er i liten grad brukt i dag. Dette er et problem, både med hensyn på tilgjengeliggjøring av samme datapost i forskjellige systemer og referering og sporbarhet til data. Ikke minst vanskeliggjør dette også kobling av informasjon mellom ulike datasett. Det bør derfor stilles krav til at ulike elementer av et datasett kodes opp med slike nøkler. Med ulike elementer menes her for eksempel poster som beskriver ulike hierarkiske nivåer med innsamling og laboratorieanalyse, funn av sekvenser (artsfunn), fysisk materiale (DNA ekstrakter, miljøprøver, se kapittel 2.3) etc. Dette muliggjør sporbarhet av data på tvers av systemer, også når ulike systemer har ulik praksis på aggregering. Det eksisterer flere ulike alternativer for slike globalt unike nøkler, for eksempel bruk av Universelle Unike Identifikatorer (UUID). Relatert til dette bør det også settes krav til for hvordan ulike biologiske og tekniske replikater skal kodes i standardiserte / harmoniserte data.

Basert på den ovenfor diskuterte dataflyt og standard-implementering vurderes det som realistisk å etablere en nasjonal portal for data fra miljø-DNA undersøkelser og tilhørende miljøprøver. Dette inkluderer både egne dedikerte portaler og implementering i sentrale forvaltningsverktøy. Internasjonalt er den mest utviklede løsningen i denne sammenheng det GBIF initierte Living Atlas nettverket (https://living-atlases.gbif.org/) (Lecoq et al. 2019). Living Atlas er en plattform for aggregering, visualisering, analyse og distribusjon av biodiversitetdata basert på Atlas of Living Australia (https://www.ala.org.au/), utviklet av the Commonwealth Scientific and Industrial Research Organisation (CSIRO) i Australia. Den betjener i dag kunnskapsbehovet med hensyn på biodiversitetdata til både australske forsknings-, undervisnings- og forvaltningsorganisasjoner, samt industri. Dataflyt inn til en Living Atlas portal kan enkelt konfigureres til hente alle nye datasett som publiseres i GBIF/Artskart via IPT (med miljø-DNA data fra Norge eller tilsvarende kriterier). Løsningen kan også konfigureres med tilsvarende datastrøm, fra eller til andre systemer (slik som for eksempel Vannmiljø) under forutsetning av noe IT-utvikler tid for å etablere kompatible programmeringsgrensesnitt (API) på den ene eller begge sidene.

Under LivingNorway paraplyen (https://livingnorway.no/) går nå flere sentrale norske institusjoner sammen inn i Living Atlas nettverket og etablerer en egen Living Atlas installasjon for Norge. Biodiversity Atlas Sweden jobber per i dag med nødvendige justeringer for å implementere framvisning av miljø-DNA i sin Living Atlas installasjon, og, vil i henhold til grunnprinsippene for Living Atlas Nettverket, gjøre disse tilgjengelige for andre under åpen programvarelisens. Det vurderes derfor at dette vil være en svært kostnadseffektiv plattform for å etablere en nasjonal portal for fremvisning og formidling av resultater fra miljø-DNA baserte undersøkelser, samt informasjon om fysiske prøver knyttet til dette.

Artskart er i dag det viktigste nasjonale fagsystemet for framvisning av artsdata. Plattformen med tilhørende visningsportal (https://artskart.artsdatabanken.no/) vil utvikles videre for å håndtere og vise stadig nye datatyper/format, bedre koblinger til og visning av metadata. Det er sannsynligvis bare mindre justeringer som skal til for å bedre håndtere datasett basert på miljø-DNA. Artskart baserer seg på de internasjonale standardene for biodiversitetsdata som utvikles i TDWG og benytter seg av teknologi utviklet av GBIF. Datasett basert på miljø-DNA med kjent taksonomi (i.e. DwC feltet “scientificName” fylt ut med godkjent vitenskapelig

navn) vil derfor allerede vises i Artskart i dag. Kommende nødvendige endringer vil imidlertid spesielt være knyttet til håndtering av taksonomi. Frem til nå har Artskart bare vist funn med kjent taksonomi fra Artsnavnebasen, men skal nå tilpasses for å ta med funn med ukjent taksonomi, herunder også artshypoteser. Utover Artsnavnebasen kan det gjøres oppslag mot andre internasjonale databaser som f.eks. Catalogue of Life+. Videre vil Artskart forsøke å gjøre oppslag i og forholde seg til internasjonale registre med sekvensdata – som BOLD via deres Api'er - og dermed hente ut taksonomi basert på de id'er og referanser som oppgis i publiserte datasett. Det siste vil være viktig da treff i disse databasene kan endre seg etter hvert som flere arter blir sekvensert.

Å gjøre tilpasninger i Artskart har stor betydning for mange bruksområder, da infrastrukturen rundt Artskart er tett integrert med andre nasjonale tjenester. Plattformen leverer data til bl.a. Miljødirektoratets portal for sensitive artsfunn, Naturbase, Miljøstatus og NIBIO's Skogportal via tilpasset Api. Det publiseres også data via standarder fra The Open Geospatial Consortium (OGC: https://www.opengeospatial.org/) som benyttes av bl.a. Statskog, Viken skog og lignende aktører. Plattformen er sentral i tjenester knyttet til tidlig oppdagelse og overvåking av fremmede arter der automatisk varsling går for å fange opp nye arter for Norge og spredning av fremmede arter (NINA rapport 1729 – Tidlig oppdagelse av nye fremmede arter i Norge – under publisering). Funn av rødlistede- og fremmede arter publiseres i den nasjonale infrastrukturen GeoNorge.

2.3 Lagring og tilgjengeliggjøring av prøver

Lagring av miljøprøver krever utbygging av dertil egnet infrastruktur (ISBER 2018). Miljøprøvene kan ofte være voluminøse og krever derfor store arealer i frysere eller fryserom, mens DNA-ekstrakter fra miljøprøver gjerne utgjør små volumer og kan oppbevares i små kryorør eller egnede mikroplater. Mens de to sistnevnte relativt enkelt kan inkorporeres i eksisterende biobanker, eventuelt utgjøre separate samlinger i egne frysere, vil lagring av råmaterialet kreve en helt annen infrastruktur.

I noen tilfeller (f.eks. permafrostprøver) har råmaterialet så stor potensiell verdi for videre forskning og som dokumentasjon at dette også må tas vare på i sin helhet. I andre tilfeller kan mellomprodukter i prosessen bevares for og sikre fremtidig gjenbruk. Svært mye av metodikken er imidlertid foreløpig i rask utvikling. For eksempel kan ulike ekstraksjonsmetoder gi forskjellig resultat. En mulighet i forbindelse med miljø-DNA basert på filtrerte vannprøver kan derfor være og ta ekstra replikater i forbindelse med filtrering og lagre filter der hvor prøver/resultater anses for å være av særlig betydning i et langtidsperspektiv (for eksempel tidsserie-overvåkning).

Det er et behov for omforente retningslinjer på hvordan ulike miljøprøver bør tas og oppbevares om DNA skal preserveres optimalt for gjenbruk i andre forsknings- og forvaltningsprosjekter. Per i dag finnes det ikke etablerte standarder, men enkelte er under utvikling i internasjonale konsortier (slik som DNAqua-Net), og publiserte anbefalinger finnes (f.eks. Gemeinholzer et al. 2010). Det er også behov for retningslinjer på hvilke prøver som skal lagres videre.

Selv om innsamling av prøver finansieres over for eksempel overvåkningsprosjekter, vil det også påløpe betydelige kostnader med eventuell oppbevaring av fysiske miljøprøver, samt uthenting og behandling. Miljøprøver skiller seg således fra data, hvor lagring og tilgjengeliggjøring er relativt sett billig, og gjenbruk ikke forringer råvaren. Skal det implementeres prinsipper for leverandøravhengighet for tilgang til slike prøver, er derfor grunnfinansiering for ikke bare utbygging av infrastruktur, men også drift, sannsynligvis nødvendig. Det må også stilles krav til sikring og kuratering av slike samlinger for å sikre relevans.

Universitetsmuseene har et samfunnsoppdrag som innbefatter lagring og tilgjengeliggjøring av biodiversitetsmateriale. Disse har også kompetanse og infrastruktur for langvarig sikring, kuratering og kvalitetskontroll, men mangler per i dag ressurser for lagring og håndtering av et større antall slike prøver. Mange institusjoner og private firma har også egne samlinger og lagerfasiliteter, men har ikke krav til sikring, langtidsoppbevaring eller faglig kuratering på lik linje med universitetsmuseene.

Standardiserte metadata knyttet til miljøprøver må loggføres i egnede systemer. Det er også en stor fordel om beskrivelser av gjennomførte analyser og publiserte resultater knyttes til prøven i databasen. På den måten kan en unngå duplisering av analyser og unødvendig forbruk av verdifullt forskningsmateriale. For å sikre at innsamlede prøver og resulterende DNA-ekstrakter er tilgjengelige for miljøforvaltning (og forskning), må som et minimum informasjon om prøvene legges ut i offentlig tilgjengelige, søkbare databaser. Det eksisterer i dag løsninger for miljøprøver og miljø-DNA som er utviklet eller er under utvikling i GBIF og GGBN, og som følger samme dataflyt som skissert ovenfor under kapittel 2.2.

En forutsetning er at miljø-prøver må gis en stabil og globalt unik ID (for eksempel i form av URN:UUID). Dette for at prøver kan henvises til i separate datasett med resultatdata (f.eks. fra miljø-DNA analyser). Det vil da i etterkant være mulig å knytte sammen slik informasjon selv om den publiseres fra ulike datasett. Det finnes gode verktøy for å koble slik distribuert informasjon, blant annet tilgjengelig i R og tilsvarende skriptspråk med utbredt anvendelse i både grunnleggende og anvendte forskningsmiljøer. Den før nevnte Living Atlas plattformen kan også med noe implementering konfigureres for å indeksere og levere innsyn til slike data hentet fra multiple distribuerte datasett (diskutert under 2.2.5). Spesifisering av ny kjernestruktur i DwC-A, gjennom en egen kjernetabell dedikert til miljøprøver og belegg av organismer (“MaterialSample Core”), vil også kunne gjøre harmonisering av data enklere.

3. Oppsummering / konklusjon

Studier basert på miljø-DNA prøver inneholder mange ulike steg og har potensiale for mange feilkilder. Resultater herfra kan derfor være krevende å tolke og krever et bredt spekter av ekspertise for kvalitetssikring. Selv om metoden er laboratoriebasert og teknisk i utførelse, må sluttresultater evalueres av økologiske og taksonomiske eksperter. Det er derfor vesentlig at resultater kan etterprøves på alle nivåer av prosessen. Dette gjør at tilgjengeliggjøring av både primærdata og deriverte resultatdata (artslister) er viktige fra et rent kvalitetssikringsperspektiv, men også ut fra perspektivet om at data skal kunne gjenbrukes. En minstestandard på henvisning til og bruk av referansedatabaser i studier basert på DNA-strekkoding må innbefatte tydelig dokumentert DNA-identifikasjoner med presis referanse til referansesekvensen, angi likhet mellom sekvens og referansesekvens, samt å stille krav til brukere av referansedatabaser i forhold til organismekunnskap. Det må også stilles krav til referansedatabaser som benyttes. Disse må være offentlig tilgjengelige og referansesekvens lenket til belegg i offentlig tilgjengelige samlinger som er kuraterbare med hensyn på taksonomi. Med hensyn på studier basert på qPCR og ddPCR er det vesentlig at disse baserer seg på primere som er dokumentert og offentlig tilgjengelige i henhold til etablerte fagstandarer.

En felles standard for innrapportering og offentlig tilgjengeliggjøring av DNA- sekvenser og tilhørende metadata fra oppdrag er gjennomførbar basert på eksisterende åpne internasjonale standarder og infrastrukturer for utveksling av biodiversitetdata. Noe utvikling med hensyn på tekniske løsninger for dataflyt mellom ulike fagsystemer må påregnes, og det må defineres klare spesifikasjoner for bruk av standarder der hvor disse ikke er stringente nok til å dekke forvaltningens behov for koherens mellom datasett. Dette involverer nødvendigvis noe utviklingsarbeid, for eksempel i forbindelse med spesifisering av krav til bruk av kodeverk / vokabular.

Oppbevaring av innsamlede prøver, som ikke analyseres umiddelbart, over lengre tid for å sikre tilgjengelighet for framtidig forskning og forvaltning, uavhengig av opprinnelig oppdragsinstitusjon, må være et viktig bidrag for å sikre materiale. Imidlertid må dette nøye veies opp mot kostnader. Per i dag eksisterer det ikke kapasitet i dagens infrastruktur for at dette skal implementeres i stor skala. I tillegg til investeringer må det beregnes betydelig kostnader for drift. Tilgjengeliggjøring av informasjon om tilstedeværelse og innhold av slike prøver vurderes som mindre kostnadskrevende, og infrastruktur og standarder skissert for formidling av data fra miljø-DNA studier kan gjenbrukes til dette formålet.

4. Referanser

Alberdi, Antton, Ostaizka Aizpurua, M. Thomas P. Gilbert, and Kristine Bohmann. 2018. "Scrutinizing Key Steps for Reliable Metabarcoding of Environmental Samples." Edited by Andrew Mahon. *Methods in Ecology and Evolution / British Ecological Society* 9 (1): 134–47.

Andersen, Kenneth, Karen Lise Bird, Morten Rasmussen, James Haile, Henrik Breuning-Madsen, Kurt H. Kjaer, Ludovic Orlando, M. Thomas P. Gilbert, and Eske Willerslev. 2012. "Meta-Barcoding of 'Dirt' DNA from Soil Reflects Vertebrate Biodiversity." *Molecular Ecology* 21 (8): 1966–79.

Auriac, Marc B. Anglès d', Marc B. Anglès d'Auriac, David A. Strand, Marit Mjelde, Benoit O. L. Demars, and Jens Thaulow. 2019. "Detection of an Invasive Aquatic Plant in Natural Water Bodies Using Environmental DNA." *PLOS ONE*. https://doi.org/10.1371/journal.pone.0219700.

Bergsten, Johannes, David T. Bilton, Tomochika Fujisawa, Miranda Elliott, Michael T. Monaghan, Michael Balke, Lars Hendrich, et al. 2012. "The Effect of Geographical Scale of Sampling on DNA Barcoding." *Systematic Biology* 61 (5): 851–69.

Biggs, Jeremy, Naomi Ewald, Alice Valentini, Coline Gaboriaud, Tony Dejean, Richard A. Griffiths, Jim Foster, et al. 2015. "Using eDNA to Develop a National Citizen Science-Based Monitoring Programme for the Great Crested Newt (Triturus Cristatus)." *Biological Conservation* 183 (0): 19–28.

Bista, Iliana, Gary R. Carvalho, Min Tang, Kerry Walsh, Xin Zhou, Mehrdad Hajibabaei, Shadi Shokralla, et al. 2018. "Performance of Amplicon and Shotgun Sequencing for Accurate Biomass Estimation in Invertebrate Community Samples." *Molecular Ecology Resources*, April. https://doi.org/10.1111/1755-0998.12888.

Braukmann, Thomas W. A., Natalia V. Ivanova, Sean W. J. Prosser, Vasco Elbrecht, Dirk Steinke, Sujeevan Ratnasingham, Jeremy R. de Waard, Jayme E. Sones, Evgeny V. Zakharov, and Paul D. N. Hebert. 2019. "Metabarcoding a Diverse Arthropod Mock Community." *Molecular Ecology Resources* 19 (3): 711–27.

Bustin, Stephen A., Vladimir Benes, Jeremy A. Garson, Jan Hellemans, Jim Huggett, Mikael Kubista, Reinhold Mueller, et al. 2009. "The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments." *Clinical Chemistry* 55 (4): 611–22.

Buttigieg, Pier Luigi, Felix Janssen, James Macklin, and Kathleen Pitz. 2019. "The Global Omics Observatory Network: Shaping Standards for Long-Term Molecular Observation." *Biodiversity Information Science and Standards* 3 (June): e36712.

Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13 (7): 581–83.

Calvignac-Spencer, Sébastien, Kevin Merkel, Nadine Kutzner, Hjalmar Kühl, Christophe Boesch, Peter M. Kappeler, Sonja Metzger, Grit Schubert, and Fabian H. Leendertz. 2013. "Carrion Fly-Derived DNA as a Tool for Comprehensive and Cost-Effective Assessment of Mammalian Biodiversity." *Molecular Ecology* 22 (4): 915–24.

Campbell, Lori D., Jonas J. Astrin, Yvonne DeSouza, Judith Giri, Ashokkumar A. Patel, Melissa Rawley-Payne, Amanda Rush, and Nicole Sieffert. 2018. "The 2018 Revision of the ISBER Best Practices: Summary of Changes and the Editorial Team's Development Process." *Biopreservation and Biobanking*. https://doi.org/10.1089/bio.2018.0001.

CEN. 2018. "Water Quality - Technical Report for the Management of Diatom Barcodes." CEN/TR 17244 .

Chakrabarty, Prosanta, Melanie Warren, Lawrence Page, and Carole Baldwin. 2013. "GenSeq: An Updated Nomenclature and Ranking for Genetic Sequences from Type and Non-Type Sources." *ZooKeys*. https://doi.org/10.3897/zookeys.346.5753.

Clare, Elizabeth L. 2014. "Molecular Detection of Trophic Interactions: Emerging Trends, Distinct Advantages, Significant Considerations and Conservation Applications." *Evolutionary Applications* 7 (9): 1144–57.

Crampton-Platt, Alex, Douglas W. Yu, Xin Zhou, and Alfried P. Vogler. 2016. "Mitochondrial Metagenomics: Letting the Genes out of the Bottle." *GigaScience* 5 (March): 15.

DiBattista, Joseph D., Darren J. Coker, Tane H. Sinclair-Taylor, Michael Stat, Michael L. Berumen, and Michael Bunce. 2017. “Assessing the Utility of eDNA as a Tool to Survey Reef-Fish Communities in the Red Sea.” *Coral Reefs*. https://doi.org/10.1007/s00338-017-1618-1.

Doi, Hideyuki, Keiichi Fukaya, Shin-Ichiro Oka, Keiichi Sato, Michio Kondoh, and Masaki Miya. 2019. “Evaluation of Detection Probabilities at the Water-Filtering and Initial PCR Steps in Environmental DNA Metabarcoding Using a Multispecies Site Occupancy Model.” *Scientific Reports* 9 (1): 3581.

Doi, Hideyuki, Ryutei Inui, Yoshihisa Akamatsu, Kazuki Kanno, Hiroki Yamanaka, Teruhiko Takahara, and Toshifumi Minamoto. 2017. “Environmental DNA Analysis for Estimating the Abundance and Biomass of Stream Fish.” *Freshwater Biology*. https://doi.org/10.1111/fwb.12846.

Dougherty, Matthew M., Eric R. Larson, Mark A. Renshaw, Crysta A. Gantz, Scott P. Egan, Daniel M. Erickson, and David M. Lodge. 2016. “Environmental DNA (eDNA) Detects the Invasive Rusty Crayfish at Low Abundances.” *The Journal of Applied Ecology* 53 (3): 722–32.

Droege, G., K. Barker, O. Seberg, J. Coddington, E. Benson, W. G. Berendsohn, B. Bunk, et al. 2016. “The Global Genome Biodiversity Network (GGBN) Data Standard Specification.” *Database: The Journal of Biological Databases and Curation* 2016 (October). https://doi.org/10.1093/database/baw125.

Edge, Thomas A., Donald J. Baird, Guillaume Bilodeau, Nellie Gagné, Charles Greer, David Konkin, Glen Newton, et al. 2019. “The Ecobiomics Project: Advancing Metagenomics Assessment of Soil Health and Freshwater Quality in Canada.” *The Science of the Total Environment* 710 (December): 135906.

Ekrem, Torbjørn, and Markus Majaneva. 2019. “DNA-Metastrekkoding Til Undersøkelser Av Invertebrater I Ferskvann.” *NTNU Vitenskapsmuseet Naturhistorisk Notat*. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2612638.

Elbrecht, Vasco, and Florian Leese. 2017. “Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment.” *Frontiers in Environmental Science* 5: 11.

Epp, Laura S., Sanne Boessenkool, Eva P. Bellemain, James Haile, Alfonso Esposito, Tiayyba Riaz, Christer Erséus, et al. 2012. “New Environmental Metabarcodes for Analysing Soil DNA: Potential for Studying Past and Present Ecosystems.” *Molecular Ecology* 21 (8): 1821–33.

Ficetola, Gentile F., Johan Pansu, Aurélie Bonin, Eric Coissac, Charline Giguet-Covex, Marta De Barba, Ludovic Gielly, et al. 2015. “Replication Levels, False Presences and the Estimation of the Presence/absence from eDNA Metabarcoding Data.” *Molecular Ecology Resources* 15 (3): 543–56.

Ficetola, Gentile Francesco, Claude Miaud, François Pompanon, and Pierre Taberlet. 2008. “Species Detection Using Environmental DNA from Water Samples.” *Biology Letters*. https://doi.org/10.1098/rsbl.2008.0118.

Field, Dawn, Linda Amaral-Zettler, Guy Cochrane, James R. Cole, Peter Dawyndt, George M. Garrity, Jack Gilbert, et al. 2011. “The Genomic Standards Consortium.” *PLoS Biology* 9 (6): e1001088.

Fossøy, Frode, Hege Brandsegg, Rolf Sivertsgård, Oskar Pettersen, Brett K. Sandercock, Øyvind Solem, Kjetil Hindar, and Tor Atle Mo. 2019. “Monitoring Presence and Abundance of Two Gyrodactylid Ectoparasites and Their Salmonid Hosts Using Environmental DNA.” *Environmental DNA*. https://doi.org/10.1002/edn3.45.

Fossøy, Frode, Sondre Dahle, Line Birkeland Eriksen, Merethe Hagen Spets, Sten Karlsson, and Trygve Hesthagen. 2017. “Bruk Av Miljø-DNA for Overvåking Av Fremmede Fiskearter. Utvikling Av Artsspesifikke Markører for Gjedde, Mort Og ørekyt.” *Norsk Institutt for Naturforskning*. NINA Rapport 1299.

Fossøy, Frode, Jens Thaulow, Marc Anglès d’Auriac, Hege Brandsegg, Rolf Sivertsgård, Tor Atle Mo, Odd Terje Sandlund, and Trygve Hesthagen. 2018. “Bruk Av Miljø-DNA Som Supplerende Verktøy for Overvåkning Og Kartlegging Av Fremmed Ferskvannsfisk.” *Norsk Institutt for Naturforskning*. NINA Rapport 1586.

Frøslev, T., and R. Ejrnæs. 2018. “BIOWIDE eDNA Fungi Dataset. Danish Biodiversity Information Facility. Occurrence Dataset.” https://doi.org/10.15468/nesbvx.

Gauthier, Mailys, Lara Konecny-Dupré, Agnès Nguyen, Vasco Elbrecht, Thibault Datry,

Christophe Douady, and Tristan Lefébure. 2019. “Enhancing DNA Metabarcoding Performance and Applicability with Bait Capture Enrichment and DNA from Conservative Ethanol.” *Molecular Ecology Resources*, September. https://doi.org/10.1111/1755-0998.13088.

Gemeinholzer, Birgit, Isabel Rey, Kurt Weising, M. Grundman, Alexandra N. Muellner, Holger Zetzsche, Gabriele Droege, et al. 2010. “Organizing Specimen and Tissue Preservation in the Field for Subsequent Molecular Analyses.” *Manual on Field Recording Techniques and Protocols for All Taxa Biodiversity Inventories*.

Gómez-Rodríguez, Carola, Alex Crampton-Platt, Martijn J T, Andrés Baselga, and Alfried P. Vogler. 2015. “Validating the Power of Mitochondrial Metagenomics for Community Ecology and Phylogenetics of Complex Assemblages.” *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210x.12376.

Hajibabaei, Mehrdad, Teresita M. Porter, Michael Wright, and Josip Rudar. 2019. “COI Metabarcoding Primer Choice Affects Richness and Recovery of Indicator Taxa in Freshwater Systems.” *PloS One* 14 (9): e0220953.

Harper, Lynsey R., Lori Lawson Handley, Christoph Hahn, Neil Boonham, Helen C. Rees, Kevin C. Gough, Erin Lewis, et al. 2018. “Needle in a Haystack? A Comparison of eDNA Metabarcoding and Targeted qPCR for Detection of the Great Crested Newt (Triturus Cristatus).” *Ecology and Evolution*. https://doi.org/10.1002/ece3.4013.

Hebert, Paul D. N., Alina Cywinska, Shelley L. Ball, and Jeremy R. deWaard. 2003. “Biological Identifications through DNA Barcodes.” *Proceedings. Biological Sciences / The Royal Society* 270 (1512): 313–21.

Huggett, Jim F., Carole A. Foy, Vladimir Benes, Kerry Emslie, Jeremy A. Garson, Ross Haynes, Jan Hellemans, et al. 2013. “The Digital MIQE Guidelines: Minimum Information for Publication of Quantitative Digital PCR Experiments.” *Clinical Chemistry* 59 (6): 892–902.

Johnson, Matthew G., Lisa Pokorny, Steven Dodsworth, Laura R. Botigué, Robyn S. Cowan, Alison Devault, Wolf L. Eiserhardt, et al. 2019. “A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using K-Medoids Clustering.” *Systematic Biology* 68 (4): 594–606.

Jones, Matthew B., Mark P. Schildhauer, O. J. Reichman, and Shawn Bowers. 2006. “The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere,” November. https://doi.org/10.1146/annurev.ecolsys.37.091305.110031.

Kartzinel, Tyler R., Patricia A. Chen, Tyler C. Coverdale, David L. Erickson, W. John Kress, Maria L. Kuzmina, Daniel I. Rubenstein, Wei Wang, and Robert M. Pringle. 2015. “DNA Metabarcoding Illuminates Dietary Niche Partitioning by African Large Herbivores.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (26): 8019–24.

Knudsen, Steen, Martin Hesselsoe, Dorte Bekkevold, Søren Jensen, Peter Moller, and Jesper Andersen. 2018. *Tekniske Anvisninger for eDNA-Baseret Overvågning Af Ikke-Hjemmehørende Marine Arter*.

Knudsen, Steen Wilhelm, Rasmus Bach Ebert, Martin Hesselsøe, Franziska Kuntke, Jakob Hassingboe, Peter Bondgaard Mortensen, Philip Francis Thomsen, et al. 2019. “Species-Specific Detection and Quantification of Environmental DNA from Marine Fishes in the Baltic Sea.” *Journal of Experimental Marine Biology and Ecology* 510 (January): 31–45.

Kunnskapsdepartementet. 2017. “Nasjonal Strategi for Tilgjengeliggjøring Og Deling Av Forskningsdata.”

Kvist, Sebastian. 2013. “Barcoding in the Dark? A Critical View of the Sufficiency of Zoological DNA Barcoding Databases and a Plea for Broader Integration of Taxonomic Knowledge.” *Molecular Phylogenetics and Evolution* 69 (1): 39–45.

Lacoursière-Roussel, Anaïs, Maikel Rosabal, and Louis Bernatchez. 2016. “Estimating Fish Abundance and Biomass from eDNA Concentrations: Variability among Capture Methods and Environmental Conditions.” *Molecular Ecology Resources* 16 (6): 1401–14.

Laurendz, Charlotte. 2017. “Impact of Temperature, Food Availability and Life-History Stages on the eDNA Emission from Pacifastacus Leniusculus and Its Obligate Parasite Aphanomyces Astaci.” MsC, UiO.

Lecoq, Marie-Elise, Anne-Sophie Archambeau, Rui Figueira, David Martin, Sophie Pamerlon, Tim Robertson, Régine Vignes Lebbe, and Cristina Villaverde. 2019. “The Living Atlases Community of Practice.” *Biodiversity Information Science and Standards* 3 (June): e35779.

Lehmann, Sabine, Fiorella Guadagni, Helen Moore, Garry Ashton, Michael Barnes, Erica Benson, Judith Clements, et al. 2012. “Standard Preanalytical Coding for Biospecimens: Review and Implementation of the Sample PREanalytical Code (SPREC).” *Biopreservation and Biobanking*. https://doi.org/10.1089/bio.2012.0012.

Macklin, James, Donald Baird, and Keith Newton. 2019. “Ecobiomics: Environmental Metagenomic Biomonitoring.” *Biodiversity Information Science and Standards*. http://search.proquest.com/openview/b8e2a081a2ac8b2f4eb5faec3d7fb559/1?pq-origsite=gscholar&cbl=2049297.

Mahon, Andrew R., Christopher L. Jerde, Matthew Galaska, Jennifer L. Bergner, W. Lindsay Chadderton, David M. Lodge, Margaret E. Hunter, and Leo G. Nico. 2013. “Validation of eDNA Surveillance Sensitivity for Detection of Asian Carps in Controlled and Field Experiments.” *PloS One* 8 (3): e58316.

Majaneva, Markus, Ola H. Diserud, Shannon H. C. Eagle, Erik Boström, Mehrdad Hajibabaei, and Torbjørn Ekrem. 2018. “Environmental DNA Filtration Techniques Affect Recovered Biodiversity.” *Scientific Reports* 8 (1): 4682.

Majaneva, Markus, Ola H. Diserud, Shannon H. C. Eagle, Mehrdad Hajibabaei, and Torbjørn Ekrem. 2018. “Choice of DNA Extraction Method Affects DNA Metabarcoding of Unsorted Invertebrate Bulk Samples.” *Metabarcoding and Metagenomics* 2. https://doi.org/10.3897/mbmg.2.26664.

Meier, Rudolf. 2017. “Citation of Taxonomic Publications: The Why, When, What and What Not.” *Systematic Entomology*. https://doi.org/10.1111/syen.12215.

Minamoto, Toshifumi, Hiroki Yamanaka, Teruhiko Takahara, Mie N. Honjo, and Zen’ichiro Kawabata. 2012. “Surveillance of Fish Species Composition Using Environmental DNA.” *Limnology*. https://doi.org/10.1007/s10201-011-0362-4.

Moore, Helen M., Andrea Kelly, Lisa M. McShane, and Jim Vaught. 2013. “Biospecimen Reporting for Improved Study Quality (BRISQ).” *Transfusion*.

OECD. 1999. “Final Report on the OECD Megascience Forum Working Group on Biological Informatics.” http://www.oecd.org/science/inno/2105199.pdf.

Ogram, Andrew, Gary S. Sayler, and Tamar Barkay. 1987. “The Extraction and Purification of Microbial DNA from Sediments.” *Journal of Microbiological Methods*. https://doi.org/10.1016/0167-7012(87)90025-x.

Pietramellara, G., J. Ascher, F. Borgogni, M. T. Ceccherini, G. Guerri, and P. Nannipieri. 2009. “Extracellular DNA in Soil and Sediment: Fate and Ecological Relevance.” *Biology and Fertility of Soils*. https://doi.org/10.1007/s00374-008-0345-8.

Pilat, Dirk, and Yukiko Fukasaku. 2007. “OECD Principles and Guidelines for Access to Research Data from Public Funding.” *Data Science Journal*. https://doi.org/10.2481/dsj.6.od4.

Pleijel, F., U. Jondelius, E. Norlinder, A. Nygren, B. Oxelman, C. Schander, P. Sundberg, and M. Thollesson. 2008. “Phylogenies without Roots? A Plea for the Use of Vouchers in Molecular Phylogenetic Studies.” *Molecular Phylogenetics and Evolution*. https://doi.org/10.1016/j.ympev.2008.03.024.

PlutoF. 2018. “UNITE - Unified System for the DNA Based Fungal Species Linked to the Classification.” https://doi.org/10.15156/bio/587474.

Pompanon, Francois, Bruce E. Deagle, William O. C. Symondson, David S. Brown, Simon N. Jarman, and Pierre Taberlet. 2012. “Who Is Eating What: Diet Assessment Using next Generation Sequencing.” *Molecular Ecology* 21 (8): 1931–50.

Puillandre, N., P. Bouchet, M. -C. Boisselier-Dubayle, J. Brisset, B. Buge, M. Castelin, S. Chagnoux, et al. 2012. “New Taxonomy and Old Collections: Integrating DNA Barcoding into the Collection Curation Process.” *Molecular Ecology Resources*. https://doi.org/10.1111/j.1755-0998.2011.03105.x.

Ratnasingham, Sujeevan, and Paul D. N. Hebert. 2007. “Bold: The Barcode of Life Data System (http://www.barcodinglife.org).” *Molecular Ecology Notes* 7 (3): 355–64.

Robertson, Tim, Markus Döring, Robert Guralnick, David Bloom, John Wieczorek, Kyle Braak, Javier Otegui, Laura Russell, and Peter Desmet. 2014. “The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet.” *PloS One* 9 (8): e102623.

Rusch, Johannes C., Haakon Hansen, David A. Strand, Turhan Markussen, Sigurd Hytterød, and Trude Vrålstad. 2018. “Catching the Fish with the Worm: A Case Study on eDNA Detection of the Monogenean Parasite Gyrodactylus Salaris and Two of Its Hosts, Atlantic

Salmon (Salmo Salar) and Rainbow Trout (Oncorhynchus Mykiss).” *Parasites & Vectors*. https://doi.org/10.1186/s13071-018-2916-3.
Schigel, Dmitry, Thomas Jeppesen, Robert Finn, Guy Cochrane, Urmas Kõljalg, Christian Quast, Jerry Lanfear, Thomas Orrell, Donald Hobern, and Joseph Miller. 2019. “Going Molecular: Sequence-Based Spatiotemporal Biodiversity Evidence in GBIF.” *Biodiversity Information Science and Standards* 3: e37036.
Schmidt, Benedikt R., Marc Kéry, Sylvain Ursenbacher, Oliver J. Hyman, and James P. Collins. 2013. “Site Occupancy Models in the Analysis of Environmental DNA Presence/absence Surveys: A Case Study of an Emerging Amphibian Pathogen.” *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210x.12052.
Schnell, Ida Bærholm, Philip Francis Thomsen, Nicholas Wilkinson, Morten Rasmussen, Lars R. D. Jensen, Eske Willerslev, Mads F. Bertelsen, and M. Thomas P. Gilbert. 2012. “Screening Mammal Biodiversity Using DNA from Leeches.” *Current Biology: CB* 22 (8): R262–63.
Shehzad, Wasim, Tiayyba Riaz, Muhammad A. Nawaz, Christian Miquel, Carole Poillot, Safdar A. Shah, François Pompanon, Eric Coissac, and Pierre Taberlet. 2012. “Carnivore Diet Analysis Based on next-Generation Sequencing: Application to the Leopard Cat (Prionailurus Bengalensis) in Pakistan.” *Molecular Ecology* 21 (8): 1951–65.
Strand, David A., Arne Holst-Jensen, Hildegunn Viljugrein, Bente Edvardsen, Dag Klaveness, Japo Jussila, and Trude Vrålstad. 2011. “Detection and Quantification of the Crayfish Plague Agent in Natural Waters: Direct Monitoring Approach for Aquatic Environments.” *Diseases of Aquatic Organisms* 95 (1): 9–17.
Strand, David A., Stein Ivar Johnsen, Johannes C. Rusch, Sune Agersnap, William Brenner Larsen, Steen Wilhelm Knudsen, Peter Rask Møller, and Trude Vrålstad. 2019. “Monitoring a Norwegian Freshwater Crayfish Tragedy: eDNA Snapshots of Invasion, Infection and Extinction.” *Journal of Applied Ecology*. https://doi.org/10.1111/1365-2664.13404.
Strand, David A., Johannes Rusch, Stein Ivar Johnsen, Attila Tarpai, and Trude Vrålstad. 2019. “The Surveillance Programme for Aphanomyces Astaci in Norway 2018.” Norwegian Veterinary Institute.
Strickland, Garret J., and James H. Roberts. 2019. “Utility of eDNA and Occupancy Models for Monitoring an Endangered Fish across Diverse Riverine Habitats.” *Hydrobiologia*. https://doi.org/10.1007/s10750-018-3723-8.
Sweetlove, Maxime, Yi Ming Gan, Alison Murray, and Anton Van de Putte. 2019. “The Microbial Antarctic Resource System: Integrating Discoverability and Preservation of Environmentally-Annotated Microbial’omics Data.” *Biodiversity Information Science and Standards*. http://search.proquest.com/openview/5732e7509d452ad5a950687f9d0e166d/1?pq-origsite=gscholar&cbl=2049297.
Taberlet, Pierre, Aurélie Bonin, Eric Coissac, and Lucie Zinger. 2018. *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford University Press.
Taberlet, Pierre, Eric Coissac, Mehrdad Hajibabaei, and Loren H. Rieseberg. 2012. “Environmental DNA.” *Molecular Ecology* 21 (8): 1789–93.
Takahara, Teruhiko, Toshifumi Minamoto, Hiroki Yamanaka, Hideyuki Doi, and Zen ’ichiro Kawabata. 2012. “Estimation of Fish Biomass Using Environmental DNA.” *PloS One* 7 (4): e35868.
Taugbøl, A., Børre K. Dervo, Rolf Sivertsgård, Hege Brandsegg, and Frode Fossøy. 2018. “Bruk Av Miljø-DNA Til Overvåkning Av Små- Og Storsalamander.” *Norsk Institutt for Naturforskning* NINA-Rapport 1476.
Telfer, A. 2019. “Centre for Biodiversity Genomics - Canadian Specimens. Version 1.5. University of Guelph. Occurrence Dataset.” https://doi.org/10.15468/mbwnw9.
The International Barcode of Life Consortium. 2016. “International Barcode of Life Project (iBOL) Barcode Index Numbers (BINs).” https://doi.org/10.15468/wvfqoi.
Thomsen, Philip Francis, J. O. S. Kielgast, Lars L. Iversen, Carsten Wiuf, Morten Rasmussen, M. Thomas P. Gilbert, Ludovic Orlando, and Eske Willerslev. 2012. “Monitoring Endangered Freshwater Biodiversity Using Environmental DNA.” *Molecular Ecology* 21 (11): 2565–73.
Thomsen, Philip Francis, Peter Rask Møller, Eva Egelyng Sigsgaard, Steen Wilhelm Knudsen, Ole Ankjær Jørgensen, and Eske Willerslev. 2016. “Environmental DNA from Seawater

Samples Correlate with Trawl Catches of Subarctic, Deepwater Fishes." *PloS One* 11 (11): e0165252.
Thomsen, Philip Francis, and Eske Willerslev. 2015. "Environmental DNA – An Emerging Tool in Conservation for Monitoring Past and Present Biodiversity." *Biological Conservation* 183 (March): 4–18.
Valentini, Alice, Pierre Taberlet, Claude Miaud, Raphaël Civade, Jelger Herder, Philip Francis Thomsen, Eva Bellemain, et al. 2016. "Next-Generation Monitoring of Aquatic Biodiversity Using Environmental DNA Metabarcoding." *Molecular Ecology* 25 (4): 929–42.
Wacker, Sebastian, Frode Fossøy, Bjørn Mejdell Larsen, Hege Brandsegg, Rolf Sivertsgård, and Sten Karlsson. 2019. "Downstream Transport and Seasonal Variation in Freshwater Pearl Mussel (Margaritifera Margaritifera) eDNA Concentration." *Environmental DNA*. https://doi.org/10.1002/edn3.10.
Ward, R. D., R. Hanner, and P. D. N. Hebert. 2009. "The Campaign to DNA Barcode All Fishes, FISH-BOL." *Journal of Fish Biology*. https://doi.org/10.1111/j.1095-8649.2008.02080.x.
Weigand, Hannah, Arne J. Beermann, Fedor Čiampor, Filipe O. Costa, Zoltán Csabai, Sofia Duarte, Matthias F. Geiger, et al. 2019. "DNA Barcode Reference Libraries for the Monitoring of Aquatic Biota in Europe: Gap-Analysis and Recommendations for Future Work." *The Science of the Total Environment* 678 (August): 499–524.
Wieczorek, John, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. 2012. "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard." *PloS One* 7 (1): e29715.
Wilkinson, Mark D., Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (March): 160018.
Yates, Matthew C., Dylan J. Fraser, and Alison M. Derry. 2019. "Meta-analysis Supports Further Refinement of eDNA for Monitoring Aquatic Species-specific Abundance in Nature." *Environmental DNA*. https://doi.org/10.1002/edn3.7.
Yilmaz, Pelin, Renzo Kottmann, Dawn Field, Rob Knight, James R. Cole, Linda Amaral-Zettler, Jack A. Gilbert, et al. 2011. "Minimum Information about a Marker Gene Sequence (MIMARKS) and Minimum Information about Any (x) Sequence (MIxS) Specifications." *Nature Biotechnology* 29 (5): 415–20.

5. Vedlegg

Vedlegg 1: Eksempler på teknisk beskrivelse av miljø-DNA https://github.com/NTNU-VM/Darwinize-eDNA).

Miljødirektoratet
Telefon: 03400/73 58 05 00 | Faks: 73 58 05 01
E-post: post@miljodir.no
Nett: www.miljødirektoratet.no
Post: Postboks 5672 Torgarden, 7485 Trondheim
Besøksadresse Trondheim: Brattørkaia 15, 7010 Trondheim
Besøksadresse Oslo: Grensesvingen 7, 0661 Oslo

Miljødirektoratet jobber for et rent og rikt miljø. Våre hovedoppgaver er å redusere klimagassutslipp, forvalte norsk natur og hindre forurensning.

Vi er et statlig forvaltningsorgan underlagt Klima- og miljødepartementet og har mer enn 700 ansatte ved våre to kontorer i Trondheim og Oslo, og ved Statens naturoppsyn (SNO) sine mer enn 60 lokalkontor.

Vi gjennomfører og gir råd om utvikling av klima- og miljøpolitikken. Vi er faglig uavhengig. Det innebærer at vi opptrer selvstendig i enkeltsaker vi avgjør, når vi formidler kunnskap eller gir råd. Samtidig er vi underlagt politisk styring.
Våre viktigste funksjoner er at vi skaffer og formidler miljøinformasjon, utøver og iverksetter forvaltningsmyndighet, styrer og veileder regionalt og kommunalt nivå, gir faglige råd og deltar i internasjonalt miljøarbeid.