Research Ideas and Outcomes 10: e120483
doi: 10.3897/rio.10.e120483

OPEN ACCESS

Forum Paper

A lab-centric, workflow-based data management system for environmental DNA research

Alex Borisenko[‡], Robert G Young[§], Robert Hanner[§]

‡ Biodetics Data Solutions, Guelph, Canada
§ University of Guelph, Guelph, Canada

Corresponding author: Alex Borisenko (aborisenko@biodetics.net)

Reviewed v 1

Academic editor: Editorial Secretary

Received: 07 Feb 2024 | Accepted: 28 Feb 2024 | Published: 28 Mar 2024

Citation: Borisenko A, Young R, Hanner R (2024) A lab-centric, workflow-based data management system for environmental DNA research. Research Ideas and Outcomes 10: e120483. https://doi.org/10.3897/rio.10.e120483

Abstract

The adoption of environmental DNA approaches as a standard tool for biodiversity monitoring leads to the increase in the number of eDNA-based species occurrence records; however, considerable disparity remains in the nature and quality of associated information, much of it unpublished and/or poorly parametrised. A robust system for tracking biological materials from their point of origin through laboratory analyses is required to connect inferred taxon occurrences with analytical history and provenance data. The bulk of eDNA research is currently driven by small-scale operations where the tasks of digitisation, organisation and cross-referencing field records with laboratory analytical data and biomaterial sample location, are often performed manually and disconnected.

We present an integrative, full-stack data management solution that provides a structured ontological concept, a minimalist data schema for eDNA research and a software application prototype designed to facilitate real-time digitisation, parsing, annotation and archival of eDNA data. The system tracks the provenance and analytical history of biological samples through a structured hierarchy of events, linked with associated digital file attachment archives, such as images and raw sequence files, and with inferred taxonomic occurrence records. The data entry process is compartmentalised and incorporated into the corresponding stages of standard operations used in fieldwork, biological collection management and laboratory analysis. Resulting data records can be integrated into various output formats required for large-scale analytics, publication and/or submission to global data aggregators. The prototype is implemented on the Microsoft 365 platform as a relational database (Access) linked to cloud-based data tables (SharePoint) and a set of associated data conversion spreadsheets (Excel). The system is designed primarily around the data management needs of small research labs; however, it is scalable to larger institutions and inter-institutional academic networks.

Keywords

eDNA, database, Microsoft 365, Access, SharePoint, Excel, digitisation, fieldwork, collection management, LIMS

Background

Environmental DNA (eDNA) approaches have gained considerable traction in biodiversity research and monitoring (Schenekar 2023). Often considered superior to conventional biological surveys (e.g. Fediajevaite et al. (2021)), they are expanding into a widening gamut of applications, such as conservation (Barnes and Turner 2016, Belle et al. 2019, Beng and Corlett 2020), environmental impact assessments (Hinz et al. 2022) and One Health initiatives (Farrell et al. 2021, Ríos-Castro et al. 2021). The growing volume of eDNA case studies, accompanied by rapid development and gradual adoption of relevant methodologies as the new standard for environmental research, have bolstered the value of eDNA data as a source of species occurrences (Beng and Corlett 2020) and other Essential Biodiversity Variables (Hoban et al. 2022). This trend is prompting global biodiversity data aggregators, such as the Ocean Biodiversity Information System (OBIS) and the Global Biodiversity Information Facility (GBIF), to develop relevant data standards and to adjust the data schema (Berry et al. 2021), with the aim of hosting eDNA-derived datasets (Finstad et al. 2023, Powers and Hampton 2019).

While much effort has been devoted to workflow automation for processing already deposited biodiversity data (e.g. Mathew et al. (2014), Young et al. (2017)), such workflows are not equipped to assist with new data curation by field researchers, collection managers and laboratory staff responsible for recording, digitising and publishing field or laboratory data. Thus, despite the increasing accumulation rate of DNA-inferred species occurrence records, most of them remain unstructured and scattered amongst research papers or unpublished reports, while provenance information associated with published eDNA analytical results remains inconsistent and sometimes lacking (Nicholson et al. 2020). At the same time, the need for standardising data outcomes has been clearly identified (e.g. De Brauwer et al. (2023)).

Data provenance has long been a critical consideration in computer science, regarded as a potential major source of ambiguity and error in downstream analysis (Buneman et al. 2000). With respect to eDNA research and other field-based biological disciplines, this is particularly true for provenance data. Despite optimistic projections that the challenges of capturing provenance information for biological samples would be solved by large-scale deployment of automated field sampling devices, analytical pipelines and scientific workflow systems (Bohan et al. 2017, Reichman et al. 2011), the tasks of data logging, conversion and integration largely remain time-consuming and manually driven (Michener and Jones 2012). Thus, incorporation of eDNA-derived data into Essential Biodiversity Variables (Kissling et al. 2018, Hardisty et al. 2019) requires more efficient, structured approaches towards digitisation, storage and publication of eDNA-inferred occurrence records.

Several large institutions and research networks are developing centralised field survey data management platforms (ten Hoopen et al. 2022); and efforts are underway to deploy institution, project-wide or international frameworks to collate and organise information from field survey activities (e.g. Hackett et al. (2019), Kõljalg et al. (2019), Penev et al. (2022)). These trends align with emerging requirements imposed by major funding agencies for grant applicants to develop and adhere to comprehensive data management plans for their research projects. Although put in place with institution and networks in mind, these responsibilities are often “passed down” to individual applicants. Unfortunately, many small-scale operations, including most government or academic research labs and environmental assessment companies, may not possess the workforce needed to set up and operate high-maintenance data management systems. Furthermore, many biologists have limited experience with relational database management (Philippi and Köhler 2006); hence, they are not technically equipped to do much more than to deposit poorly structured “data dumps” on to centralised data repositories, while paying lip-service to the data management requirements of their project research networks and funders.

While advocating for better resourcing of data management efforts deployed by smaller-scale eDNA research operations, we posit that increasing their efficiency as providers of accurate and standardised genomic biodiversity data requires overcoming several operational challenges outlined below.

Data Collection Challenges

Challenges to efficient data collection stem from the inherently complicated nature of biodiversity informatics (Morris 2005), as exemplified by the experimental design of eDNA research projects. These are usually comprised of several hierarchically interconnected stages (field collecting of samples and environmental data, sample processing, molecular analyses, informatics pipelines) that span varied dates and locations and are often performed by different agents (people and/or organisations).

Despite the multitude of biological data management systems developed to date, most of them are not readily deployable in small labs or lack the intuitive structure that make them available for a particular application (Anonymous 2006, Saarnak et al. 2013). “Off-the-shelf” database solutions that are ready for deployment tend to be expensive and typically have a pre-defined data schema and field structure, thereby lacking the flexibility required to accommodate specialised user needs. This leads to database field “co-opting” (misappropriation) and other workarounds (Thomer et al. 2017) — a problem that is likely to be exacerbated in disciplines that are actively evolving or undergoing rapid methodological development, such as genomics. In the end, researchers, students and technicians often have to develop their own *ad hoc* task-specific solutions for data collection, storage and exchange. Consequently, data capture in small labs often represents a patchwork of printed or hand-written labels, field notes, lab journals and computer spreadsheets, which do not always conform to existing data standards and best practices. Furthermore, detailed record-keeping often competes for researcher's attention and resources in a time-sensitive and logistically complicated fieldwork or laboratory setting. Procedural shortcuts taken under operational duress can lead to incomplete or inaccurate data records, compromising data quality.

Data Integration Challenges

When the outcomes of eDNA research are communicated through scientific publications or technical reports, associated raw data archives may remain in proprietary custody. If published, they may be structured according to a multitude of disparate publisher or client requirements. Publication data standards for biodiversity and ecology advocate the use of non-relational ("flat-file") spreadsheets for data submission (Costello and Wieczorek 2014) that are easily stored, parsed and managed by the publisher or aggregator.

Such publication datasets are often manually collated by researchers at the end of their study or even later. Data may be sourced from disparate, disconnected and sometimes poorly-validated records made by different people during different stages of the project. The validation of researcher data against a publisher's standards usually happens during the data submission process (e.g. Robertson et al. (2014)), rather than the data collection phase. Consequently, changes or corrections made to the data during publication may not be reflected in the archived field notes or lab books, causing disparities between original and published data.

Due to the complicated nature of eDNA field sampling techniques and molecular analytical pathways, information pertaining to sourcing, managing, processing and analysing eDNA samples may comprise hundreds of data fields, many of which can be specific to particular sampling or analytical methodologies. A single research lab often hosts several projects simultaneously, each with its unique research design and methods, which may change over time. Parsing and transferring this diverse information while keeping track of the different projects is a daunting undertaking, especially when manual data manipulations are required to transcribe personal records and notes. Consolidating disparate and unstructured field/lab records retrospectively into a single dataset can also be time-consuming and mentally taxing.

Finally, when the nested relational hierarchy of research stages and material transformations is "flattened" into a single non-relational spreadsheet during integration (Philippi and Köhler 2006), data contained in most fields are necessarily duplicated, resulting in considerable redundancy of records (Morris 2005), which obscures the experimental logic model. When done manually, this may provoke more procedural shortcuts and lead to further data migration errors, omissions, mix-ups and/or disconnection between analysed biological materials and associated data records.

Addressing Data Challenges

The above challenges particularly affect small-scale research operations, which constitute a major part of the eDNA research establishment. As a result, a large proportion of generated genetic and survey data, even if technically published, remains practically unusable for large-scale parametrised meta-analysis. Although this is a universal problem plaguing biodiversity datasets at large (Blair et al. 2020), it seems to attract little mainstream attention.

For example, a recent comprehensive review of eDNA metabarcoding in the assessment of aquatic ecosystems (Pawlowski et al. 2018) focused on technical aspects of molecular and bioinformatic analyses, but did not regard the integration of structured data pertaining to samples and their provenance amongst noteworthy considerations or recommended actions. Likewise, field/lab data collection and management are not even mentioned in recently published DNA-based biodiversity assessment/monitoring guidelines (Bruce et al. 2021, Liu et al. 2020). Furthermore, although major thematic networks such as DNAqua-Net (Leese et al. 2016) and national strategies (De Brauwer et al. 2023, Kelly et al. 2023) are emerging to coordinate eDNA research, they seem to lack coherent plans for consolidating provenance-linked eDNA-inferred occurrence records and providing them in a structured and coordinated way to global biodiversity data aggregators.

To overcome or, at least, to alleviate these shortfalls, more attention needs to be paid to structured data digitisation. In particular, efforts should concentrate on facilitating the data capture and management needs of eDNA research operations that perform these tasks. An important step in their adherence to current standards and best practices would be the development of data management tools that are intuitive, user-friendly, locally deployable and customisable for small-scale operations, while providing downstream integration with data aggregators. Such tools should facilitate efficient tracking of biological samples and real-time data entry while reflecting the logic of each lab's operational workflows and supporting connectivity between different stages — particularly between fieldwork and laboratory experiments. Finally, these tools should be seamlessly integrated within each eDNA research operation into a single coherent data management system built on a commonly used software platform that does not require specialised technical background or IT staff to deploy and maintain. A working prototype for such a system is described herein.

Conceptual Framework

We propose using a single relational laboratory-wide database with compartmentalised, staged data entry protocols that map the operational complexity of eDNA projects. Real-time data recording and validation is facilitated by breaking it down into manageable partitions, corresponding in sequential order and content to the individual stages of the research workflow. This makes it easier for different researchers and staff members to relay information between projects and research phases using a common data standard. Under this scenario, publication datasets and summary reports can be generated using automated data queries, with moderate added effort and minimal data loss. Real-time and unambiguous linking of data records with biological materials facilitates efficient access to them when additional analyses are required.

Operational Framework

To develop the relational data architecture that would facilitate structured data entry, it is important to conceptualise its general **operational framework** (what core objects or entities we are dealing with) and **ontological framework** (broad categories of data that are being recorded). Below, we outline these conceptual considerations in more detail. We further provide an overview of a minimalist data schema and present examples of implementing it as a standard for practical application in a small laboratory context.

Biological Objects: Specimens, Lots, Bulk Samples and Environmental DNA

Conventional zoological and botanical collecting activities usually preserve target organisms as representative biological individuals (in unitary organisms), clones (in modular organisms) or as fragments thereof. Such preserved organisms, conventionally referred to as **voucher specimens** (e.g. Culley (2013), Martin (1990)), are generally assumed to possess a single taxonomic identity, i.e. to belong to a certain species (whether known or not). [Voucher] specimens were regarded as key elements of biorepository-underpinned genomic diversity research at large (Hanner and Gregory 2007) and recognised as central to the logistical framework of DNA barcoding workflows (Borisenko et al. 2009) where reference DNA sequences were derived from isolated, taxonomically identified organisms. This voucher-centric framework remains salient for many current genomic studies and initiatives (Buckner et al. 2021, Lewin et al. 2018).

Recent syn-ecological advances, aided by rapidly developing DNA technologies, are expanding the perception of an organism beyond its core taxonomic identity. Instead, organisms are increasingly recognised as hosts to diverse microbiomes (Gibson et al. 2014), pathobiomes (Vayssier-Taussat et al. 2014) or symbiomes (Thompson et al. 2021) whose metagenomes harbour genetic information from a consortium of multiple taxa, which may be studied in the context of their trophic (Anonymous 2013, Wells et al. 2022), mutualistic (Bell et al. 2017, Pornon et al. 2017) or other ecological relationships. This paradigm shift extends to the corresponding collection voucher specimens, obscuring their operational distinction from other types of biological materials discussed below.

Much of ecological genomic research deals with field-collected aggregations of multiple uncounted, sometimes undiscernible organisms of different, often unknown taxonomic identities. Such aggregations are commonly referred to as "**bulk samples**" (Gibson et al. 2014, Leray and Knowlton 2015, Taberlet et al. 2018). Although increasingly used in genomic literature, the term "bulk sample" has not been applied in the broader environmental sampling context (Hoffmann 1994); instead, it has been used to describe aggregate non-biological samples (Zhang 2007). To avoid terminological confusion, we suggest that the qualifier "bulk" is best reserved for cases when unaltered portions of substrate or water are taken for analysis; however, the composition and concentration of biological materials within that "sample" remains the same as in the environment from which it has been taken. By contrast, most biological research involves some form of differential extraction and/or condensing of organisms or their derivatives, such as DNA, metabolites or other organic matter, relative to their original occurrence or concentration in the environment. For example, the contents of a collecting container in a Malaise trap, Berlese funnel, plankton tow or organic slough on an eDNA water filter represent concentrated organisms or organic matter derived from the air, water or substrate. Thus, we suggest that it is conceptually important to discern extracted/concentrated biological materials from unaltered bulk portions removed from the environment.

The term "sample" has been widely used to describe organismal parts or pieces of tissue destined for laboratory analysis (e.g. Kilian et al. (2015), Plitzner et al. (2017)). This terminological overlap is sometimes obviated by using "subsample" to discern lab-derived materials from field-sourced "bulk samples" (Gibson et al. 2014, Leray and Knowlton 2015); however, the term "subsample" may also be used to characterise portions of an organismal tissue sample, adding to further terminological ambiguity. Thus, we propose to avoid using the term "sample" to describe materials sourced from the field, at least as a technical definition for operational entities in the data schema (see below).

The term "lot" is commonly used in biological collection management to categorise a batch of multiple organisms derived from a single collecting event. It is sometimes restricted to characterise taxonomically sorted aggregations of specimens and juxtaposed to unsorted "bulk samples", such as trap contents sourced from the field (Anonymous 2016). We think that the distinction between "sorted" and "unsorted" collections is arbitrary because the very act of field collecting (including eDNA filtration) is, in essence, a form of initial sorting imposed by the chosen collecting methodology. To account for these collecting/sampling scenarios and to obviate the potential terminological confusion between field-sourced and laboratory-derived biological materials, we propose to expand the term "Lot" to characterise all types of field-sourced biological materials, including eDNA; but to restrict the term "Sample" to its laboratory-derived partition, which may include DNA filtered from the preservation medium. Correspondingly, within our proposed data architecture, we will apply the term "collecting" to field sourcing of biological materials and will restrict the term "sampling" to processing/partitioning these materials for laboratory analysis. Finally, within this terminological context, we see no need to define "subsamples" or "subsampling" as distinct operational categories.

We posit that, despite the fundamental biological difference between lots, individual organisms and environmental DNA, the logistics of field sourcing, processing and analysing biological materials of different nature are fundamentally similar. For example, molecular analytical protocols applied in environmental DNA research can also be used for DNA-based biodiversity analysis of aggregate specimen collections, such as arthropod traps or plankton tows. Analyses of such lots can be done by picking out and sequencing individual specimens (e.g. Young et al. (2017)), by homogenising the entire contents of the storage container for bulk DNA extraction, followed by metabarcoding (e.g. Leray and Knowlton (2015)) or by filtering and metabarcoding the organic suspension diffused into the fixative (Milián-García et al. 2021, Milián-García et al. 2021, Zizka et al. 2019). In the latter case, preserved specimens remain relatively intact, allowing for subsequent morphological or organism-based DNA analyses. A robust data management approach should accommodate all these research designs, whether DNA is taken directly from the environment, the storage medium or extracted from preserved organisms.

Operational Entities: Darwin Core MaterialSample vs. Event Class

Data records hosted by biodiversity data aggregators, such as GBIF, are centred around "species occurrences" or "observations" (*sensu* Lindström (2006)), which report taxonomically identified biological organisms collected or otherwise detected in a certain location at a certain time. This is also known as the Occurrence Core (Wieczorek et al. 2014) approach. The material entity upon which the Occurrence record is based is known as MaterialSample, defined as a "physical result of a sampling (or subsampling) event" in the Darwin Core data schema—https://dwc.tdwg.org/terms/#materialsample (Mayfield-Meyer et al. 2022, Wieczorek et al. 2014). The different categories of biological objects discussed above are all examples of the MaterialSample class. Although not a necessary component of an observational record, the MaterialSample is a critical element of both collection specimen-based and eDNA-inferred taxonomic occurrences.

Once a MaterialSample is collected in the field (Lot), it may be processed/subdivided (Sample) and transformed, for example, through DNA extraction (Aliquot, see below). It may further be transferred between agents, research teams, labs, institutions etc. during different phases of the analytical process. During each of these stages, associated data must "pass through" the data management system of the next processing facility efficiently and without information loss. A laboratory should be able to use the same data management system to track eDNA research, to facilitate metagenomic analyses of lots (e.g. invertebrate trap contents) and to contribute reference DNA sequences derived from taxonomically curated voucher specimens. In a "simple" eDNA research scenario, the same water filter with field-collected organic slough may be registered as a Lot or as a Sample, depending on its processing stage. The proposed data management framework provides sufficient flexibility required to accommodate the various collection processing pathways for eDNA research and other emerging fields of enquiry. At the same time, it conforms to the logic model of conventional collection-based biodiversity research, which reduces potential connectivity issues during future "crosswalking" with data schemas used in natural history collection databases (Thomer et al. 2017).

The second, operationally critical part of an occurrence record is the Event, broadly defined within Darwin Core as an "action that occurs at some location during some time" — https://dwc.tdwg.org/terms/#event (Wieczorek et al. 2014). Within the narrower context of biological collecting, an Event, sometimes defined as "gathering" (Lindström 2006), could be characterised as an action aimed at acquiring a MaterialSample. If successful, the event can, thus, be regarded as the source or origin of derived MaterialSample(s) and the corresponding occurrence record(s); however, even if unsuccessful, it provides an important methodological context on the broader collecting effort deployed to obtain an occurrence dataset.

The taxonomic identity of the MaterialSample constitutes the central piece of information contained in an occurrence record; however, it is of tangential importance to the logistics of an eDNA research project. The detection of certain taxa in a sample depends on the sampling methodology used (e.g. study site choice, filtration technique, preservation parameters) and is derived from a certain procedural outcome (e.g. targetted PCR detection, Sanger sequencing or metabarcoding). Each sample can be subdivided and processed using different analytical and/or bioinformatic pipelines or as several replicates using the same pipeline. As the limit of detection for different taxa may vary between methods and/or analytical parameters used, these analyses may yield varying taxonomic outcomes. Thus, although taxonomic occurrence records are the end-point of many eDNA research projects, they are best treated as context-dependent annotations of a MaterialSample and only meaningful if underpinned by a robust and adequately parametrised “Event—MaterialSample” data dyad. This data management approach is congruent with the emerging Collecting Event Core concept (Kissling et al. 2018, Wieczorek et al. 2014, M and RJ 2017) that emphasises the critical role of methodologies in producing specific occurrence records.

In eDNA research, as with other taxonomic inferences derived from collected and analysed biological objects, it is practical to shift the emphasis of the data model from Occurrence Core to Event Core. Under the Event Core logic model (Kissling et al. 2018), biodiversity survey activities (e.g. expeditions or field trips) can be broken down into a series of [collecting] Events, each of which typically results in a MaterialSample that, in turn, undergoes a series of analytical procedures to infer biodiversity information (occurrences). Additional data elements are necessary to attain Event Core parametrisation. For example, geospatial localisation of collecting events is needed to link a MaterialSample with a pre-defined collecting locality (e.g. site, station), corresponding to Darwin Core’s Location Class (https://dwc.tdwg.org/terms/#location).

From a pragmatic laboratory data management point of view, it is important to acknowledge that the Darwin Core schema employed by GBIF was designed to facilitate biodiversity data publication (Robertson et al. 2014, Wieczorek et al. 2014), rather than data capture. While it is an important minimalist, universal data standard, it is not necessarily sufficient to accommodate all data elements pertinent to particular use cases (Chapman et al. 2020) and does not offer the relational structure required to track the nested hierarchy of field collecting, laboratory operations or other logistical aspects of biodiversity research. For example, both MaterialSample and Event classes have been used to characterise field collecting efforts and outcomes (e.g. Kissling et al. (2018), Wieczorek et al. (2014)). As mentioned earlier, DNA-inferred taxon detections are also dependent on the outcome of molecular analyses of derived DNA aliquots, which should be regarded as separate “analytical events” and “material sub-samples”, respectively. Within this context, the Collecting Event would be distinct from the Analytical Event. This approach mirrors the distinction between two different classes of events characterising the "material sampling process" and the "identification process" recognised within the Biological Collections Ontology (Walls et al. 2014).

To maintain semantic distinction between field collecting and laboratory analyses, we will refer to the former as Events and to the latter as Analyses, each characterised by a defined methodology and localised in space and time. Operationally, this allows breaking the data entry process into stages corresponding to phases of field collecting, post-field processing and laboratory analyses. Keeping track of unsuccessful Events and Analyses ("negative results") further parametrises the methodological context for the sought taxonomic occurrence outcomes. For example, it may be useful to know that the detection of a certain taxon in a certain locality is linked to several unsuccessful attempts to recover its sequence using alternative collecting protocols or analytical parameters. Darwin Core does not accommodate for this relational complexity (Walls et al. 2014), although it offers semantic provisions for reporting genomic information through its associatedSequences term—https://dwc.tdwg.org/list/#dwc_associatedSequences and associated terms.

Ontological Framework

From a broad philosophical perspective, contemporary field-based biological disciplines, including eDNA research, span two classical domains of enquiry: Natural History, which aims to accrue empirical knowledge about the natural world and Natural Philosophy, which aims to infer abstract universal patterns (Anstey 2012). Although eDNA research is methodologically rooted in Natural Philosophy principles, for operational purposes, it could be regarded as an extension of the Natural History domain. An ontological framework that characterises this type of research should be conceptually and semantically rooted in the Biological Collections Ontology (Blair et al. 2020, Walls et al. 2014) and should align with the logic schema for conventional sourcing of natural history objects (Bölling et al. 2022, Miller et al. 2020, Thomer et al. 2017) when representative whole organisms or organismal fragments are collected. A data management system based on this ontological framework should have the flexibility to track the provenance and analytical history of environmental samples, lots and collection voucher specimens using the same overarching data architecture.

The Distinction Between Data and Metadata

It is important to contextualise our ontological framework by providing semantic clarification on our use of the terms "data" and "metadata". We apply the original and currently predominant definition of the term "metadata" as "data about data" (Furner 2020). Applied to ecological datasets, metadata would, thus, be restricted to information describing the content, context, structure, quality and accessibility of data (Hampton et al. 2017, Michener 2006, Michener and Jones 2012).

Several recent works have confounded the scope of the term "metadata" to denote sampling and provenance information (e.g. ten Hoopen et al. (2022)), to define data on environmental conditions and other circumstantial parameters of field sampling (e.g. Bockrath et al. (2022), Felczykowska et al. (2015)) and to include methodological information on sampling and laboratory analyses (Nicholson et al. 2020). In our view, this approach obscures the otherwise clear semantic distinction between data and metadata or even makes it very context-based, for example, sequencing or qPCR data, vs. all other information (Abbott et al. 2021). To avoid this "terminological creep", we will refer to information about the provenance of biological materials, parameters of the collecting effort or analytical methods as different categories of data proper. For example, geocoded location or water quality measurements from the collecting station from which the samples originated would constitute eDNA-associated provenance data, but not metadata.

Main Ontological Categories

Within the context of eDNA data ontologies and within the scope of data associated with natural objects or observations, we can define three major categories characterised by the nature of data (Table 1): provenance, attributes and history. It is important to note that ontological categories should be discerned from operational entities, such as Events or MaterialSamples. Neither should these abstract groupings be conflated with specific tables in the data schema that will be discussed below. The three ontological categories defined below are best construed as classes of data that could characterised using sets of data fields within the hierarchy of tables in the data management system. As such, they constitute important dimensions or qualifiers that can help to define the operational entities and to identify their relationships within the data architecture.

Table 1.

Practical application and examples of three broad ontological categories of eDNA data (history, provenance and attributes), as they relate to the two operational entities (sampling Events and MaterialSamples).

	Event (Activity)	**MaterialSample (Biological Object)**
Provenance: **Where**? **When**? **How**?	Applies to: Spatiotemporal and circumstantial properties of the field sampling effort. Examples: sampling locality, GPS coordinates, sampling date/time, sampling method, habitat classification; molecular analytical methodology.	Applies to: Relationship to sampling effort; record of material transactions, processing and analysis. Examples: associations between lots (field samples), laboratory samples, sub-samples, aliquots etc.
Attributes: **What**?	Applies to: Qualitative or metric data pertaining to the sampling effort. Examples: sampling depth, water temperature, turbidity, weather conditions, volume of water sampled, sampling duration.	Applies to: Intrinsic or relational properties the biological materials (objects) collected. Examples: taxonomic position or biological condition of the specimen from which the sample was obtained, aliquot volume or DNA concentration.
History: **Why**? **Who**?	Applies to: Agent(s) and organisation(s) undertaking collecting/sampling activities and associated data collection. Examples: institution executing the expedition; field crew members.	Applies to: Agent(s) and organisation(s) taking custody of materials and performing processing/ analytical procedures. Examples: collection repository, collectors, analytical laboratory, sample processing technicians.

Provenance

Provenance circumscribes the spatiotemporal and circumstantial properties of the collecting or analytical events. This is the core part of the biodiversity ontology, providing details on the origin and transformations of biological objects and inferred taxonomic occurrences. Provenance data can be grouped into three broad categories of properties that describe the collecting event's localisation in space ("where?"), time ("when?") and the method used ("how?"). This information should be recorded at the time when the collecting or analytical event occurs and applies by extension to all biological objects (MaterialSamples) that are collected or produced as a result: lots, specimens, samples, aliquots and their derivatives.

Attributes

Attributes characterise intrinsic (e.g. organismal) or relational (e.g. ecological) properties of the MaterialSample ("what?") or related circumstantial properties of their origin. Unlike provenance information, which applies to an entire event and all derived materials, attributes may characterise a collection lot as a whole or may be restricted to individual biological objects or their derivatives (e.g. size of an organism or form of sample preservation). Data acquired during subsequent analysis, such as DNA concentration, sequence quality and interpretation of analytical results (e.g. presence/absence of target taxa) will fall into this category as well. Relevant information may be recorded at the time of collecting or during subsequent processing and analysis and may be stored in the form of structured data fields or file attachments. In the context of eDNA research, field-collected data may include a description and/or images of the filter containing the water sample.

History

Once a biological object is removed from nature and is transferred into human custody, it also becomes a cultural object. Historic context provides an account of agents (persons) and organisations behind the events, for example, staff undertaking the sampling activities and performing subsequent processing/analyses of MaterialSample. Thus, "historic" properties record and contextualise human interactions with biological objects, rather than their natural origin or intrinsic properties. This information provides background on the purpose of the events and overall experimental design ("why?"), the actors involved ("who?"), a record of transactions (e.g. change of ownership), processing status, storage conditions and physical location(s) of materials. It should be stressed that any information about the biological object constitutes an integral part of its research value to the scientific enterprise and, thus, by extension, of its cultural value to society at large.

Certain data types may fall into a "grey area". For example, photos taken at the collection site can be used to parametrise provenance data; however, they also depict attributes of the collecting station and/or collecting event (see below). Likewise, a scanned page from a field journal may depict provenance information, attributes of the materials collected and historic context of the collecting process.

Database Prototype

We present a prototype data management system aligned with the operational and ontological frameworks described above that implements the data architecture for environmental research design, integrates with standard field and laboratory workflows and is deployable in a typical eDNA research setting. This system facilitates the following operational needs:

- project planning and preparation (e.g. experimental design, defining unique identifiers, pre-printing field labels with ID codes);
- real-time digitisation of field sample provenance data and associated environmental characteristics;
- tracking the chain of custody and analytical history for collected biological materials and all their derivatives (e.g. DNA extracts);
- connecting each environmental sample with its analytical results;
- linking structured database records to external digital objects (files) archived in a searchable online repository of images, documents, spreadsheets and DNA sequence files.

Below is a more detailed account of the prototype database.

Overall System Requirements

To address the operational needs outlined above, a data management system for eDNA research operations should meet the following functionality requirements:

- Fieldwork and provenance data capture: facilitate real-time and/or retrospective digitisation of provenance data in a way that integrates with fieldwork operations (Events);
- Sample collection management: facilitate cataloguing, tracking, management and curation of collected biological materials (MaterialSamples) and their derivatives;
- Laboratory information management: link provenance information with downstream molecular analytical stages and facilitate tracking of the analytical history of each MaterialSample;
- Hosting and linking DNA sequences: provide a searchable repository of genomic data (sequences, primers, raw sequence files) linked to the MaterialSample and relevant provenance information;
- Taxonomic observations: provide a log of parsed taxonomic detections (Occurrence records) inferred from eDNA analyses linked to MaterialSamples, associated provenance data, analytical methodology, essential qualitative and quantitative parameters;
- Hosting and linking images and other file attachments: provide an easily accessible repository of linked images, documents, spreadsheets and other file attachments containing original raw data and metadata;

- Data validation: provide a suite of built-in validation tools to check for internal data consistency, relational integrity and alignment with external references (e.g. geocoding, taxonomy);
- External data connectivity: provide intuitive tools for efficient data conversion and exchange with other data management systems;
- Data publication and submission to global aggregators: allow for downstream compatibility with global data aggregator requirements for data publication and archival.

Below is an outline of specific technical solutions that we have developed to address these requirements.

Data Architecture

Our proposed data architecture is based upon minimum data requirements currently established for biodiversity research, with emphasis on eDNA and other genomic-derived data (Abbott et al. 2021, Finstad et al. 2023), with the understanding that universally established, structured data standards for eDNA are presently lacking (Loeza-Quintana et al. 2020). While emerging data standards centre around the needs of data aggregators and end-users, rather than data providers, the focus of this data schema is to support data management needs of eDNA research operations; therefore, it aims to reflect the relational structure and sequential pattern of their typical workflows. As mentioned earlier, this data architecture also has the capacity to accommodate other kinds of biodiversity genomic research activities (e.g. organism-based reference library building or DNA metabarcoding), to address a laboratory's essential needs for biodiversity collection curation and information management. Finally, it has the potential to incorporate additional enterprise resource planning modules, such as purchasing or shipping, if required for scaling up operational throughput.

This conceptual data framework has been implemented as a prototype eDNA Laboratory Operations Tracking Database with a MS Access ***front-end graphical user interface*** consisting of ***Forms***, ***Reports*** and ***Queries*** linked to data contained in ***back-end Tables***, which may be stored locally on the workstation running the database or, preferably, hosted as ***SharePoint Lists*** on a corporate Microsoft 365 SharePoint site. User access to data contained in these tables through the database front-end or through the SharePoint website is managed by site administrators. This set-up is easily deployable across organisations with Microsoft 365 for Business, but could also work, with proper adjustments, on a locally accessible network or on a compatible cloud server. This set-up allows real-time multi-user collaboration, without the need for file versioning or manual backups. Furthermore, it requires no additional hosting and maintenance overhead or dedicated IT infrastructure or staff to manage access, permissions and security.

Relational Structure and Primary Keys

At its core, eDNA research is the process of inferring digital genomic data from analogue biological samples. Therefore, the referential integrity of the entire research project hinges on the researcher's ability to discern individual samples and to track their derivatives through all stages of processing and analyses. Each material entity must be unambiguously associated with corresponding data records over the project's entire life cycle. Thus, establishing a proper numbering convention at the source is essential. The database prototype addresses this critical step by requiring the users to devise a robust and intuitive schema of unique, human-readable identifying codes ("IDs" or Primary Keys) for all physical and ontological entities at the inception of each project and/or experiment, a step that is often neglected with "convenience" sampling (Bockrath et al. 2022).

Many database designers (e.g. Morris (2005)) advocate using surrogate (machine-generated) Primary Keys to ensure their uniqueness within each data table and to maintain referential integrity of data records across the relational structure — an approach that has been implemented in many biological databases. However, this approach has been found less effective in facilitating relationship updates (Anonymous 2010) and in building first-order relationship queries (Link et al. 2010). We concur with this and further argue that basing relational structure of a biological research database on surrogate keys is operationally counterproductive for two important reasons.

Firstly, to avoid mixing up lots or samples in the field and/or lab, each biological object must be assigned a unique ID code (e.g. "Field ID", Sample ID", "Catalogue Number"). This is often accompanied by affixing a label with the pre-printed ID code and almost invariably pre-dates the moment when the database record is generated; hence, the surrogate Primary Key is generally unavailable when the collection object needs to be labelled. As a result, keeping accurate track of the manually-assigned ID number — and not the random surrogate key — becomes critical to ensuring data integrity.

Secondly, most machine-generated Primary Keys represent long integer numbers incrementing from 1 to infinity and are, thus, prone to overlap (not globally unique). They are only meaningful within the context of the database table where they have been generated. When migrating data between tables and/or data management systems and especially when integrating data from multiple sources into large data aggregators, such as GBIF, new surrogate keys are generated by the system, whereas original surrogate primary keys cannot be used to identify such collated records unambiguously within the new context.

By contrast, using operator-generated, or natural, Primary Keys in biological databases, while not without its challenges (Pop 2011), offers a more intuitive relational architecture and facilitates greater user engagement in building their study design around a standardised data framework. This is due to natural keys being typically based on human-readable ID strings that are established by and generally self-explanatory to users (https://www.endpointdev.com/blog/2021/03/database-design-using-natural-keys/). Practically, establishing static and unique identifiers for field records (e.g. Lot IDs) alleviates most operational challenges to relational integrity of natural keys. It also provides a more intuitive process for revising relationships between data records; for example, if a Primary Key needs to be changed during data curation (e.g. as a result syntax correction). Consequently, the relational structure in our database prototype is based on natural Primary Keys. Finally, natural Primary Keys help to maintain data integrity when records from multiple tables and/or databases are aggregated for analysis or publication.

While not intending to revisit the discussions regarding the feasibility of using persistent Globally Unique Identifiers in biological databases (e.g. Guralnick et al. (2014), Klump et al. (2021)), we strongly urge researchers and students to conceptualise a robust, human-readable numbering schema at the inception of their research project, congruent with the data relationship model that reflects their proposed study design. It is also incumbent upon lab and/or institutional leadership to facilitate and coordinate this process to ensure a consistent data management approach within their organisation. In an academic lab setting and in larger collaborative projects, it is hard to mandate and enforce a uniform syntax of globally-unique identifiers. However, it helps to have a standardised numbering schema with an intuitive and transparent logic, at least within each research group. Institution and/or network-level coordination is required to ensure that the ID number syntax does not overlap and is used consistently across research projects. This will help to avoid registry conflicts during subsequent data publication and submission to aggregators. This database prototype is intended to provide a common core data schema that adds transparency and facilitates coordination/networking within and between small eDNA operations.

Data Schema

A simplified proposed schema of key operational entities of an eDNA data management system is provided on Fig. 1. These entities are implemented in the database as relational tables accessible as forms, subforms and queries in the user interface. Some of these tables functionally overlap with Darwin Core's Event and MaterialSample categories, although none of them should be equated with these categories. A more detailed outline of the main tables and data fields is provided in Suppl. material 1. Below is an account of these tables, grouped according to their position and function in the data schema. Optional modules not implemented in the current prototype are marked with an asterisk (*).

Geographic reference

The following two data entities provide the geographic reference for the Collecting Event Core. Although field adoption of GIS-based data capture in wildlife census has been proposed early on (Travaini et al. 2007), its application in sample-based operations is far from common. The database prototype contains Visual Basic modules allowing real-time GPS data capture using MS Windows-enabled tablets, but is predominantly intended for manual data entry or batch conversion from existing records.

Sites

A **Site** is a medium-high level of geographic localisation of project activities. Using Fisheries and Oceans Canada (DFO) standard terminology (Abbott et al. 2021), it is defined as “a specific area, within a selected sample location, where water (or other environmental substrate) will be collected”. A site represents the general geographic location of the research area; therefore, it is not necessarily defined by a research project. This table also incorporates information relevant to the DFO “geographic location” and “regions” data definitions (Abbott et al. 2021). Typically, the site would be associated with a named geographic area, limited by geomorphological characteristics (e.g. water body, valley or catchment basin) or administrative boundaries (e.g. conservation area or municipality). If justified by experimental design, it may represent a more restricted research area within this geographic location. Within the proposed data architecture, several projects may use the same site, while a single project may contain field activities conducted at multiple sites.

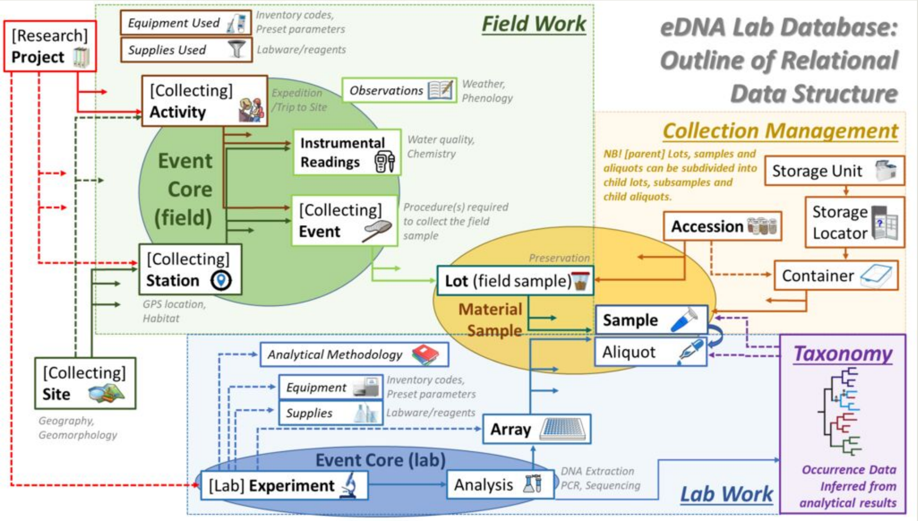

Figure 1. doi

Schematic representation of key ontological entities of an eDNA data management system.

The Site registry in the current database prototype can be cross-referenced against automatically downloadable official gazetteers of geographic localities for Canada and the United States:

the Canadian Geographical Names Database (CGNDB) provided by the Geographical Names Board of Canada – https://natural-resources.canada.ca/earth-sciences/geography/download-geographical-names-data/9245

and the USGS Geographic Names Information System (GNIS) provided by the U.S. Board on Geographic Names – https://prd-tnm.s3.amazonaws.com/index.html?prefix=StagedProducts/GeographicNames/DomesticNames/

Stations

A **Station** identifies an exact geolocated spot where field samples are taken. As per DFO terminology (Abbott et al. 2021), it “refers to spatially distinct sampling locations within a site”. It is generally research project-specific (each station is linked to only one project) and is characterised by distinct GPS coordinates, which may or may not be tied to a geographically defined landmark. Each station is further parametrised by fine-grained habitat characteristics. Multiple stations may be established within each site, whereas the same station may be surveyed one or multiple times; therefore, it may be associated with one or several sampling events. Built-in functionality within the database prototype allows real-time direct capture of geographic coordinates using a GPS-enabled Windows laptop or tablet; additional scripts further allow reverse geocoding of the geographic location of each station based on the coordinates entered. Finally, several built-in tools allow individual or batch validation of these geographic coordinates by plotting them in Open Street Map, Microsoft Maps or by generating a KML file for displaying in Google Earth.

Collecting Event Core Entities

The following database tables contain information directly related to the Darwin Core Event data class. Note that only one of them (Events proper) is directly linked to MaterialSamples, whereas the remainder are used to provide further parametrised context (Readings and Observations) and to help structure this information into the experimental logic model (Activities).

Activities

An Activity represents a series of collecting events, measurements or observations undertaken as part of a project within a specified site, usually over a restricted timespan (e.g. one to several days); for example, a field trip or short-term expedition. Each Activity is associated (unambiguously linked) to one project (through the reference Project ID) and to one site (through the reference Site ID). An Activity is carried out within a specified Collecting Site as part of a single Project.

[*Collecting] Events*

The Events table characterises the specific targetted field collecting effort that results in the acquisition of biological materials (Lot; see below) at a particular Station over a specified time interval. As the name implies, it is the key element of the event-based data management schema; it is also the key point of reference linking biological materials with their provenance information. Each Event is linked to a single parent Activity and Station. Within the eDNA research context, a typical example would be the collection of aquatic DNA on to a water filter. Each sampling replicate, repeat or replication, as per DFO definition (Abbott et al. 2021), should be recorded as a distinct Event. If successful, it would typically be associated with a single eDNA Lot (see below) and all its derivatives.

Readings (Instrumental Reads)

Many ecological sampling activities involve recording chemical, physical or other parameters of the environment (water, soil, air) at the locality where sampling occurs. These measurements are usually taken with specialised equipment, using a set of standards established as part of the study design. An example would be water quality measurements taken with a digital probe. The Readings table is designed to accommodate this information. Although often considered part of sample “metadata”, this information does not fit the strict metadata definition (see discussion above). It is not necessarily linked to any particular sampling Event; but may be indirectly associated with one or several Events through the corresponding Station and collection date.

Observations

Although sometimes used as an alternative name for the occurrence record (Lindström 2006), the Observations table is used here as a collection of optional ancillary data pertaining to any of the other tables. Observation notes generally are not part of the standard data recorded, but may affect the outcome of the analyses. Examples include phenology, weather conditions, wildlife presence etc. Unlike Instrumental Reads, an Observation may pertain to a specific Event, Site, Station or Activity on a certain date or within a specified date range. By extension, it may provide further parametrised context to all corresponding MaterialSample units.

Collection Management — MaterialSample Entities (Biological Objects)

The following tables characterise operational relationships within the MaterialSample Core data class, operationally separated into three categories: Lots (field-derived MaterialSamples), Samples (resulting from concentrating, subdividing or otherwise processing Lots at the research facility) and Aliquots (laboratory derivatives of samples destined for analysis).

Lots

The Lots table houses a registry of field-sourced biological materials (Lots) originating from a field collecting Event. As mentioned previously, the term has been co-opted from natural history collection management practice where it is used to define a set of specimens and/or samples from one or multiple organisms originating from the same collecting event that are catalogued and stored together as a single unit (Anonymous 2016), but extended here to eDNA samples taken from the environment. Each Lot must be unambiguously linked to a single collecting Event, which provides the necessary provenance context. An additional module incorporated in the database prototype allows users to register a predefined quota of Lot ID numbers before a field trip and to pre-print sticky lot labels in several common formats. If a Lot is later subdivided into sub-lots, this can be accommodated within the proposed data architecture by adding new Lot records linked to the parent Lot ID through a corresponding foreign key field.

**Specimens*

For research operations focused on building genomic reference collections linked to preserved voucher specimens (as discussed in the Operational Framework section), it may be optimal to designate a separate Specimens table within the proposed data architecture. However, for the purpose of most field-based eDNA research, the Lots table can accommodate essential provenance information on voucher specimens (e.g. opportunistically collected organisms), without the need to designate a separate data entity. The table is, therefore, not implemented in the prototype data schema.

Samples

The Samples table stores information about field- or laboratory-derived MaterialSamples prepared and preserved for archival storage and/or partitioned for laboratory analysis. In cases when DNA is filtered from the Lot preservation medium, the Sample would constitute a portion of that parent Lot; however, under many eDNA research scenarios, it may represent the same physical object as the entire Lot (e.g. DNA filter). In some cases, Samples may originate from external collaborators and not directly from the field. Samples are often grouped together into Containers (see below) for processing or storage efficiency; however, the latter should not be conflated with Lots. The table accommodates subdividing each sample into subsamples by creating new records linked to the parent Sample ID via the designated foreign key field.

Aliquots

Aliquots are laboratory derivatives of Samples: DNA extracts, PCR products etc. In many cases, these are transient substances that are used up during analyses. The purpose of an Aliquot is to identify the portion of each sample that is destined for a specific analytical pipeline (e.g. to sequence a particular gene region). Aliquots are arranged into Arrays (see below) for streamlined batch processing. Similar to Lots and Samples, sub-aliquots can be accounted for by creating new records linked to the parent Aliquot ID through a foreign key field.

MaterialSample Organisation

The following two data entities are used to organise MaterialSample units for storage, batch processing and/or analysis. Unlike MaterialSample units proper, these entities are provenance-agnostic, allowing us to aggregate materials from multiple collecting Events, Activities, Sites etc., provided that each associated MaterialSample and, if applicable, its position (e.g. processing order) within the batch are unambiguously tracked.

Containers

Containers are physical objects used to store biological materials, which can be archived, relocated or processed as a single unit. Each container can be used to house one or many biological collection items (e.g. whole Lots, Samples, Aliquots or portions thereof). Containers are designed to facilitate organisation of samples together within a processing or storage batch, their localisation within the research facility and their transfer within or outside the lab. Examples include tube racks, boxes, trays, Tupperware, removable drawers etc.

Arrays

Arrays, or processing batches, are operational (logistical) counterparts of Containers that are used to organise Aliquots or their virtual derivatives in sequential order for laboratory analyses. As such, they may be somewhat “ephemeral” as physical objects, for example, PCR plates that are used up and discarded after DNA sequencing. They may also be purely virtual, for example, batches of raw DNA sequence files run through an informatics pipeline. Additional built-in functionality in the prototype database allows the user to map aliquots within an array (processing batch) and display these maps in several common formats, for example, 12 × 8 wells in a microplate or 10 × 10 sample tubes in a rack.

Lab Work—Analytical Event

The following two data entities do not conform to any existing Darwin Core data classes; however, they are operationally essential for laboratory information management. As discussed earlier, we use the term “Analytical Event” to emphasise that they represent a separate category of “events”, which corresponds to the “identification process” category recognised within the Biological Collections Ontology (Walls et al. 2014). Together with the Collecting Event Core information, it constitutes a critical piece of information for inferring taxonomic Occurrences from the MaterialSample.

Analyses

The Analyses table provides a registry of analytical procedures and stages used in laboratory analyses, for example, DNA extraction, PCR reactions, sequencing runs etc. The prototype data schema provides for many-to-many relationships between Analyses registry and associated Arrays, thereby allowing flexibility in tracking the processing of a single Array through multiple analytical stages or assembling multiple Arrays for a single analytical procedure, for example, multiplexing several PCR plates for the same sequencing run. Analyses table fields are further parametrised by ancillary registries for target Markers, PCR Primer combinations and a Multiplexing schema to map the Aliquots used in Next-Generation sequencing runs.

Experiments

Experiments represent sets of Analyses aimed at a particular research goal; for example, grouping together sets of Analyses that use the same protocols. As such, they represent abstract entities used to facilitate operational logistics and may be placed in the category below.

Research Administration

The following tables serve to facilitate overall operational logistics of research projects, thereby parametrising the historic component of the ontological framework. Portions of the information contained in these tables fall into the metadata category, as related to Event and MaterialSample Core components.

Projects

The Project table houses records pertaining to the logistics of administration and/or management of research and survey activities; therefore, it is not subordinate to any other database module. Projects are registered before the beginning of any related field activities or laboratory experiments. All activities, Collecting Event Core and MaterialSample Core tables are associated with the respective Project by linking each of them to the corresponding Project ID. However, Projects may have a many-to-many relationship with laboratory Analyses and Experiments, depending on experimental design and laboratory management logistics. As such, the Project may be considered as the logistical counterpart of the Experiment.

Agents

The Agents table hosts names, institutional affiliations and contact details of persons recorded in other database tables (collectors, data recorders, processing staff, collaborators, project managers, expedition leads etc.). Information from this table is linked to agent drop-down menus available in other tables.

Organizations

The Organizations table holds information about institutions, laboratories, companies and other organisations affiliated with or responsible for different projects, experiments, corresponding activities and analytical stages.

Accessions

The Accessions table is adopted from biological collection management practices (Berendsohn et al. 2011, Miller et al. 2020) to document batches of biological materials (e.g. Lots or Samples) received together on the same day under the auspices of the same project. Information contained within an accession record thus extends to each associated MaterialSample (Lot, Sample and all derivatives). While bearing no direct relevance to research design, accessions are important for tracking the administrative aspects of biological material transactions, such as ownership, destination, mutually agreed terms and applicable restrictions on analysis or data publication. Tracking this information is particularly important if biological materials are sourced from another institution and, especially, from another country.

**Loans*

Although not currently implemented in the prototype database, loans (batches of biological materials dispatched to external users) constitute an important component of collection management logistics (Berendsohn et al. 2011). If a laboratory plans to engage in biological material transactions with third parties, it should consider incorporating a separate registry of loans in its data management strategy. The corresponding change can be easily implemented in the database prototype.

[*Storage] Units*

Storage **Units** are items of furniture and/or equipment used for material storage (freezers, refrigerators, cabinets, shelving units etc.). Typically, they have a fixed location in a specific building, floor, room etc. within an organisation. Each Storage Unit is linked to multiple Storage Locators (see below).

[*Storage] Locators*

Storage **Locators** are fixed compartments within Storage Units housing various physical or biological objects, specifically, collection items (Containers with Lots, Samples or Aliquots). Examples include fixed drawers, shelves or slots within freezers, shelving units or storage cabinets. Locators are important in ensuring that biological materials housed and processed by lab members can be easily found and tracked within the laboratory or collection facility. Locators have a one-to-many relationship with storage Containers and, by extension, with all associated Lots, Samples and/or Aliquots.

Equipment

Most research activities use specialised equipment, which may impact field collecting and analytical outcomes. An **Equipment** inventory helps to control for biases that may be introduced by using generic equipment types (e.g. technical specifications of different brands of eDNA samplers) or particular equipment items (e.g. working condition or calibration). Individual Events, Instrumental Reads and Analyses could be linked to utilised Equipment items through dedicated foreign key fields. Depending on laboratory setting, this module could be further parametrised by adding separate registries of calibration, maintenance or sign-out for use by laboratory staff and/or external collaborators.

**Supplies*

Basic information on standard **Supplies** used in particular Events (e.g. eDNA filters) and Analyses (e.g. PCR reagents) is incorporated within the respective Events, Analyses and other data tables. For larger-scale operations, it may be useful to establish a separate registry of supplies and/or reagents that would allow evaluating the relative performance of separate supply batches or reagent stocks over time. Although not implemented in the prototype database, this data module could be added and further parametrised by logging accrued stock and its use for field or laboratory work, linked to Activities, Events, Experiments or Analyses. It could also be integrated with other enterprise resource planning modules, such as a registry of purchase orders. Such modules could be custom-built or adopted from existing off-the-shelf enterprise resource planning solutions.

Ancillary Data Modules

The following tables provide annotations for files associated with existing data records that are either unstructured (e.g. raster images) or cannot be adequately parsed and incorporated into existing data fields without significant information loss (e.g. original Excel tables). Each data entry includes a reference (foreign key) linking it to the "parent" record (e.g. collecting Event ID) and an absolute URL to the online resource where the file is hosted. By default, attachment files are named in a self-explanatory way (i.e. by incorporating the foreign key into the file name) and are hosted in a designated folder on SharePoint or other cloud server that ensures reliable data hosting for the project's life cycle. This allows effective retrieval of external files associated with each database record, as well as direct browsing through data folders on the cloud server and, as necessary, batch processing, backup or migration of these files. Built-in database functionality allows the user to perform batch renaming of files and automated generation of links, based on a set of standard algorithms.

Attachments

The **Attachments** table provides annotation for generic file attachments, such as documents, digital images or collaborator-provided Excel spreadsheets. Attachments may be linked to any record in any of the core tables within the prototype data schema using their primary ID as a foreign key. If the same primary ID is used to identify records in two or more different tables (e.g. if the syntax of the collecting Event ID is identical to the derived Lot ID), then the same attachment file (e.g. photo of the collected water filter) is linked to all corresponding database records. Hosting these files in dedicated folders on the database SharePoint server allows direct batch viewing and download through the online SharePoint interface or using OneDrive file manager applications.

Sequences

The **Sequences** table is used specifically to annotate DNA sequence file attachments (e.g. FASTA and FASTQ files) linked to records of individual Aliquots from which they have been generated. In addition to providing links to the Aliquot ID and file URL, this table includes fields that provide additional parametrisation, in line with the Darwin Core's AssociatedSequences (https://dwc.tdwg.org/list/#dwc_associatedSequences) and related fields. While the database prototype does not offer built-in functionality for analysing stored sequence data files, it facilitates their direct download and processing using external software applications.

Protocols

The **Protocols** table provides a registry of Analytical Protocols and SOPs used in the organisation's research operations linked to Collecting and Analytical Event Core modules. This module currently provides only basic annotation functionality; however, it offers potential for future parametrisation of research outcomes by adding custom tables with project- or laboratory-specific qualitative or quantitative metrics that vary, according to the collection or analytical protocols selected.

Taxonomic Annotation and Occurrence Records

Several additional tables and fields within the prototype data schema allow basic taxonomic annotation for the MaterialSample (Lot, Sample or Aliquot), including modules that validate the taxonomy used against existing taxonomic references (currently, GBIF and NCBI taxonomy). The **NGS_Taxonomy** table provides a detailed breakdown of taxonomic occurrence records inferred from analysing Aliquot-derived raw sequence data using different informatics pipelines. By extension, these results are linked to field-sourced Lots with associated Collecting Events and other provenance information. They are also linked to laboratory-assembled arrays and associated analytical protocol parameters (Analytical Events), allowing us to backtrace the field provenance and/or methodological and procedural origin of each taxonomic occurrence record.

Additional Database Features and Best Practices

The overarching goal of the data management system that supports eDNA research and other work based on analysing biological materials is to ensure that each analogue MaterialSample is unambiguously linked to its corresponding Event digital data record and that all information pertaining to its provenance, attributes and history is accurately captured and parsed in real time and in adequate detail. To meet these requirements, the process of data capture should be integrated with research operations in a way that minimises additional databasing effort and provides immediate incentives to the person(s) recording the data. This could be achieved through workflow optimisation (e.g. sequential structuring of operational and data entry phases), automation (batch file renaming/linking, direct instrumental input) or procedural guidance (integration of pop-up SOPs and checklists into the user interface).

Some of this functionality has been implemented as a suite of data management tools and modules in the prototype database; however, the feasibility of their practical deployment will depend on the specifics of user organisations, their infrastructure, workforce and research settings. Below, we outline some basic principles of how the proposed data architecture could be used to address the data management needs during different research phases and suggest best practices for streamlining the process and increasing data quality.

Mapping the Data Schema Against the Operational Framework

Table 2 provides a rough breakdown of what happens to biological materials (MaterialSample) and associated data during typical phases of the research project. It is intended to provide context for best practices outlined below.

Pre-printed MaterialSample Labels

A good practice with respect to ensuring the uniqueness of the identifiers used as primary keys (e.g. Lot ID or Container ID numbers) is to generate them in advance of a field trip or experiment using a dedicated module of the data management system. This will ensure both uniqueness and accuracy of the syntax used for any given activity and will also “preoccupy” this syntax pattern and not allow it to be registered accidentally by another user or field crew.

Pre-printed Analogue Field Data Journals

We have developed an MS Excel template that could be pre-filled and printed in a 4.625” x 7” weatherproof 5-inch binder format where core blocks of data and data fields are structured similarly to the database, allowing for subsequent manual database entry from hand-filled templates. The template is organised as a set of predefined forms, rather than as a non-relational spreadsheet. Each form mirrors the operating procedure performed in the field: arriving on site, confirming the location, identifying and characterising sampling stations, performing water quality tests, sample collection and recording ancillary observations.

Table 2.

Main workflow stages involved in MaterialSample-based research and their relationship to the sample, associated data and corresponding database tables in the data management framework. Asterisks (*) mark optional tables of potential use for collection repositories that were not implemented in the prototype database.

Research Phases	**MaterialSample**	**Associated Data**	**Relevant Data Entities (Database Tables)**
Field Collecting	Field sourcing (collecting), labelling of biological materials.	Assignment of unique Lot/Specimen identifiers, field capture of provenance data (geospatial information, observations, instrumental readings and metadata).	Sites Stations Activities Collecting Events Instrumental Reads Observations Lots Specimens*
Pre-lab Processing and Preparation	Preservation, sorting, labelling of biological materials; subsampling and/or preparation of (sub)samples for analysis.	Recording associations between Lots, Specimens, Samples and Aliquots and aggregating them into corresponding Container and/or Array records.	Lots Specimens* Samples Aliquots Containers Arrays

Research Phases	**MaterialSample**	**Associated Data**	**Relevant Data Entities (Database Tables)**
Laboratory Analyses	Analytical procedures to detect target DNA signatures and reconstruct taxonomic position/ taxonomic lists.	Tracking and digitisation of laboratory analytical procedures (lab books, LIMS etc.).	Aliquots Arrays Experiments Analyses Protocols
Post-laboratory Informatics Analysis	Not applicable	Informatics analysis of qPCR and/or DNA sequencing data, including quality scoring, demultiplexing, NGS pipelines, taxonomic queries.	Aliquots Arrays Experiments Analyses Protocols Taxonomy Sequences
Transfer/ Acquisition	Movement of biological materials between organisations and/or agents.	Data migration between management systems, material transfer agreements, accessioning by the recipient.	Lots Specimens* Samples Aliquots Containers Arrays Accessions Loans*
Archival/ Deposition	Long-term preservation of materials for potential future re-examination and/or analysis.	Data upload/archival in collection database.	Locators Storage Units
Data Publication	Not applicable	Batch data query and conversion into data packages and/or data submission spreadsheets formatted to the requirements of the publisher or data repository.	All tables (potentially)

Data Connectivity and Conversion

The database prototype offers a suite of pre-defined MS Excel templates to assist users with standardised field data capture, batch data conversions and validation. The tools are being constantly updated to address emerging user needs. Several modules currently available or under development are listed below:

- Batch geographic coordinate conversions from degree, minute, second (DMS) and Universal Transverse Mercator (UTM) formats to decimal degrees;
- Batch reverse geocoding script to provide geographic annotations for sets of coordinates;
- Batch data converter from an MS Excel generic non-relational field data template;
- Batch data converter from an MS Excel standard non-relational data template used by the GEN-FISH network — https://gen-fish.ca/;
- Batch data converter from Excel table output from the Sample and Field data collection Information System for the Hanner Lab (sFISH) prototype mobile field app — https://github.com/HannerLab/sFish;

- Batch data converter from the database to the Molecular Detection Mapping and Analysis Platform for R (MDMAPR) Shiny web app (Yu et al. 2020);
- Batch Excel conversion file for manual 96-well plate array assembly from isolated samples;
- Batch data converter from custom Next-Generation Sequencing pipeline outputs — under development.

Automation and Integration

Under an ideal scenario, most metric data should be digitised in the field and in the lab through direct instrumental input, by feeding the digital output from measuring devices (e.g. water quality probes) and analytical instruments (e.g. DNA sequencers) into the corresponding data tables. In practice, this is not always logistically feasible and very rarely implemented, especially in remote field settings. The database prototype has built-in functionality that allows some basic data manipulations; however, it presently offers limited support for direct instrumental input. For example, it can capture the geolocation of the device running the database using its Wi-Fi, cellular connection or built-in GPS receiver. One of the logistical bottlenecks identified by beta-users of the database prototype is batch renaming and annotated archival of images associated with field collecting Stations, Events, Lots and Samples. This requirement has been addressed for database installations run on MS Windows tablets using the Microsoft Camera app. Identifying other priority areas of development, based on user input, is key to improving the database's operational utility for small lab applications.

General Data Maintenance and Validation

Data management systems benefit from active engagement of users in the process of their development (Glöckler et al. 2020). This engagement is particularly important for new databasing projects aiming to facilitate emerging and actively developing research areas, such as eDNA. Active input is sought from current and prospective database users and collaborators to strengthen the system's ontological and semantic conceptual framework, to review and improve its architecture and data schema, to improve its user interface, built-in tools and additional functionality.

Continued curation and management are essential for maintaining a database's utility over time (Blair et al. 2020). Although the database and its additional data conversion tools are designed to maximise efficiency in data upload and validation, its use requires a significant time commitment to familiarise with the user interface and custom functions and continued attention towards data curation.

Sustained efforts should be devoted towards building robust, standardised, logically consistent and intuitively comprehensible naming conventions for natural Primary Keys used throughout the data management system, especially when digital records refer to analogue biological objects that are being collected, stored and analysed.

Unstructured and Analogue Data Curation

Curation of unstructured and/or analogue data (e.g. images, hand-written field notes) requires digital capture of representative data files (e.g. photos or scans), which are then appended to core database records as annotated attachments. Associated metadata, if available, could be used to parametrise such files. As mentioned earlier, the database prototype allows storing and annotating diverse types of file attachments; however, detailed user input and continued curation are required to ensure that archived files remain properly organised, referenced and readily accessible through individual database records or directly from the hosting server.

Making eDNA Data Findable, Accessible, Interoperable and Reusable (FAIR)

To date, robust standards have been developed for biodiversity data (Berendsohn et al. 1999, Chapman et al. 2020) and specifically for DNA-derived occurrence data (e.g. Finstad et al. (2023)); and best practices have been established for making them findable, accessible, interoperable and reusable, or FAIR (Reyserhove et al. 2020, Wilkinson et al. 2016). Tools are being developed to facilitate adherence of eDNA data to FAIR principles (e.g. Kimble et al. (2022)), with similar developments underway addressing taxonomic occurrence and biodiversity data at large (Sandall et al. 2022, Reyserhove et al. 2020). Unfortunately, there is currently limited awareness and/or slow uptake of these principles amongst field and laboratory biologists. In practice, adherence to these standards is difficult, particularly for small research groups lacking integrative data management systems and dedicated databasing staff. We hope that the proposed data management solution will be a step towards facilitating the generation of FAIR eDNA data records. Amongst the priority areas for future development of the database prototype are further mapping of the data schema on to Darwin Core and the development of a semi-automated data submission pipeline from the database to GBIF and other biodiversity data aggregators.

General Conclusions

The proposed data management system aims to address the basic, yet specialised needs of eDNA data tracking that have been identified through extensive consultations with our colleagues engaged in this research. As eDNA is an actively developing field with emerging methodological standards, there is a need for structural flexibility of the data schema that could accommodate data management to support academic research and development. At the same time, eDNA's potential for planning and regulatory applications (Anonymous 2020, De Brauwer et al. 2023, McDonald et al. 2020) demands robust standardisation of its core data elements. Our approach attempts to strike this balance by providing a rigid baseline data schema rooted in an existing (rather than *ad hoc*) Biological Collections Ontology, while providing a modular structure of data tables that could facilitate detailed parametrisation of methodological approaches that may be specific to individual experimental designs. This should allow eDNA researchers to integrate their results seamlessly into a larger body of biodiversity occurrence data, while retaining important details needed to drive further methodological advancements in their specialised field.

From a technical aspect, the format of any software applications/databases used for data management and archival should be non-proprietary and data schema should be intuitive enough to allow migrating datasets in their entirety from one system to another, for example, as may be necessitated by database software becoming obsolete or cloud storage providers going out of service. This is particularly important for image and raw data archives (e.g. FASTQ files) associated with database records, which must remain directly accessible for batch download or transfer, while retaining their association with the corresponding data and metadata records, for example, through robust and transparent file-naming conventions.

We should emphasise that the publication of aggregated eDNA-derived taxonomic observations, however important, cannot be regarded as an adequate substitute for proper archival of complete, properly referenced and parametrised datasets by the organisations from where they have been generated. When possible, such comprehensive data archives should be backed by properly stored and curated biological samples from which the eDNA originated. The quality of the data and samples, thus archived, requires an initial investment in relevant staffing and infrastructure and further depends on a continued commitment to maintaining the accuracy and accessibility of biological materials and data records. This may be particularly hard to achieve for small-scale research operations and time-restricted surveys or monitoring projects. Their specific challenges and essential role in human understanding of planetary health across time should be more broadly acknowledged and addressed by relevant administrators, regulators and funders.

Finally, we hope that this paper will help to draw the attention of researchers to the importance of further harmonising data strategies for eDNA research with those established for more “traditional” approaches to surveying and monitoring biodiversity.

Acknowledgements

The data management system for eDNA research presented here was developed, based on the needs identified by and with support from varied large-scale research partnerships including Agnico Eagle Gold Corp, the Canadian Food Inspection Agency (via the Federal Assistance Partnership), GEN-FISH (funded by Genome Canada and the Ontario Genomics Institute; OGI-184), the Great Lakes Fisheries Commission (via the “Field-ready environmental DNA (eDNA) protocols and tools for sea lamprey assessment” project) and the Nuclear Waste Management Organisation. We also thank our colleagues for their detailed input on critical data elements that need to be captured during different stages of eDNA research and for testing the application prototypes developed during this study.

The following members of the University of Guelph research team and associates offered helpful insights into their research workflows and provided user feedback on the initial database prototype: Kate Lindsay, Erika Myler, Cameron Brown, Kathleen Nolan, Nathan

Zeinstra, Kayley Head, Morgan Humphrey, Liam Lalonde, Abinaya Yogasekaram, Ian Murphy, Danielle Bourque, Tzitziki Loeza Quintana, Yoamel Milian Garcia, Sarah Adamowicz.

Further input on eDNA data management requirements and the usability of the database prototype was provided by the GEN-FISH team and associates during the deployment and testing of the database at the Great Lakes Institute for Environmental Research, University of Windsor: Matthew Yates, Joe Branget, Jonathon Leblanc, Alex Van Nynatten, Paige Breault, Keta Patel, Mohammed Zain.

As well, we thank Jarrett Phillips, Morgan Humphrey and Kayley Head for their constructive comments on the earlier drafts of the paper, Daniel Mietchen for editorial remarks and Rolando Blanco and for reviewing the submitted manuscript.

Publication of the article was supported by the first author's affiliation as Departmental Associate of the Royal Ontario Museum.

Hosting institution

University of Guelph

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Abbott C, Coulson M, Gagné N, Lacoursière-Roussel A, Parent GJ, Bajno R, Dietrich C, May-McNally S (2021) Guidance on the Use of Targeted Environmental DNA (eDNA) Analysis for the Management of Aquatic Invasive Species and Species at Risk. DFO Can. Sci. Advis. Sec. Res. Doc. 2021/019. iv + 42 p. URL: https://publications.gc.ca/collections/collection_2021/mpo-dfo/fs70-5/Fs70-5-2021-019-eng.pdf
- Aleksic S, Celikovic M, Link S, Lukovic I, Mogin P (2010) Faceoff: Surrogate vs. Natural Keys. Advances in Databases and Information Systems. 6295. [ISBN 978-3-642-15575-8 978-3-642-15576-5]. https://doi.org/10.1007/978-3-642-15576-5_41
- Anonymous (2016) Guidance for Cataloging Department of the Interior Museum Collections. Unites States Department of the Interior. URL: https://www.doi.gov/sites/doi.gov/files/uploads/museum_cataloging_guidance_march_2016_fnl.pdf
- Anstey P (2012) Francis Bacon and the Classification of Natural History. Early Science and Medicine 17 (1-2): 11-31. https://doi.org/10.1163/157338212X631765
- Bani A, De Brauwer M, Creer S, Dumbrell A, Limmon G, Jompa J, Von Der Heyden S, Beger M (2020) Informing marine spatial planning decisions with environmental DNA. Advances in Ecological Research. 62. [ISBN 978-0-12-821134-2]. https://doi.org/10.1016/bs.aecr.2020.01.011

- Barnes M, Turner C (2016) The ecology of environmental DNA and implications for conservation genetics. Conservation Genetics 17 (1): 1‑17. https://doi.org/10.1007/s10592-015-0775-4
- Belle C, Stoeckle B, Geist J (2019) Taxonomic and geographical representation of freshwater environmental DNA research in aquatic conservation. Aquatic Conservation: Marine and Freshwater Ecosystems 29 (11): 1996‑2009. https://doi.org/10.1002/aqc.3208
- Bell K, Fowler J, Burgess K, Dobbs E, Gruenewald D, Lawley B, Morozumi C, Brosi B (2017) Applying Pollen DNA Metabarcoding to the Study of Plant–Pollinator Interactions. Applications in Plant Sciences 5 (6). https://doi.org/10.3732/apps.1600124
- Beng K, Corlett R (2020) Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. Biodiversity and Conservation 29 (7): 2089‑2121. https://doi.org/10.1007/s10531-020-01980-0
- Berendsohn W, Anagnostopoulos A, Hagedorn G, Jakupovic J, Nimis PL, Valdés B, Güntschl A, Pank-hurst R, White R (1999) A comprehensive reference model for biological collections and surveys. TAXON 48 (3): 511‑562. https://doi.org/10.2307/1224564
- Berendsohn W, Güntsch A, Hoffmann N, Kohlbecker A, Luther K, Müller A (2011) Biodiversity information platforms: From standards to interoperability. ZooKeys 150: 71‑87. https://doi.org/10.3897/zookeys.150.2166
- Berry O, Jarman S, Bissett A, Hope M, Paeper C, Bessey C, Schwartz M, Hale J, Bunce M (2021) Making environmental DNA (eDNA) biodiversity records globally accessible. Environmental DNA 3 (4): 699‑705. https://doi.org/10.1002/edn3.173
- Blair J, Gwiazdowski R, Borrelli A, Hotchkiss M, Park C, Perrett G, Hanner R (2020) Towards a catalogue of biodiversity databases: An ontological case study. Biodiversity Data Journal 8 https://doi.org/10.3897/BDJ.8.e32765
- Bockrath K, Maloy AP, Rees C, Tuttle-Lau M, Mize E (2022) Environmental DNA (eDNA) Best Management Practices for Project Planning, Deployment, and Application. Unpublished policy document https://doi.org/10.13140/RG.2.2.12200.96006
- Bohan D, Vacher C, Tamaddoni-Nezhad A, Raybould A, Dumbrell A, Woodward G (2017) Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks. Trends in Ecology & Evolution 32 (7): 477‑487. https://doi.org/10.1016/j.tree.2017.03.001
- Bölling C, Bilkhu S, Gendreau C, Glöckler F, Macklin J, Shorthouse D (2022) Representation of Object Provenance for Research on Natural Science Objects: Samples, parts and derivatives in DINA-compliant collection data management. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.94531
- Borisenko A, Sones J, Hebert P (2009) The front-end logistics of DNA barcoding: challenges and prospects. Molecular Ecology Resources 9: 27‑34. https://doi.org/10.1111/j.1755-0998.2009.02629.x
- Bruce K, Blackman R, Bourlat S, Hellström AM, Bakker J, Bista I, Bohmann K, Bouchez A, Brys R, Clark K, Elbrecht V, Fazi S, Fonseca V, Hänfling B, Leese F, Mächler E, Mahon A, Meissner K, Panksep K, Pawlowski J, Schmidt Yáñez P, Seymour M, Thalinger B, Valentini A, Woodcock P, Traugott M, Vasselon V, Deiner K (2021) A practical guide to DNA-based methods for biodiversity assessment. Pensoft Publishers https://doi.org/10.3897/ab.e68634

- Buckner JC, Sanders RC, Faircloth BC, Chakrabarty P (2021) The critical importance of vouchers in genomics. eLife 10 https://doi.org/10.7554/eLife.68264
- Buneman P, Khanna S, Tan W (2000) Data Provenance: Some Basic Issues. FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science. 1974. [ISBN 978-3-540-41413-1 978-3-540-44450-3]. https://doi.org/10.1007/3-540-44450-5_6
- Chapman A, Belbin L, Zermoglio P, Wieczorek J, Morris P, Nicholls M, Rees ER, Veiga A, Thompson A, Saraiva A, James S, Gendreau C, Benson A, Schigel D (2020) Developing Standards for Improved Data Quality and for Selecting Fit for Use Biodiversity Data. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.50889
- Costello M, Wieczorek J (2014) Best practice for biodiversity data management and publication. Biological Conservation 173: 68-73. https://doi.org/10.1016/j.biocon.2013.10.018
- Culley T (2013) Why Vouchers Matter in Botanical Research. Applications in Plant Sciences 1 (11). https://doi.org/10.3732/apps.1300076
- De Brauwer M, Clarke L, Chariton A, Cooper M, de Bruyn M, Furlan E, MacDonald A, Rourke M, Sherman CH, Suter L, Villacorta-Rath C, Zaiko A, Trujillo-González A (2023) Best practice guidelines for environmental DNA biomonitoring in Australia and New Zealand. Environmental DNA 5 (3): 417-423. https://doi.org/10.1002/edn3.395
- Farrell JA, Whitmore L, Duffy DJ (2021) The Promise and Pitfalls of Environmental DNA and RNA Approaches for the Monitoring of Human and Animal Pathogens from Aquatic Sources. BioScience 71 (6): 609-625. https://doi.org/10.1093/biosci/biab027
- Fediajevaite J, Priestley V, Arnold R, Savolainen V (2021) Meta-analysis shows that environmental DNA outperforms traditional surveys, but warrants better reporting standards. Ecology and Evolution 11 (9): 4803-4815. https://doi.org/10.1002/ece3.7382
- Felczykowska A, Krajewska A, Zielińska S, Łoś J (2015) Sampling, metadata and DNA extraction - important steps in metagenomic studies. Acta Biochimica Polonica 62 (1): 151-160. https://doi.org/10.18388/abp.2014_916
- Finstad AG, Andersson A, Bissett A, Fossøy F, Grosjean M, Hope M, Kõljalg U, Lundin D, Nilsson H, Prager M, Jeppesen TS, Svenningsen C, Schigel D, Abarenkov K, Provoost P, Suominen S, Frøslev TG (2023) Publishing DNA-derived data through biodiversity data platforms. v1.3. Copenhagen: GBIF Secretariat https://doi.org/10.35035/DOC-VF1A-NR22
- Furner J (2020) Definitions of “Metadata”: A Brief Survey of International Standards. Journal of the Association for Information Science and Technology 71 (6). https://doi.org/10.1002/asi.24295
- Gibson J, Shokralla S, Porter T, King I, Van Konynenburg S, Janzen D, Hallwachs W, Hajibabaei M (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proceedings of the National Academy of Sciences 111 (22): 8007-8012. https://doi.org/10.1073/pnas.1406468111
- Glöckler F, Macklin J, Shorthouse D, Bölling C, Bilkhu S, Gendreau C (2020) DINA—Development of open source and open services for natural history collections & research. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59070

- Guralnick R, Conlin T, Deck J, Stucky B, Cellinese N (2014) The Trouble with Triplets in Biodiversity Informatics: A Data-Driven Case against Current Identifier Practices. PLoS ONE 9 (12). https://doi.org/10.1371/journal.pone.0114069
- Hackett R, Belitz M, Gilbert E, Monfils A (2019) A data management workflow of biodiversity data from the field to data users. Applications in Plant Sciences 7 (12). https://doi.org/10.1002/aps3.11310
- Hampton S, Jones M, Wasser L, Schildhauer M, Supp S, Brun J, Hernandez R, Boettiger C, Collins S, Gross L, Fernández D, Budden A, White E, Teal T, Labou S, Aukema J (2017) Skills and Knowledge for Data-Intensive Environmental Research. BioScience 67 (6): 546-557. https://doi.org/10.1093/biosci/bix025
- Hanner R, Gregory TR (2007) Genomic Diversity Research and the Role of Biorepositories. Cell Preservation Technology 5 (2): 93-103. https://doi.org/10.1089/cpt.2007.9993
- Hardisty A, Michener W, Agosti D, Alonso García E, Bastin L, Belbin L, Bowser A, Buttigieg PL, Canhos DL, Egloff W, De Giovanni R, Figueira R, Groom Q, Guralnick R, Hobern D, Hugo W, Koureas D, Ji L, Los W, Manuel J, Manset D, Poelen J, Saarenmaa H, Schigel D, Uhlir P, Kissling WD (2019) The Bari Manifesto: An interoperability framework for essential biodiversity variables. Ecological Informatics 49: 22-31. https://doi.org/10.1016/j.ecoinf.2018.11.003
- Hinz S, Coston-Guarini J, Marnane M, Guarini J (2022) Evaluating eDNA for Use within Marine Environmental Impact Assessments. Journal of Marine Science and Engineering 10 (3). https://doi.org/10.3390/jmse10030375
- Hoban S, Archer F, Bertola L, Bragg J, Breed M, Bruford M, Coleman M, Ekblom R, Funk WC, Grueber C, Hand B, Jaffé R, Jensen E, Johnson J, Kershaw F, Liggins L, MacDonald A, Mergeay J, Miller J, Muller-Karger F, O'Brien D, Paz-Vinas I, Potter K, Razgour O, Vernesi C, Hunter M (2022) Global genetic diversity status and trends: towards a suite of Essential Biodiversity Variables (EBVs) for genetic composition. Biological Reviews 97 (4): 1511-1538. https://doi.org/10.1111/brv.12852
- Hoffmann P (1994) General Aspects of Environmental Sampling. Environmental Sampling for Trace Analysis. [ISBN 978-3-527-61587-2 978-3-527-30051-8]. https://doi.org/10.1002/9783527615872.ch2
- Kelly R, Lodge D, Lee K, Theroux S, Sepulveda A, Scholin C, Craine J, Andruszkiewicz Allan E, Nichols K, Parsons K, Goodwin K, Gold Z, Chavez F, Noble R, Abbott C, Baerwald M, Naaum A, Thielen P, Simons AL, Jerde C, Duda J, Hunter M, Hagan J, Meyer RS, Steele J, Stoeckle M, Bik H, Meyer C, Stein E, James K, Thomas A, Demir-Hilton E, Timmers M, Griffith J, Weise M, Weisberg S (2023) Toward a national eDNA strategy for the United States. Environmental DNA https://doi.org/10.1002/edn3.432
- Kilian N, Henning T, Plitzner P, Müller A, Güntsch A, Stöver B, Müller K, Berendsohn W, Borsch T (2015) Sample data processing in an additive and reproducible taxonomic workflow by using character data persistently linked to preserved individual specimens. Database 2015 https://doi.org/10.1093/database/bav094
- Kimble M, Allers S, Campbell K, Chen C, Jackson LM, King BL, Silverbrand S, York G, Beard K (2022) medna-metadata: an open-source data management system for tracking environmental DNA samples and metadata. Bioinformatics 38 (19): 4589-4597. https://doi.org/10.1093/bioinformatics/btac556
- Kissling WD, Ahumada J, Bowser A, Fernandez M, Fernández N, García EA, Guralnick R, Isaac NB, Kelling S, Los W, McRae L, Mihoub J, Obst M, Santamaria M, Skidmore A,

Williams K, Agosti D, Amariles D, Arvanitidis C, Bastin L, De Leo F, Egloff W, Elith J, Hobern D, Martin D, Pereira H, Pesole G, Peterseil J, Saarenmaa H, Schigel D, Schmeller D, Segata N, Turak E, Uhlir P, Wee B, Hardisty A (2018) Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biological Reviews 93 (1): 600-625. https://doi.org/10.1111/brv.12359

- Klump J, Lehnert K, Ulbricht D, Devaraju A, Elger K, Fleischer D, Ramdeen S, Wyborn L (2021) Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number. Data Science Journal 20 https://doi.org/10.5334/dsj-2021-033
- Kõljalg U, Abarenkov K, Zirk A, Runnel V, Piirmann T, Pöhönen R, Ivanov F (2019) PlutoF: Biodiversity data management platform for the complete data lifecycle. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37398
- Leese F, Altermatt F, Bouchez A, Ekrem T, Hering D, Meissner K, Mergen P, Pawlowski J, Piggott J, Rimet F, Steinke D, Taberlet P, Weigand A, Abarenkov K, Beja P, Bervoets L, Björnsdóttir S, Boets P, Boggero A, Bones A, Borja Á, Bruce K, Bursić V, Carlsson J, Čiampor F, Čiamporová-Zatovičová Z, Coissac E, Costa F, Costache M, Creer S, Csabai Z, Deiner K, DelValls Á, Drakare S, Duarte S, Eleršek T, Fazi S, Fišer C, Flot J, Fonseca V, Fontaneto D, Grabowski M, Graf W, Guðbrandsson J, Hellström M, Hershkovitz Y, Hollingsworth P, Japoshvili B, Jones J, Kahlert M, Kalamujic Stroil B, Kasapidis P, Kelly M, Kelly-Quinn M, Keskin E, Kõljalg U, Ljubešić Z, Maček I, Mächler E, Mahon A, Marečková M, Mejdandzic M, Mircheva G, Montagna M, Moritz C, Mulk V, Naumoski A, Navodaru I, Padisák J, Pálsson S, Panksep K, Penev L, Petrusek A, Pfannkuchen M, Primmer C, Rinkevich B, Rotter A, Schmidt-Kloiber A, Segurado P, Speksnijder A, Stoev P, Strand M, Šulčius S, Sundberg P, Traugott M, Tsigenopoulos C, Turon X, Valentini A, van der Hoorn B, Várbíró G, Vasquez Hadjilyra M, Viguri J, Vitonytė I, Vogler A, Vrålstad T, Wägele W, Wenne R, Winding A, Woodward G, Zegura B, Zimmermann J (2016) DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. Research Ideas and Outcomes 2 https://doi.org/10.3897/rio.2.e11321
- Leray M, Knowlton N (2015) DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Sciences 112 (7): 2076-2081. https://doi.org/10.1073/pnas.1424997112
- Lewin H, Robinson G, Kress WJ, Baker W, Coddington J, Crandall K, Durbin R, Edwards S, Forest F, Gilbert MTP, Goldstein M, Grigoriev I, Hackett K, Haussler D, Jarvis E, Johnson W, Patrinos A, Richards S, Castilla-Rubio JC, van Sluys M, Soltis P, Xu X, Yang H, Zhang G (2018) Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences 115 (17): 4325-4333. https://doi.org/10.1073/pnas.1720115115
- Lindström J (2006) Database Model for Taxonomic and Observation Data. Proceedings of the 2nd IASTED International Conference on Advances in Computer Science and Technology. [ISBN 0-88986-545-0].
- Link S, Luković I, Mogin P (2010) Performance Evaluation of Natural and Surrogate Key Database Architectures. URL: https://ecs.wgtn.ac.nz/foswiki/pub/Main/TechnicalReportSeries/ECSTR10-06.pdf
- Liu M, Clarke L, Baker S, Jordan G, Burridge C (2020) A practical guide to DNA metabarcoding for entomological ecologists. Ecological Entomology 45 (3): 373-385. https://doi.org/10.1111/een.12831

- Loeza-Quintana T, Abbott C, Heath D, Bernatchez L, Hanner R (2020) Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS)—Advancing collaboration and standardization efforts in the field of eDNA. Environmental DNA 2 (3): 255-260. https://doi.org/10.1002/edn3.112
- Martin NA (1990) Voucher specimens: A way to protect the value of your research. Biology and Fertility of Soils 9 (2): 93-94. https://doi.org/10.1007/BF00335789
- Mathew C, Güntsch A, Obst M, Vicario S, Haines R, Williams A, de Jong Y, Goble C (2014) A semi-automated workflow for biodiversity data retrieval, cleaning, and quality control. Biodiversity Data Journal 2 https://doi.org/10.3897/BDJ.2.e4221
- Mayfield-Meyer T, Baskauf SJ, Endresen D, Bölling C, Wieczorek J, Pyle R, Buschbom J (2022) MaterialSample and its Properties. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.91407
- McDonald J, Wellington C, Coupland G, Pedersen D, Kitchen B, Bridgwood S, Hewitt M, Duggan R, Abdo D (2020) A united front against marine invaders: Developing a cost-effective marine biosecurity surveillance partnership between government and industry. Journal of Applied Ecology 57 (1): 77-84. https://doi.org/10.1111/1365-2664.13557
- Michener W (2006) Meta-information concepts for ecological data management. Ecological Informatics 1 (1): 3-7. https://doi.org/10.1016/j.ecoinf.2005.08.004
- Michener W, Jones M (2012) Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution 27 (2): 85-93. https://doi.org/10.1016/j.tree.2011.11.016
- Milián-García Y, Janke LA, Young R, Ambagala A, Hanner R (2021) Validation of an Effective Protocol for Culicoides Latreille (Diptera: Ceratopogonidae) Detection Using eDNA Metabarcoding. Insects 12 (5). https://doi.org/10.3390/insects12050401
- Milián-García Y, Young R, Madden M, Bullas-Appleton E, Hanner R (2021) Optimization and validation of a cost-effective protocol for biosurveillance of invasive alien species. Ecology and Evolution 11 (5): 1999-2014. https://doi.org/10.1002/ece3.7139
- Miller SE, Barrow LN, Ehlman SM, Goodheart JA, Greiman SE, Lutz HL, Misiewicz TM, Smith SM, Tan M, Thawley CJ, Cook JA, Light JE (2020) Building Natural History Collections for the Twenty-First Century and Beyond. BioScience 70 (8): 674-687. https://doi.org/10.1093/biosci/biaa069
- Morris P (2005) Relational Database Design And Implementation For Biodiversity Informatics. Retrieved from website https://doi.org/10.5281/ZENODO.59796
- M W, RJ S (Eds) (2017) The GEO Handbook on Biodiversity Observation Networks. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-27288-7
- Nicholson A, McIsaac D, MacDonald C, Gec P, Mason BE, Rein W, Wrobel J, Boer M, Milián-García Y, Hanner R (2020) An analysis of metadata reporting in freshwater environmental DNA research calls for the development of best practice guidelines. Environmental DNA 2 (3): 343-349. https://doi.org/10.1002/edn3.81
- Pawlowski J, Kelly-Quinn M, Altermatt F, Apothéloz-Perret-Gentil L, Beja P, Boggero A, Borja A, Bouchez A, Cordier T, Domaizon I, Feio MJ, Filipe AF, Fornaroli R, Graf W, Herder J, van der Hoorn B, Iwan Jones J, Sagova-Mareckova M, Moritz C, Barquín J, Piggott J, Pinna M, Rimet F, Rinkevich B, Sousa-Santos C, Specchia V, Trobajo R, Vasselon V, Vitecek S, Zimmerman J, Weigand A, Leese F, Kahlert M (2018) The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. Science of The Total Environment 637-638: 1295-1310. https://doi.org/10.1016/j.scitotenv.2018.05.002

- Penev L, Koureas D, Groom Q, Lanfear J, Agosti D, Casino A, Miller J, Arvanitidis C, Cochrane G, Hobern D, Banki O, Addink W, Kõljalg U, Copas K, Mergen P, Güntsch A, Benichou L, Benito Gonzalez Lopez J, Ruch P, Martin C, Barov B, Demirova I, Hristova K (2022) Biodiversity Community Integrated Knowledge Library (BiCIKL). Research Ideas and Outcomes 8 https://doi.org/10.3897/rio.8.e81136
- Philippi S, Köhler J (2006) Addressing the problems with life-science databases for traditional uses and systems biology. Nature Reviews Genetics 7 (6): 482-488. https://doi.org/10.1038/nrg1872
- Plitzner P, Müller A, Güntsch A, Berendsohn W, Kohlbecker A, Kilian N, Henning T, Stöver B (2017) The CDM Applied: Unit-Derivation, from Field Observations to DNA Sequences. Proceedings of TDWG 1 https://doi.org/10.3897/tdwgproceedings.1.20366
- Pop D (2011) Natural versus Surrogate Keys. Performance and Usability. Database Systems Journal 2 (2): 55-63. URL: http://www.dbjournal.ro/archive/4/6_Dragos_Pop.pdf
- Pornon A, Andalo C, Burrus M, Escaravage N (2017) DNA metabarcoding data unveils invisible pollination networks. Scientific Reports 7 (1). https://doi.org/10.1038/s41598-017-16785-5
- Powers S, Hampton S (2019) Open science, reproducibility, and transparency in ecology. Ecological Applications 29 (1). https://doi.org/10.1002/eap.1822
- Reichman OJ, Jones M, Schildhauer M (2011) Challenges and Opportunities of Open Data in Ecology. Science 331 (6018): 703-705. https://doi.org/10.1126/science.1197962
- Reyserhove L, Desmet P, Oldoni D, Adriaens T, Strubbe D, Davis AJS, Vanderhoeven S, Verloove F, Groom Q (2020) A checklist recipe: making species data open and FAIR. Database 2020 https://doi.org/10.1093/database/baaa084
- Ríos-Castro R, Romero A, Aranguren R, Pallavicini A, Banchi E, Novoa B, Figueras A (2021) High-Throughput Sequencing of Environmental DNA as a Tool for Monitoring Eukaryotic Communities and Potential Pathogens in a Coastal Upwelling Ecosystem. Frontiers in Veterinary Science 8 https://doi.org/10.3389/fvets.2021.765606
- Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, Otegui J, Russell L, Desmet P (2014) The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. PLoS ONE 9 (8). https://doi.org/10.1371/journal.pone.0102623
- Saarnak CL, Utzinger J, Kristensen T (2013) Collection, verification, sharing and dissemination of data: the CONTRAST experience. Acta Tropica 128 (2): 407-411. https://doi.org/10.1016/j.actatropica.2013.05.008
- Sandall E, Maureaud A, Guralnick R, McGeoch M, Sica Y, Rogan M, Booher D, Costello M, Edwards R, Franz N, Ingenloff K, Lucas M, Marsh C, McGowan J, Pinkert S, Ranipeta A, Uetz P, Wieczorek J, Jetz W (2022) Getting the GIST: Testing an integrative data structure for linking taxonomy, biodiversity and conservation. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.94209
- Schenekar T (2023) The current state of eDNA research in freshwater ecosystems: are we shifting from the developmental phase to standard application in biomonitoring? Hydrobiologia 850 (6): 1263-1282. https://doi.org/10.1007/s10750-022-04891-z
- Taberlet P, Bonin A, Zinger L, Coissac E (2018) Analysis of bulk samples. Oxford University Press URL: https://academic.oup.com/book/32663/chapter/270611440

- ten Hoopen P, Peat H, Ward P, Tarling G (2022) Polar biodiversity data: From a national marine platform to a global data portal. Patterns 3 (10). https://doi.org/10.1016/j.patter.2022.100566
- Thomer A, Cheng Y, Schneider J, Twidale M, Ludäscher B (2017) Logic-Based Schema Alignment for Natural History Museum Databases. KNOWLEDGE ORGANIZATION 44 (7): 545-558. https://doi.org/10.5771/0943-7444-2017-7-545
- Thompson C, Phelps K, Allard M, Cook J, Dunnum J, Ferguson A, Gelang M, Khan FAA, Paul D, Reeder D, Simmons N, Vanhove MM, Webala P, Weksler M, Kilpatrick CW (2021) Preserve a Voucher Specimen! The Critical Need for Integrating Natural History Collections in Infectious Disease Studies. mBio 12 (1). https://doi.org/10.1128/mBio.02698-20
- Traugott M, Kamenova S, Ruess L, Seeber J, Plantegenest M (2013) Empirically Characterising Trophic Networks. Advances in Ecological Research. 49. [ISBN 978-0-12-420002-9]. https://doi.org/10.1016/B978-0-12-420002-9.00003-2
- Travaini A, Bustamante J, Rodríguez A, Zapata S, Procopio D, Pedrana J, Martínez Peck R (2007) An integrated framework to map animal distributions in large and remote regions. Diversity and Distributions 13 (3): 289-298. https://doi.org/10.1111/j.1472-4642.2007.00338.x
- Vayssier-Taussat M, Albina E, Citti C, Cosson J, Jacques M, Lebrun M, Le Loir Y, Ogliastro M, Petit M, Roumagnac P, Candresse T (2014) Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. Frontiers in Cellular and Infection Microbiology 4 https://doi.org/10.3389/fcimb.2014.00029
- Vilhelm R (2006) What is the common problem that makes most biological databases hard to work with, if not useless to most biologists? Proceedings of the Gulf and Caribbean Fisheries Institute, 57. URL: http://hdl.handle.net/1834/29799
- Walls R, Deck J, Guralnick R, Baskauf S, Beaman R, Blum S, Bowers S, Buttigieg PL, Davies N, Endresen D, Gandolfo MA, Hanner R, Janning A, Krishtalka L, Matsunaga A, Midford P, Morrison N, Tuama É, Schildhauer M, Smith B, Stucky B, Thomer A, Wieczorek J, Whitacre J, Wooley J (2014) Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. PLoS ONE 9 (3). https://doi.org/10.1371/journal.pone.0089606
- Wells C, Paulay G, Nguyen B, Leray M (2022) DNA metabarcoding provides insights into the diverse diet of a dominant suspension feeder, the giant plumose anemone *Metridium farcimen*. Environmental DNA 4 (1): 147-156. https://doi.org/10.1002/edn3.225
- Wieczorek J, Bánki O, Blum S, Deck J, Döring M, Dröge G, Endresen D, Goldstein P, Leary P, Krishtalka L, Tuama ÉÓ, Robbins R, Robertson T, Yilmaz P (2014) Meeting Report: GBIF hackathon-workshop on Darwin Core and sample data (22-24 May 2013). Standards in Genomic Sciences 9 (3): 585-598. https://doi.org/10.4056/sigs.4898640
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, ’t Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K,

Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

- Young R, Abbott C, Therriault T, Adamowicz S (2017) Barcode-based species delimitation in the marine realm: a test using Hexanauplia (Multicrustacea: Thecostraca and Copepoda). Genome 60 (2): 169-182. https://doi.org/10.1139/gen-2015-0209
- Yu J, Young R, Deeth L, Hanner R (2020) Molecular Detection Mapping and Analysis Platform for R (MDMAPR) facilitating the standardization, analysis, visualization, and sharing of qPCR data and metadata. PeerJ 8 https://doi.org/10.7717/peerj.9974
- Zhang C((2007) Fundamentals of Environmental Sampling and Analysis. 1. Wiley https://doi.org/10.1002/0470120681
- Zizka VA, Leese F, Peinert B, Geiger M (2019) DNA metabarcoding from sample fixative as a quick and voucher-preserving biodiversity assessment method. Genome 62 (3): 122-136. https://doi.org/10.1139/gen-2018-0048

Supplementary material

Suppl. material 1: eDNA Laboratory Database Schema Outline doi

Authors: Alex Borisenko
Data type: Open Document Spreadsheet
Brief description: Outline of main tables used in the prototype eDNA Laboratory Database with a list and definitions of data fields.
Download file (28.47 kb)