Biodiversity assessment in Dutch freshwater and saltwater areas

A management and reporting system for reference DNA barcodes

Bastiaan Anker
Bachelor's student,
Bioinformatics
1095997

Educational institution
University of Applied Sciences Leiden
Zernikedreef 11,
2333 CK Leiden,
Netherlands

Institute
Naturalis Biodiversity Center
Darwinweg 2,
2333 CR Leiden
Netherlands

Educational supervisor
Koen Bossers, PhD
Lecturer-Researcher Bioinformatics

Supervisor
Rutger A. Vos, PhD
Researcher / Lecturer / Bioinformatician

Graduation period
27 November 2019
25 May 2021

Version 1.0

6 May 2021

Abstract

Freshwater and saltwater areas are regularly monitored for macrofauna diversity under the European Water (WFD) and Marine (MSFD) framework directives. The observed species composition serves as an indicator of the ecological quality of the water body. Traditionally, organisms are classified based on their morphological characteristics. However, the process of catching, sorting, and visually identifying species by experts is time-consuming and prone to error. Naturalis Biodiversity Center employs genetic techniques to identify species based on their DNA (a process known as “metabarcoding”). In this approach, accurate species-level assignments and, consequently, taxonomic coverage depend on well-curated DNA barcode reference databases. These databases provide reference sequences for species identification, but inadvertent errors in barcode generation have proven to impact their accuracy.

Internal databases at Naturalis are primarily fed by internal sequencing and are only on a project basis synchronized with international databases. This project presents a workflow to harvest data from public databases, check for reliability, and create the necessary data structures. A snapshot of this data, as it is available internally and in public databases (Barcode of Life Data Systems), provides an assessment of the total diversity of Dutch freshwater and saltwater areas.

A checklist from the Dutch Species Register provides a comprehensive overview of Dutch biodiversity. Their accepted species names and known synonyms, filtered to meet the binomial nomenclature rules, are the leading taxonomy for subsequent data retrieval processes. Aggregating all species’ genera, specimen data and sequence records are available in BOLD for 80% of taxa, covering 26% of the species from the checklist. Mollusca and Chordata (~66%) are the best-represented taxa, followed by Cnidaria and Platyhelminthes (~50%), and Annelida, Nematoda, and Arthropoda (20-35%). A total of 363 species' records can be used to complement the Naturalis database, which accounts for 46% of all Dutch annotated species retrieved from BOLD. About three-quarters of obtained public records are fully identified to the species rank, 91.4% are at least 500 bp in length, 96.4% contained less than 1% ambiguous base calls, 87% have a country annotation, 75% have latitude-longitude annotations, and 87% are assigned to a BIN.

Contents

Introduction

Biodiversity is the variation within and between all life forms on Earth or in a geographic area, including plants, animals, fungi, protists, archaea, and bacteria. The various populations of (two or more) organisms that live and interact with one another within a specified area are referred to as a community. These communities form an ecosystem along with the abiotic elements of their environment, such as landscape and climate. Ecosystem health is assessed in terms of organization (structure), vitality (function), and resilience[1]. High biodiversity aids and benefits this assessment, indicating that an ecosystem can sustain more different forms of life.

Home to a wide variety of aquatic life, two of the major types of ecosystems are Earth's freshwater and marine environments. Human well-being depends on many of the vital ecosystem services aquatic animals and plants and their ecological functions provide. Clean water, food, energy, jobs, atmospheric oxygen, buffers against new diseases, pests, predators, and protection against food shortages and climate change are essential aspects of our life[2]. However, the health of aquatic ecosystems is deteriorating due to a variety of human-induced stresses[3]. To monitor biodiversity and indicators of water quality, specific guidelines have been introduced for European aquatic ecosystems in the Water Framework Directive (WFD)[4] and Marine Strategy Framework Directive (MSFD)[5]. These directives describe which standards the quality of European waters should meet[6].

To assess ecological quality, identifying organisms at the family, genus, or species level is required. In traditional taxonomy, the classification of an organism is based on its morphological characteristics. However, the process for experts to catch, sort, and visually identify species is time-consuming and prone to error. Differences in experience, expertise, and taxon concepts between individual researchers can lead to documentation of different taxonomic groups from the same water body, resulting in discrepancies in ecological assessments[7].

Genetic markers

Serving as an alternative and less subjective approach to morphological identification, identifying species based on their DNA was first proposed by Hebert et al. in 2003[8]. The fungal community began using rDNA markers for similar purposes a decade earlier to identify unknown fungi[9]. Both groups considered DNA barcoding the only serious option for large-scale taxonomic identification, as did the bacterial community with rRNA markers decades prior[10].

Based on organisms holding shared core regions, DNA barcoding uses short sequences from specific, single-copy, genes to compare against a reference data set and identify an organism to its species. Ideally, only one gene sequence covers all taxonomic groups. However, as no such sequence exists, distinct regions each cover a subset of the organismal groups, as shown in Table 1[11]. Most commonly, nuclear internal transcribed spacer regions (ITS) are used in fungi, ribosomal RNAs (e.g., 12S, 16S, and 18S) in parts of prokaryotes and eukaryotes, chloroplast genome regions (rbcL, matK) in plants, and mitochondrial genome regions (COI) in animals.

Table 1: Overview of common DNA barcoding markers. The first column, 'organism group', represents the different organismal groups; the second column, 'marker gene/locus', the respective markers that have been used for DNA barcoding for each given group.

Organism group	Marker gene/locus
Plants	matK, rbcL, psbA-trnH, ITS
Animals	COI, Cytb, 12S, 16S
Fungi	ITS, TEF1α, RPB1, RPB2, 18S
Protists	ITS, COI, rbcL, 18S, 28S
Bacteria	COI, rpoB, 16S, cpn60, tuf, RIF, gnd

DNA metabarcoding

In contrast to DNA barcoding, which focuses on one specific organism, the technique of DNA metabarcoding determines the species composition in a sample[12]. The technique quantifies the molecular biodiversity of a sample using barcoding genes amplified and generated using high-throughput sequencing (HTS).

This process results in many different reads with partially overlapping regions. Clusters form for sets of reads that are more similar than a predefined threshold. Traditionally, a cluster serves as the operational taxonomic unit (OTU) of a species or genus. Alternatively, by using amplicon sequence variants (ASV), an OTU with a sequence similarity of 100 percent, sequence differences are resolved by even a single nucleotide change[13,14]. The selected sequence, or a representative from each cluster, is compared (using BLAST[15] searches) with the DNA profiles of species in (inter)national reference databases. A list of identifications derived from these comparisons indicates which species were found in the sample, as illustrated in Figure 1.

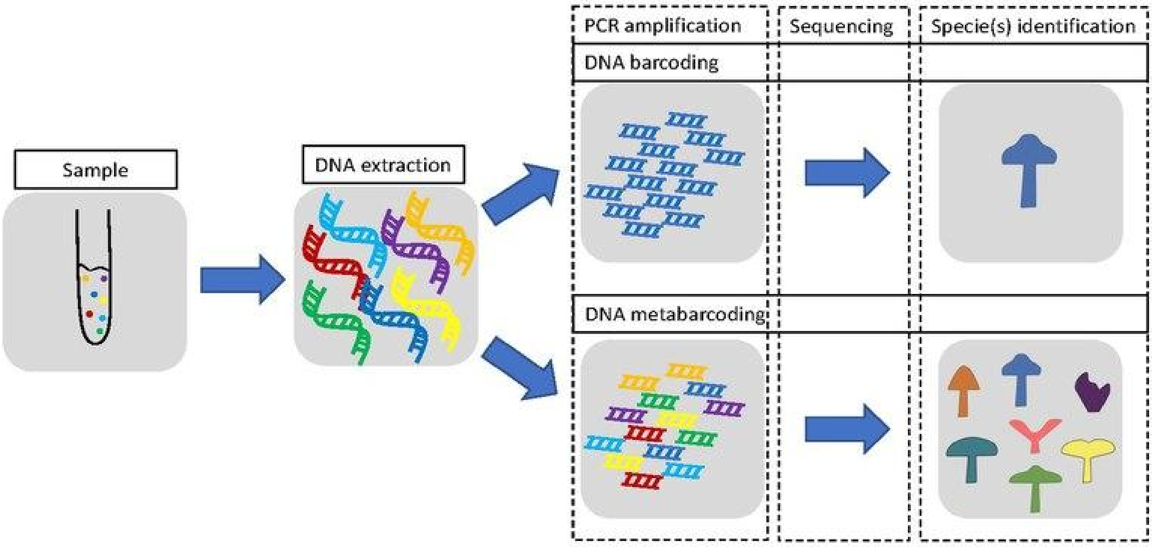


Figure 1: Illustration of the general workflow of DNA barcoding and metabarcoding and its differences in amplification/identification to find a specific species versus a whole community.

DNA barcode reference databases

Extensive DNA barcode reference databases are essential for monitoring biodiversity by comparing biological sequences. An unknown sequence can readily be searched against a public database to determine its closest species match. However, the taxonomic annotations of the records in these databases depend entirely on the data submitters. For GenBank[16], one of the most comprehensive public databases for nucleotide sequences, there is often not enough information to assess the accuracy of the annotation. Therefore, if not well-curated, reliance on this data poses the risk of misidentification - a known issue for GenBank[17-19].

The Barcode of Life Data System (BOLD)[20], a sequence database solely focused on DNA barcoding, addresses this issue by requiring more details before sequence data is approved. These requirements, often in the form of additional metadata and appropriately vouchered specimens, limit the number of species represented in BOLD. They compensate for this by periodically downloading a select and curated number of GenBank sequences. In return, BOLD allows for the automatic submission of data to GenBank. Transformed to meet the required format, GenBank assigns and returns an accession number for these records to ensure bidirectional linkage. The metadata of records in BOLD can still be refined after submission if changes occur due to newly obtained information. Changes to the BOLD records are then automatically submitted to GenBank. Due to this tight integration, according to a recent study, 11% of all COI barcode records on BOLD originate from GenBank, while 75% of the COI barcode records on GenBank derive from BOLD[21].

Many other specialized databases exist, each of which containing data for specific genetic markers (e.g., rRNA: SILVA[22]) or specific taxonomic groups (e.g., fungi: UNITE[23]). The UNITE database, centered on rDNA ITS based identification of Eukaryotes, primarily facilitates fungal species and reflects changes in their classification. They contribute to improving fungal annotation as a data and link provider for GenBank and the Global Biodiversity Information Facility (GBIF[24]).

SILVA provides a similar grade of quality for ribosomal RNA (rRNA) sequences, covering all domains of life (Archaea, Bacteria, and Eukarya). Periodical data sets contain curated and aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) rRNA sequences. The general overview of each database’s size for their respective domains is shown in Table 2.

Table 2: Overview of the mentioned public databases, from left to right: GenBank, BOLD, UNITE, and SILVA, with respectively their domains of life, number of species, and number of available sequence records.

	GenBank[16]	**BOLD**[25]	**UNITE**[26]	**SILVA**[27]
Domains of life	All	Eukarya (Animals, plants, fungi, protists)	Eukarya (Fungi)	All
# Species	450,000	320,000	102,000	-
# Sequence records	212×10^6	9.3×10^6	1.8×10^6	13×10^6 SSU, 1.9×10^6 LSU

Reliability of public reference data

The rise of DNA barcoding demands an ever-increasing need for high-quality reference data. While both GenBank and BOLD perform quality checks on submission and UNITE curation improves public records, the lack of information and the presence of inaccurate records remain a concern for identification[28-31]. According to one study from more than a decade ago, up to 20% of all fungal sequences were incorrectly identified to species level[32]. However, most studies into the accuracy of sequence data are focused on specific taxons[33-37].

Misidentification of the original material frequently results in the submission of incorrect sequences. Although often caused by synonymy- or morphological-based records, both operational and biological factors are causes of error and discrepancies[38,39]. Cross-contamination of samples, isolation techniques, the presence of endoparasites, and PCR-based errors are just a few of the contributing factors[29,40].

Nuclear encoded mitochondrial pseudogenes (numts[41]) are in a similar way of particular importance to DNA barcoding. These non-functional copies with high similarity to the mitogenome are often mistaken for genuine mitochondrial DNA[42-44]. A recent study among arthropods, one of the major phyla studied in taxonomy, found that DNA barcoding overestimates the number of species when numts are coamplified[45].

Obtaining public reference data

Sequence data of most databases is available for download through their websites or using command line applications, as it is unfeasible to fetch and identify millions of sequences one by one. Like SILVA, users can interact with the sequence data of UNITE through several data set releases, among which a general FASTA release for local BLAST searches. The BOLD ID Engine API[46] provides the ability to query the BOLD ID Engine via URL and is most often used to identify metazoan taxa sequences. On the other hand, BLAST+ searches against GenBank[47] are most often used to identify non-metazoan taxa sequences and validate the results of queries from BOLD.

Previous studies demonstrate that the usage of both GenBank and BOLD enables successful identification of metazoan taxa[48]. However, metabarcoding is known to generate data from all forms of life, not just metazoan taxa. Most non-metazoan taxa will be underrepresented if taxonomic assignments are made solely based on available BOLD records. Conversely, the use of GenBank would result in broader coverage of species, though at the risk of missing out on metazoan taxa and at the expense of lower accuracy.

Challenges in combining both data sets lay in the large amounts of data and the necessary reformatting to make them compatible. The nucleotide database of GenBank spans ~350 GB, from which a selection of specific markers (COI, 16S, 12S,18S, ITS, matK, rbcL, and trnL) are made to run blasting more efficiently. The other referenced databases available for download are BOLD (~4 GB, COI), UNITE (~1 GB, ITS), and SILVA (~6 GB, 16S/18S).

Naturalis and ARISE

The Naturalis collection, the Dutch national collection, ranks among the top five natural history collections in the world, with over 40 million objects. Its collection serves as the foundation for research in biodiversity and geology. The collection is divided into botany (5 million specimens), geology (3.2 million fossils, 0.8 million rocks and minerals), vertebrates (1.9 million objects), invertebrates (5.8 million specimens), entomology (18 million specimens), books, and type catalogs[49].

Internal databases to create reference data sets mainly consist of internal sequencing and partial synchronization with public databases. Compiling these data sets is done manually by linking specimen data and taxonomic information. The Netherlands Biodiversity Data service (NBA[50]) is the primary tool for accessing the Naturalis collection. Although its resources rely on the digitization of objects, over 8 million specimen and occurrence records are currently available. Ahead of the ARISE project[51], which aims to construct an infrastructure to identify and monitor every multicellular species in the Netherlands, it would be valuable to (periodically) complement the internal databases with public reference data.

Backbone taxonomy

A concern with combining molecular data sets from multiple sources is taxonomic incongruence. As each taxon (each species) is a hypothesis, defined in terms of its putative relatives' morphological and molecular relationships, perspective and available resources lead to differences in classification. A rudimentary selection on a singular backbone will lead to various specimens being at stake for misidentification. Additionally, each database uses its own identifier for sequence data. A (new) unique identifier is needed to build a shared BLAST database and link sequence information to the taxonomic data.

This project considers the Dutch Species Register (NSR[52]) as the leading taxonomy of species names, as it provides a current and comprehensive overview of Dutch biodiversity. The NSR is mainly coordinated and organized by Naturalis Biodiversity Center, in collaboration with the European Invertebrate Survey (EIS) - the Netherlands[53]. The available checklist contains all names of accepted species relevant to the Netherlands, including alternative names such as synonyms and expected species through recordings of their occurrence in nearby foreign areas. Collected data encompasses numerous expert-authored checklists, published papers, reports, and books, facilitating a digital entry to material that has already been reported elsewhere. An illustration of the geographic scope of species included in the NSR checklist is shown in Figure 2.

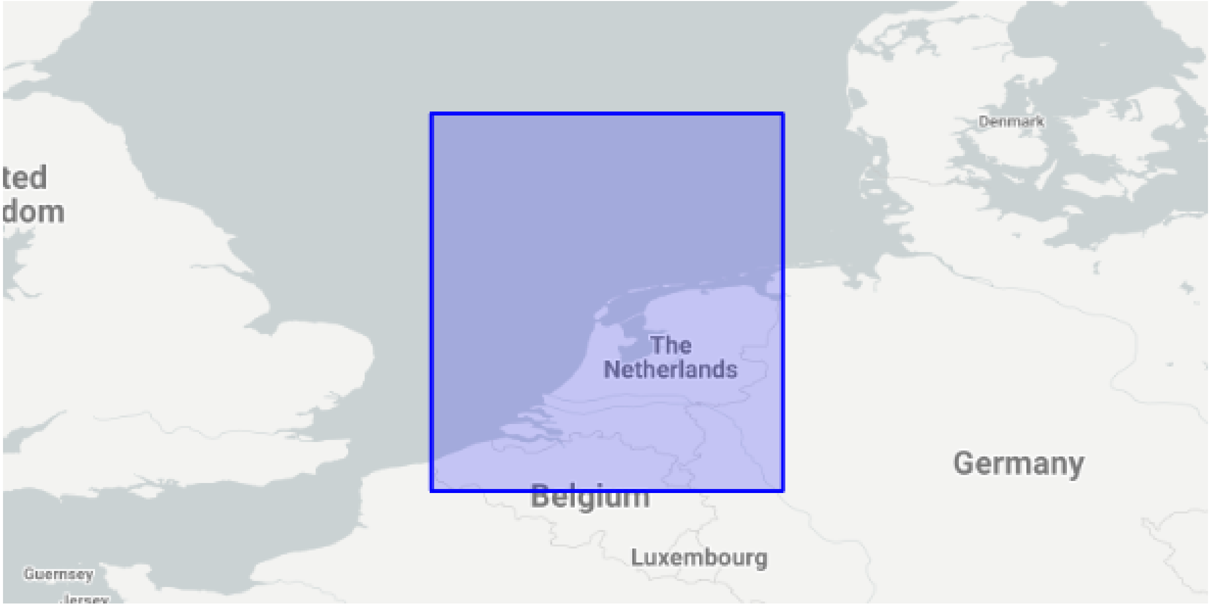


Figure 2: Geographic scope of the Dutch Species Register.

Aim of this study

Serving as a universal marker for animal species identification[54], Naturalis determines the genetic and taxonomic diversity for Dutch fauna through DNA barcoding of the COI marker of the mitochondrial genome. A selection for this project is made by exporting data from the BOLD database primarily covering metazoan taxa.

Reference databases for SILVA and UNITE are already available and easy to access. The addition of GenBank would be of little benefit due to a large number of duplicate records for COI data. The presence of pseudogenes further discourages its usage, as (opposed to BOLD) it processes nucleotide sequence submissions as small as 200 bp in length. For reference, the COI-5P marker spans 658 bp. However, GenBank could eventually be attached to this project's pipeline for data from non-COI records.

Utilizing BOLD's reference data, this study develops a system to (periodically) obtain the missing Naturalis reference records and perform an initial assessment of their quality and reliability. The accepted names and synonyms of the NSR link all sources of molecular data to create custom databases.

Research questions

Obtaining curated reference material from the Naturalis Biodiversity Center and a snapshot of public reference data, the following research questions are addressed:

1. Which data can Naturalis complement through BOLD?
2. What is the diversity of fauna in the Dutch freshwater and saltwater areas?
3. Does the accuracy and reliability of public reference data raise concerns for identification?

Material & Methods

General

The code for exporting public reference data is executed in Python 3.8[55] in a Windows environment using a command-line interface. Its ‘-h’ and ‘--help’ parameters provide a help message with examples of commands for running the script. Custom calculations in R[56] performed the steps to prepare, filter, merge and classify the species for further analysis. The output files, stored in their allocated output directory, are used to construct a custom database.

Figure 3 depicts the general steps performed in the Custom Databases pipeline. A custom export from the Dutch Species Register (NSR) contained the taxonomic classification of species of interest, including synonyms and expected species. A selection of curated taxa enabled the retrieval of BOLD specimen data and sequence records, Naturalis internal specimen records, names and phylogenetic lineages of the NCBI database, and NSR taxonomic classification. The resulting data sets provided a snapshot to assess the accuracy and reliability of BOLD’s reference data and to determine its overlap and discrepancies in comparison to Naturalis internal records.

The UTF-8 format standardized reading and writing to files. All required data, scripts, and resulting data sets are available in the Custom Databases DNA Sequences GitHub repository (https://github.com/naturalis/Custom-databases-DNA-sequences).

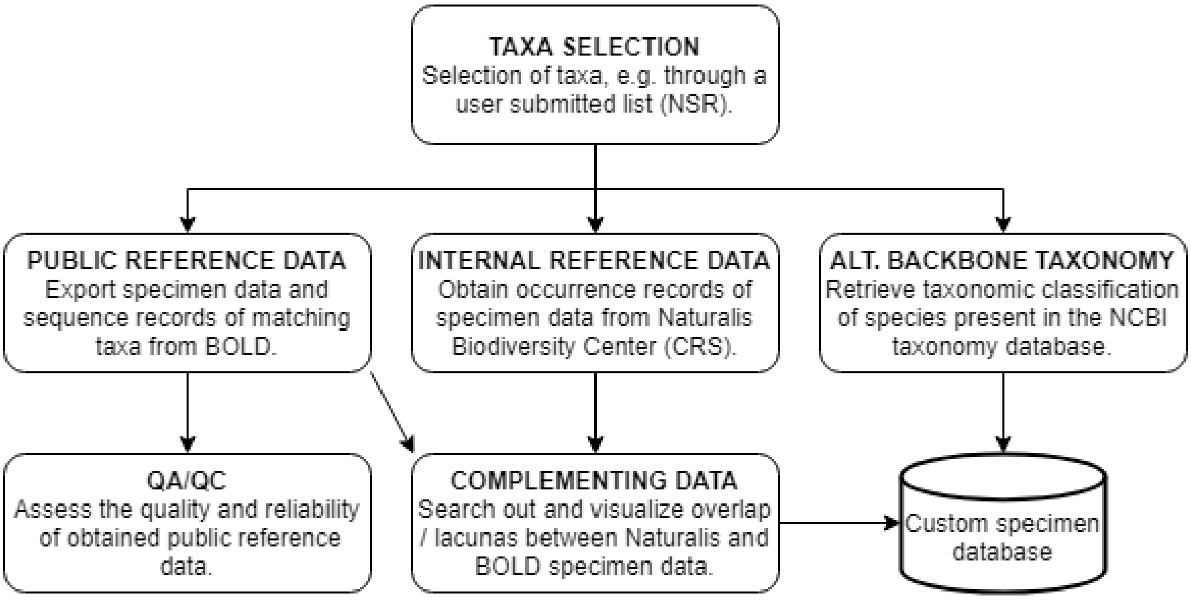

TAXA SELECTION
Selection of taxa, e.g. through a user submitted list (NSR).
PUBLIC REFERENCE DATA
Export specimen data and sequence records of matching taxa from BOLD.
INTERNAL REFERENCE DATA
Obtain occurrence records of specimen data from Naturalis Biodiversity Center (CRS).
ALT. BACKBONE TAXONOMY
Retrieve taxonomic classification of species present in the NCBI taxonomy database.
QA/QC
Assess the quality and reliability of obtained public reference data.
COMPLEMENTING DATA
Search out and visualize overlap / lacunas between Naturalis and BOLD specimen data.
Custom specimen database

Figure 3: The general workflow of the Custom Databases project, which used an export from the NSR to download corresponding reference data from BOLD and Naturalis Biodiversity Center, and taxonomic classification of species present in the NCBI Taxonomy database. The collected data was used to determine the overlap/discrepancies between data sets and assess the accuracy and reliability of public reference data. All of the molecular data were linked through NSR’s accepted species names and stored in a custom database.

Taxa selection (NSR)

A checklist from the Dutch Species Register (NSR) included all species of interest. The export of their scientific taxonomy, including expected species and all registered synonyms, generated two files: one for the current taxonomy and one for their synonyms.

Taxonomy export

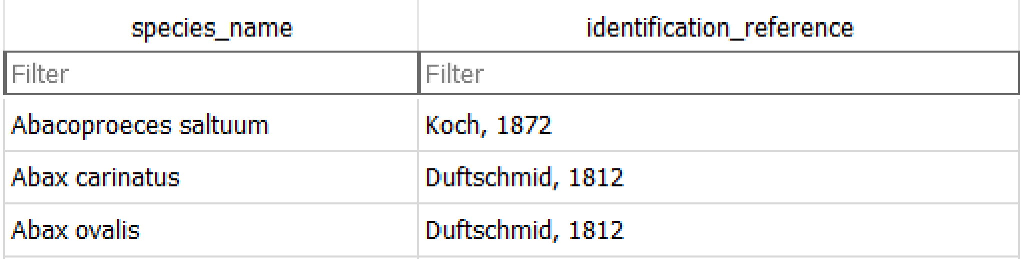
The taxonomy file included all accepted species names, specified by their scientific name and authority. Additional metadata on each record consisted of a common (Dutch) name, rank, NSR identification number, and presence status, as shown in Table 3.

Table 3: Overview of NSR taxonomy export data with, from left to right, their: scientific name, Dutch name, rank, NSR identification number, and presence status.

scientific name	common_name	rank	nsr_id	presence_status
Anseriphilus (Harrison, 1916)		genus	04B42A604D0B	3c
Abax parallelus (Duftschmid, 1812)	Echte breedborst	soort	0AHCYFBQVNNR	1a
Zapornia pusilla subsp. *intermedia* (Hermann, 1804)	Kleinst waterhoen	ondersoort	0MO551RKG4WX	1a
Alona guttata var. *tuberculata* Herrick, 1884	Gedeukte venalona	variëteit	0MJFR8X1VABW	1a
Anas platyrhynchos f. *domesticus*	Soepeend	vorm	04CA65BB4A66	2a

The selection of taxa followed the binomial nomenclature rules. Verification of at least one of the following criteria resulted in omission of the record: (a) records without a complete scientific name (including records identified solely on generic level); (b) records with ambiguous expressions present as species names (e.g., *subsp.*, *sp.*, *var.*, *f.*). As taxa are retrieved by primary taxonomic rank to species level, these criteria eliminate erroneous requests and preserve coverage rates.

By parsing each taxonomic name into its elementary components (genus, species, author, etc.), the Python Taxon parser library[57], an adaptation of the GBIF Java name-parser library[58], enabled the verification of criteria. Subsequently, the parser extracted the genus, specific epithet, and authority for all taxa adhering to the binomial nomenclature rules, as shown in Figure 4.



species_name	identification_reference
Filter	Filter
Abacoproeces saltuum	Koch, 1872
Abax carinatus	Duftschmid, 1812
Abax ovalis	Duftschmid, 1812

Figure 4: Parsed taxonomy output sample containing only binomial names and their authority.

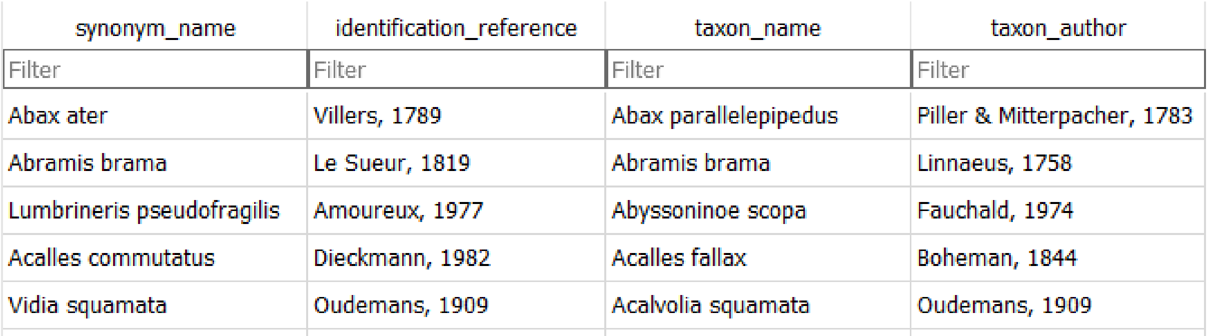
Synonyms export

The synonyms file included all known synonyms for names stored in the taxonomy file, referenced by the “taxon” column. Each known synonym contained metadata on the synonymous name, synonymy type, language, corresponding taxon, and NSR identification number, as shown in Table 4.

Table 4: Overview of NSR synonyms export data with, from left to right, their: synonym name, type of synonymy, classification of synonymy, corresponding scientific name, and NSR identification number.

synonym	type_synonym	language	taxon	taxon_nsr_id
Pituophis catenifer (Blainville, 1835)	isMisidentificationOf	Scientific	*Pituophis* sp.	029XFXRYS8SO
Rabdophaga dubia (Kieffer, 1891)	isHomonymOf	Scientific	*Rabdophaga dubiosa* Kieffer, 1913	0AHCYFBDQZRW
Pterocolus bisetatus Haller, 1881	isBasionymOf	Scientific	*Alloptes bisetatus* (Haller, 1881)	0AHCYSI10040
Onychiurus edinensis	isSynonymSLOf	Scientific	*Allonychiurus edinensis* (Bagnall, 1935)	0AHCYSI08063
Chlidonias hybridus (Pallas, 1811)	isInvalidNameOf	Scientific	*Chlidonias hybrida* (Pallas, 1811)	0AHCYFCORYXE
Tabakskleurig bloembokje	isAlternativeNameOf	Dutch	*Alosterna tabacicolor* (De Geer, 1775)	0AHCYFBRPAWZ
Anomalocera patersonii (Templeton, 1837)	isMisspelledNameOf	Scientific	*Anomalocera patersoni*	08IJVXDONB9L
Euchlora cuprea Hope, 1839	isSynonymOf	Scientific	*Anomala cuprea* (Hope, 1839)	0A1D34C040A1
Malariamug	isPreferredNameOf	Dutch	*Anopheles maculipennis*	0AHCYFBFQWGE

There were nine different synonym types, each classified as either “Dutch” or “Scientific” in language. Only scientific synonyms supplemented the species list. Dutch synonyms held a different common (vernacular) name and were not of use, thus discarded. Subsequently, all records were subject to the binomial standards outlined in the taxonomy file. The remaining synonyms paired with their respective taxa, as shown in Figure 5.



synonym_name	identification_reference	taxon_name	taxon_author
Filter	Filter	Filter	Filter
Abax ater	Villers, 1789	Abax parallelepipedus	Piller & Mitterpacher, 1783
Abramis brama	Le Sueur, 1819	Abramis brama	Linnaeus, 1758
Lumbrineris pseudofragilis	Amoureux, 1977	Abyssoninoe scopa	Fauchald, 1974
Acalles commutatus	Dieckmann, 1982	Acalles fallax	Boheman, 1844
Vidia squamata	Oudemans, 1909	Acalvolia squamata	Oudemans, 1909

Figure 5: Parsed synonym output sample containing the paired binomial names and their authority.

Public reference data (BOLD)

BOLD's Public Data Portal API[59] allows users to download specimen, sequence, and trace data using web service endpoints that follow a base URL and set of parameters. For this process five API Web Services are available, each of which tailored to retrieve the following information:

- Count data (Summary Stats Retrieval)
- Matching specimen data records (Specimen Data Retrieval)
- Matching sequences (Sequence Data Retrieval)
- Matching specimen data and sequence records (Full Data Retrieval)
- Matching trace files (Trace Data Retrieval)

Standard for URL query strings, an ampersand ('&') delimiter separates multiple input parameters, acting as a logical 'AND' clause. Correspondingly, a pipe ('|') delimited list allowed input parameters to accept multiple values, acting as a logical 'OR' clause. Table 5 lists all of the available parameters for each service.

Table 5: Overview of the type of parameters in the BOLD Public Data Portal API. The first column, 'parameter', represents the different types of parameters available to use; the second column, 'description', describes the use for each given parameter.

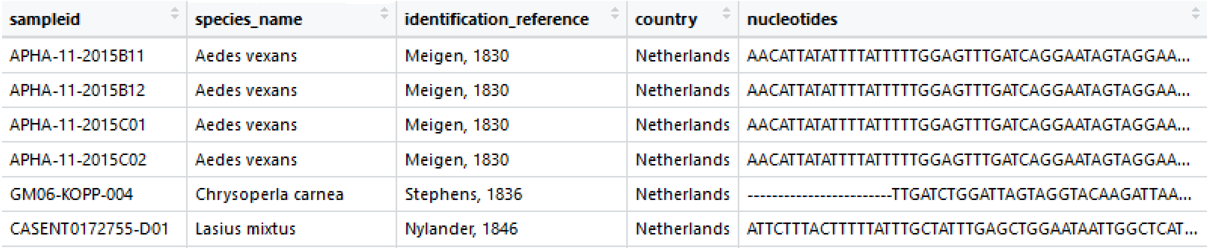
Parameter	Description
taxon	Returns all records containing matching taxa (includes scientific names at phylum, ..., genus, and species levels)
ids	Returns all records containing matching IDs (includes Sample IDs, Process IDs, Museum IDs and Field IDs)
bin	Returns all records contained in matching BINs (includes clusters defined by a Barcode Index Number URI)
container	Returns all records contained in matching projects or datasets (includes project codes and dataset codes)
institutions	Returns all records stored in matching institutions (Specimen Storing Site)
researchers	Returns all records containing matching researcher names (includes collectors and specimen identifiers)
geo	Returns all records collected in matching geographic sites (includes countries and province/states)
marker	Returns all specimen records containing matching marker codes
format	Returns the sequences in the specified formatted file (includes FASTA, TSV, XML, and JSON)

Using genera to export species with sequences present in BOLD

Aggregating all species' genera allowed for the retrieval of their subtended species present in BOLD, with their sequences. As specimen data was likely to be stored under different scientific names, the genus-level classification captured as many species as possible. Furthermore, the use of genera avoided virtually unidentified records (or those determined to a higher taxon).

Assessment of the reliability and quality of reference data required a combination of specimen data and sequence records. Records for each genus were exported one at a time using the Full Data Retrieval service. Aside from specifying a TSV format with the format parameter, the taxon parameter defined the taxa. The following request exported a Full Data Retrieval for records belonging to the taxon *Abax* in TSV format: *http://v4.boldsystems.org/index.php/API_Public/combined?taxon=Abax&format=tsv*.

The resulting formatted files contained data defined in BOLD's (original) data submission form, consisting of 80 fields. For each record, the vouchered specimen, taxonomic hierarchy, collection data, and sequence were included in the specimen data fields, as illustrated in Figure 6.



sampleid	species_name	identification_reference	country	nucleotides
APHA-11-2015B11	Aedes vexans	Meigen, 1830	Netherlands	AACATTATATTTTATTTTTGGAGTTTGATCAGGAATAGTAGGAA...
APHA-11-2015B12	Aedes vexans	Meigen, 1830	Netherlands	AACATTATATTTTATTTTTGGAGTTTGATCAGGAATAGTAGGAA...
APHA-11-2015C01	Aedes vexans	Meigen, 1830	Netherlands	AACATTATATTTTATTTTTGGAGTTTGATCAGGAATAGTAGGAA...
APHA-11-2015C02	Aedes vexans	Meigen, 1830	Netherlands	AACATTATATTTTATTTTTGGAGTTTGATCAGGAATAGTAGGAA...
GM06-KOPP-004	Chrysoperla carnea	Stephens, 1836	Netherlands	------------------------TTGATCTGGATTAGTAGGTACAAGATTAA...
CASENT0172755-D01	Lasius mixtus	Nylander, 1846	Netherlands	ATTCTTTACTTTTTATTTGCTATTTGAGCTGGAATAATTGGCTCAT...

Figure 6: BOLD formatted TSV file displaying several of the metadata fields, which, from left to right, include a sample ID, species name, identification reference, country, and sequence for each record.

Filter and compare the obtained specimen data against NSR taxonomy

The species name and identification reference of each record ensured a match with a subtended species from the original genus. As a result, specimen and sequence data were considered valid if they matched the binomial name and authority with the year of publication of an NSR taxon. In case of mismatches between species names and a synonym existed, the accepted name was adopted using the synonyms file, which covered misalignments in identification between the two databases. Specimens were considered from a Dutch freshwater or saltwater site if the "Country" field contained the location of "Netherlands".

Internal reference data (Naturalis)

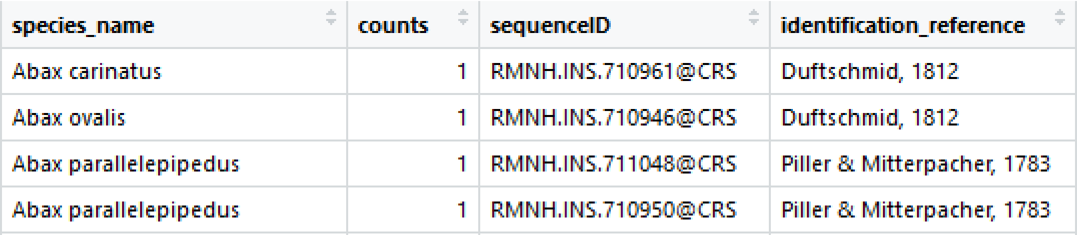
The Netherlands Biodiversity API (NBA)[60] facilitated access to the Naturalis zoological catalogs (CRS). Wrapper functions were available to query records of four different types:

- Geographic
- Multimedia
- Specimen
- Taxon

The ability to assess the overlap/discrepancies between collected public reference data and the records readily accessible at Naturalis Biodiversity Center was made possible by retrieving specimen type data. The nbaR[61] package, its specially developed R client, eased the extraction of data.

The “specimen_query” method allowed to query specific fields using a named list of query parameters. Its “sourceSystem.code” was set to match only CRS records, and a taxa parameter (“identifications.defaultClassification.genus”) contained the list of all species’ genera. The resulting data included various metadata fields, such as system data, taxonomic identification, collection data, number of specimens, and identifiers for each record.

Accessing the metadata fields enabled retrieving each specimen’s generic name (genus), specific name (species), identification reference, number of specimens, and unique CRS identifier, as shown in Figure 7. The accepted name of the NSR was adopted on any mismatches between species names, providing a synonym existed. Alternatively, if a specimen was not recognized, the record was omitted.



species_name	counts	sequenceID	identification_reference
Abax carinatus	1	RMNH.INS.710961@CRS	Duftschmid, 1812
Abax ovalis	1	RMNH.INS.710946@CRS	Duftschmid, 1812
Abax parallelepipedus	1	RMNH.INS.711048@CRS	Piller & Mitterpacher, 1783
Abax parallelepipedus	1	RMNH.INS.710950@CRS	Piller & Mitterpacher, 1783

Figure 7: Parsed NBA specimen record output sample with, from left to right, the scientific name, number of specimens, CRS identifier, and identification reference for each species.

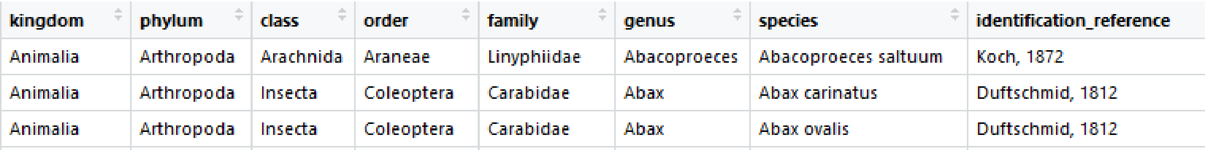
Backbone taxonomy

To facilitate taxonomic estimation by lowest common ancestor approach, the complete classification of species should ideally be present in a reference taxonomic data set. However, the NSR export only included the scientific names of the species of interest. The Netherlands Biodiversity API allows the acquisition of their higher taxa from the NSR. In addition, the retrieval of a second (external) classification allows for the ability to divert to another taxonomy.

Dutch Species Register (NSR)

The taxon services from the nbaR (specified with the “taxon_query” method) provide access to the classification of all NSR taxa. A taxon parameter (“acceptedName.scientificNameGroup”) contained the list of all relevant species, filtered (via the “sourceSystem.code”) to match only NSR records.

The resulting data included various metadata fields, such as system data, taxonomic classification, known synonyms, common names, and identifiers for each record. All main taxonomic ranks were extracted along with their unique NSR identifier, as shown in Figure 8.



kingdom	phylum	class	order	family	genus	species	identification_reference
Animalia	Arthropoda	Arachnida	Araneae	Linyphiidae	Abacoproeces	Abacoproeces saltuum	Koch, 1872
Animalia	Arthropoda	Insecta	Coleoptera	Carabidae	Abax	Abax carinatus	Duftschmid, 1812
Animalia	Arthropoda	Insecta	Coleoptera	Carabidae	Abax	Abax ovalis	Duftschmid, 1812

Figure 8: Parsed NSR taxonomic record output sample with all main taxonomic ranks and identification reference for each species.

NCBI Taxonomy

Additional taxonomic hierarchies can be retrieved using the taxizedb package[62] from various databases, including, but not limited to, NCBI, BOLD, GBIF, and ITIS. Because of their associated taxonomic identifier, NCBI served as the external reference and provided additional metadata for the NSR taxonomy.

The “name2taxid” function allowed the conversion of all species names to NCBI taxon IDs. Subsequently, the “classification” function retrieved the taxonomic hierarchies for each taxon ID and parsed out the required taxonomic levels/factors. A link between the NSR taxa and the corresponding NCBI taxon IDs maintained the referential integrity of species names.

Data structure

As each database used its unique identifier for specimen/sequence data, the NSR’s accepted names (and known synonyms) identified all molecular data and linked sequence information to its taxonomic data. The created data structure facilitated the creation of custom databases and served as the foundation for future analysis.

Unique identifiers for NSR species linked data across seven data sets, with the NSR species file as reference, as follows:

- Species ID -- Identifier corresponding to a unique NSR species name
- Species name -- Scientific name; identifying genus and species
- Reference -- Authority when first mentioned, and year of publication

The species identifiers, for example, facilitated access to corresponding records in the synonyms file, which contained all of the NSR species’ known synonyms, as follows:

- Synonym ID -- Identifier corresponding to a unique NSR synonym name
- Species ID -- Identifier corresponding to a matching NSR species name
- Synonym name -- Scientific name; identifying genus and species
- Reference -- Authority when first mentioned, and year of publication

Merging the public and internal specimen record files created a singular reference data set, with identifiers referring to each record’s details, as follows:

- Species-marker ID -- Unique identifier linking a species to their associated records
- Species ID -- Identifier corresponding to a unique NSR species name
- Database ID -- Identifier corresponding to a database storing the record
- Marker ID -- Identifier corresponding to the marker metadata of the record
- Sequence ID -- Unique identifier corresponding to the record metadata

Each marker and database ID referred to their respective data set with the various indexed names of markers/databases, as follows:

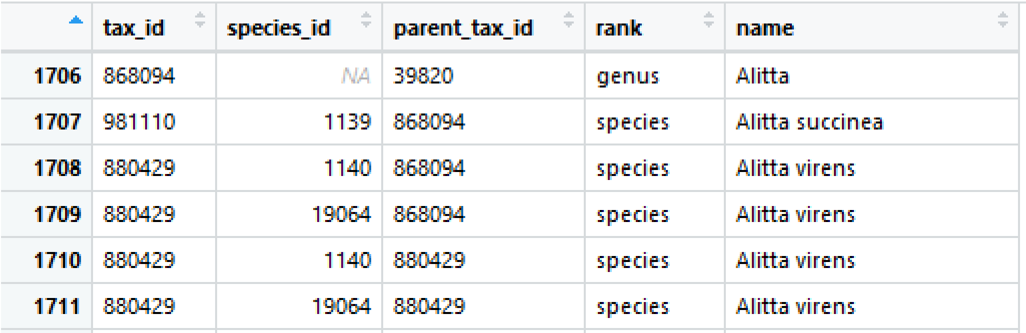
- Database ID -- Identifier corresponding to a unique database name
- Database name -- Name of the associated database

No sequence records (and as a result marker codes) were present in the CRS, as Naturalis stores sequence data in Geneious. Therefore, an *NA* description defined its marker name.

The NCBI Taxonomy database stores taxa with their name, rank, ID, and the corresponding ID of its parent taxa (in the case of species, its genus). Information extracted from the NCBI Taxonomy database (and for NSR's higher taxa) followed this layout. All species-rank taxa linked to the corresponding NSR species and accounted for the use of synonymous names within NCBI, as follows:

- Taxonomy ID -- Node id in the GenBank taxonomy database
- Species ID -- Identifier corresponding to a unique NSR species name
- Parent ID -- Parent node id in the GenBank taxonomy database
- Rank -- Rank of the associated node (kingdom, phylum, class, etc.)
- Name -- Name of the associated node

While species can occur multiple times within NCBI due to synonymy, they each hold a different parent taxon (parent_tax_id), as shown in Figure 9.



▲	tax_id	species_id	parent_tax_id	rank	name
1706	868094	NA	39820	genus	Alitta
1707	981110	1139	868094	species	Alitta succinea
1708	880429	1140	868094	species	Alitta virens
1709	880429	19064	868094	species	Alitta virens
1710	880429	1140	880429	species	Alitta virens
1711	880429	19064	880429	species	Alitta virens

Figure 9: Illustration of the NCBI taxonomy information, showing the presence of synonymy in the NCBI Taxonomy database when matched to the NSR species names. (NSR) Species IDs 1140 and 19064, while representing the same species in NCBI, each held two distinct parent taxonomic IDs.

Data visualization

Combining information from all the above data, the dplyr[63] package (used for data manipulation in R) created a taxonomic backbone and calculated the coverage of species obtained from Naturalis and BOLD.

A dynamic visualization of this complex data set, by presenting a collapsible Reingold-Tilford tree diagram[64], was eased by the D3.js[65] package. Through Shiny[66], the server observed the d3 collapsible tree library and its real-time layout. Data transferred back to Shiny was mapped to a series of logical expressions to create reactive filters. The filter rules were as follows:

- Clicking a node (taxa) interpreted it as being of interest, and a logical expression returned to it and its children.
- If specific siblings are clicked and opened, then the non-clicked siblings are not returned.

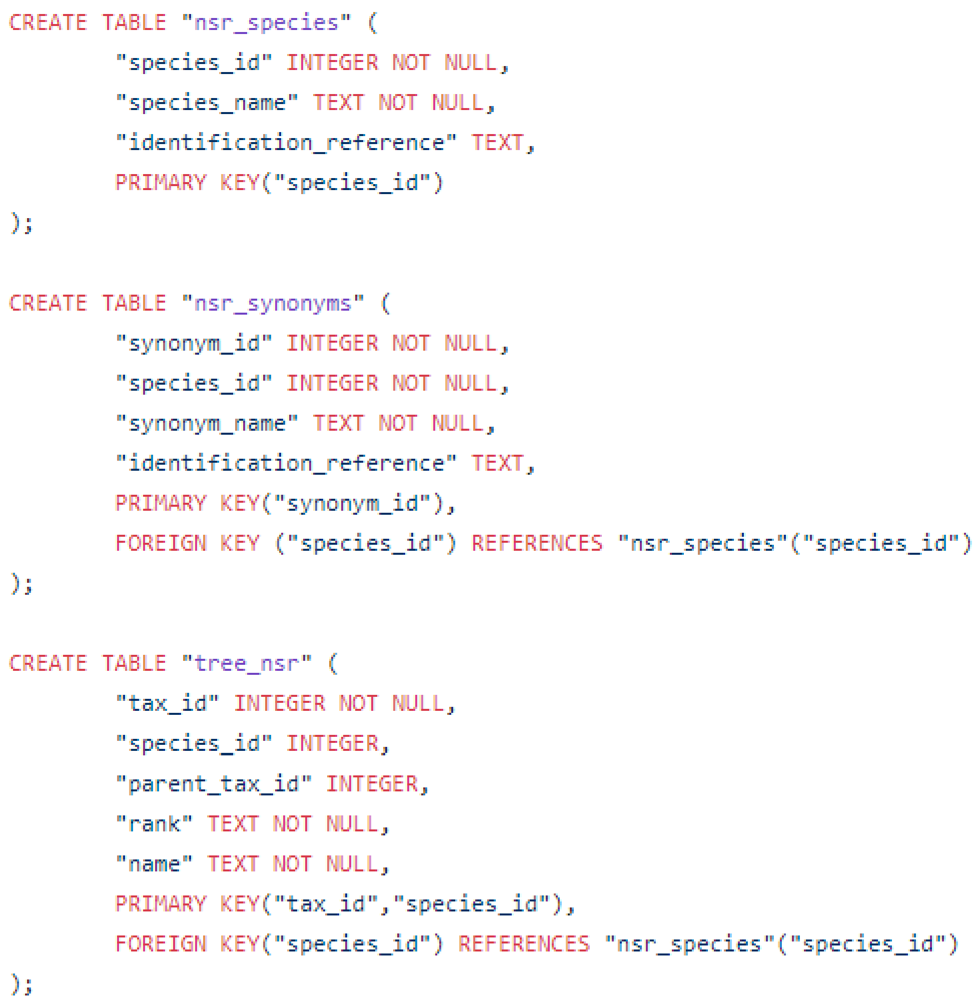
Various charts, enabled by the DataTables[67] and billboarder[68] packages, were linked to the reactive filters. By selecting nodes, the filters calculated, for example, the number of records in each taxonomic group and the presence of over-under representation for each database.

Creating custom databases

The relational database adhered to the same data structure that provided the functionality to compute and draw the hierarchical tree. SQL served as the programming language to define its database schema.

Database schema

The schema acted as a blueprint for the configuration of the database's components (e.g., tables, views, indexes, and triggers). Each data set functioned as an individual table (NSR taxonomy, NSR synonyms, NCBI taxonomy, etc.). Defined by a set of formulas/sentences, constraints imposed on the database ensured the compatibility of relationships between data sets. The SQL syntax and expressions used to create the NSR tables, and their attributes and constraints on each other are shown in Figure 10.



```
CREATE TABLE "nsr_species" (
        "species_id" INTEGER NOT NULL,
        "species_name" TEXT NOT NULL,
        "identification_reference" TEXT,
        PRIMARY KEY("species_id")
);

CREATE TABLE "nsr_synonyms" (
        "synonym_id" INTEGER NOT NULL,
        "species_id" INTEGER NOT NULL,
        "synonym_name" TEXT NOT NULL,
        "identification_reference" TEXT,
        PRIMARY KEY("synonym_id"),
        FOREIGN KEY ("species_id") REFERENCES "nsr_species"("species_id")
);

CREATE TABLE "tree_nsr" (
        "tax_id" INTEGER NOT NULL,
        "species_id" INTEGER,
        "parent_tax_id" INTEGER,
        "rank" TEXT NOT NULL,
        "name" TEXT NOT NULL,
        PRIMARY KEY("tax_id","species_id"),
        FOREIGN KEY("species_id") REFERENCES "nsr_species"("species_id")
);
```

Figure 10: The logical configuration for the nsr_species, nsr_synonyms, and tree_nsr table showcasing the relationship between the tables via the species_id attribute as a foreign key.

Populating databases

SQLite[69], a widely used database engine, served as a lightweight database management system (DBMS) to host the database. The schema and data sets respectively allowed for the creation and population of the database. While command-line options were available, the SQLite browser[70] - a high-quality and visual tool for creating, designing, and editing SQLite database files - allowed creating the database using a graphical user interface. Importing the schema allowed for the creation of the database file, as follows:

1. Start DB Browser for SQLite
2. Choose the *File -> Import -> Database from SQL file* menu option, which will open a dialog box. Use it to navigate to, select, and open the created schema.
3. Choose a filename (and location) to save the database under.

Similarly, importing the data sets populated the newly created database, as follows:

1. Choose the *File -> Import -> Table from CSV file* menu option, which will open a dialog box. Use it to navigate to, select, and open all of the project's results files.
2. Enable the checkbox next to the "Column names in first line" label, which will cause SQLite to use the names in the first line as the header.
3. If not by default chosen, select "comma" as Field separator, "double quotation mark" as Quote character, "UTF-8" as Encoding, and enable the checkboxes for the "Trim Fields" and "Separate tables" labels.
4. Click the OK button at the bottom of the tab to import each table.
5. Click on "Write Changes" (or "Ctrl+S") to commit the changes to the database file.

Quality assurance and quality control (QA/QC)

The proposed quality system aimed to collect reliable public reference data. The process of exporting specimen data and sequence records was similar to that of retrieving BOLD records, namely: (1) data selection through provided taxa (e.g., through a user-submitted list: NSR export); (2) verification and filtering of taxa through matching checklists. Segment quality criteria determined the (3) completeness of specimen data, (4) quality of sequence records, and (5) reliability of species identification.

Data selection and verification

There were two options for exporting selected taxa: downloading all available records or downloading only species records that appeared in the checklist(s). These checklists enabled the selection or exclusion of names on specific sampling areas and synonym coverage. The frequency of a binomial scientific name and its known synonyms in available records represented the number of specimen records.

Completeness of specimen data

Defining a record as complete included at least: (1) an identification to species rank; (2) a valid sampling location; (3) a complete date of occurrence; (4) a given basis of record; (5) assignment to a BIN. Along with these criteria, an initial division isolated GenBank-mined records and those without reference identification. Labeled as “B-type” records, the user could choose to include or exclude these records. The following definitions describe the use/description of the criteria:

> (1) - Taxa that were “fully identified” were those confirmed on a taxonomic identification to species rank, according to Nilsson et al. (2005)[71]. This criterion ruled out taxa identified solely at a generic level, as well as on any higher taxa. Furthermore, ambiguous expressions in the species name (e.g. *subsp.*, *sp.*, *var.*, *f.*) or sole use of secondary taxonomic ranks resulted in removing said terms.
>
> (2) - Although users can choose to include or exclude records without information on sampling location, a valid set of coordinates ensured species selection in sampling areas. The origin of a specimen must contain either country or latitude-longitude annotations.
>
> (3) - Complete dates of occurrences provide necessary information about a specimen’s observation to show the distribution of records over time. This distribution denotes the possibility of records’ bias towards specific periods or seasons.
>
> (4) - The basis of a record distinguished human observations from preserved, fossilized, or living specimens. A required assigned basis, even if *NA*, allowed users to remove records that they might not want to include.
>
> (5) - BIN (Barcode Index Number) assignments act as clusters of similar barcode references in high concordance with a species. The composition of (genetically identical) taxa for a BIN, and its relationship to a species, helped indicate the reliability of species identification.

Quality of sequence records

Sequence annotation determined the quality of the respective sequence records. However, different markers identified COI reference records, such as the COI-3P fragment (836 bp) and the COI-5P fragment (658 bp). Therefore, nucleotides in sequence records followed a set of general criteria, with the omission of records if (a) sequences were shorter than a minimum chosen user size (500 bp) or (b) sequences contained more than a specified percentage (1%) of ambiguous base calls (Ns).

The total bp length excludes gaps (-) in sequence data, as sequences in BOLD are generally internally aligned with a reference genome.

Reliability of species identification

Filtered specimen data and sequence records followed a grading system for each species, based on a ranking system[72] and its generalized implementation[73]. The level of congruence between morphospecies and COI barcode clusters (BINs) assigned records one of five grades (A to E).

In contrast to individual analyses of each species' data, a recently published modification[74] to the ranking system allowed for automated large-scale auditing and annotating custom DNA barcode reference data. The annotation grades criteria introduced in the BAGs pipeline were defined as follows (see Figure 11).

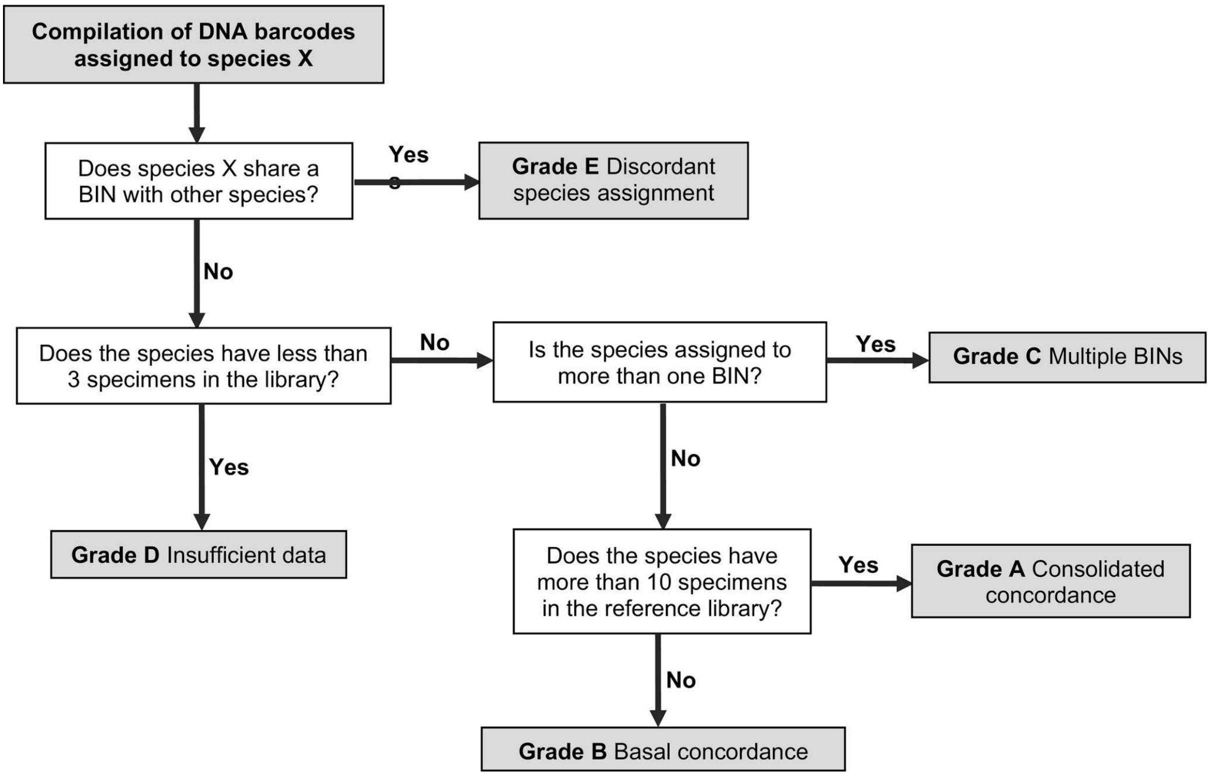

Compilation of DNA barcodes assigned to species X
Does species X share a BIN with other species?
Yes
s
Grade E Discordant species assignment
No
Does the species have less than 3 specimens in the library?
No
Is the species assigned to more than one BIN?
Yes
Grade C Multiple BINs
Yes
Grade D Insufficient data
No
Does the species have more than 10 specimens in the reference library?
Yes
Grade A Consolidated concordance
No
Grade B Basal concordance

Figure 11: Workflow for automated auditing and annotating qualitative grades to each species in a BAGS-compiled reference library.

Results

The NSR export files contained all accepted species names and known synonyms. The extraction of their binomial nomenclature, and respective authority with the year of publication, was the input for all acquired data sets and subsequent analysis.

Taxa selection (NSR)

Overview of export data

The NSR taxonomy export included 30,694 taxonomic classifications, with 30,359 identified to species level, as shown in Table 6. After applying a set of criteria to the taxonomic names, 30,322 distinct species remained. Verification of the omitted taxa revealed that 25% were subspecies with no specific (species) name, and a total of 87.5% did not include an authority reference and/or year of publication.

Table 6: Overview of the type of species in the NSR taxonomy export. The first column, 'rank', represents the different types of taxa stored; the second column, 'number of taxa', the number of taxa for each given rank.

Rank	Number of taxa
Genus	4
Species	30,359
Subspecies (subsp.)	306
Variety (var.)	19
Form (f.)	6

The species' synonyms existed in many different forms. There were 15,446 synonym names, of which 4,799 scientific as shown in Table 7, with 4,498 taxa adhering to the criteria.

Table 7: Overview of the type of synonyms in the NSR synonyms export. The first column, 'synonym type', represents the different types of synonyms stored; the second and third columns respectively portray the number of records for 'Dutch' and 'Scientific' notations for each given synonym type.

Synonym type	Dutch	Scientific
isAlternativeNameOf	1,711	
isBasionymOf		123
isHomonymOf		11
isInvalidNameOf		24
isMisidentificationOf		1
isMisspelledNameOf		227
isPreferredNameOf	8,936	
isSynonymOf		3,992
isSynonymSLOf		421

Retrieving classification (NSR)

The retrieval of higher taxonomic classification resulted in the main taxonomic ranks for 30,269 species. During the taxon record retrieval, 57 species failed to match, and three duplicates existed. Differences in species classification were either a variation in family name or a different identification reference.

The species' biological classification spanned a wide range of taxonomic ranks, all of which contributed to the Animal kingdom, as shown by the unique number of taxa in Table 8. The majority of species classified as belonging to the Arthropoda phylum (85%), as illustrated by the number of species by higher taxon in Figure 12, and then either class Insecta (72%) or Arachnida (8%).

Table 8: Overview of the classification of NSR taxa. The first column, 'taxonomic rank', represents the taxon's level in the Linnaean hierarchy; the second and third columns respectively portray the number of unique taxa for each given rank along with their most significant contributors.

Taxonomic rank	Number of taxa	Biggest contributors
kingdom	1	Animalia (100%)
phylum	25	Arthropoda (85%)
class	72	Insecta (72%), Arachnida (8%)
order	317	Diptera (19%), Hymenoptera (19%), Coleoptera (15%)
family	1,990	Ichneumonidae (5%), Staphylinidae (4%), Braconidae (3%)
genus	10,241	*Megaselia* (0.6%), *Coleophora* (0.4%)
species	30,263	

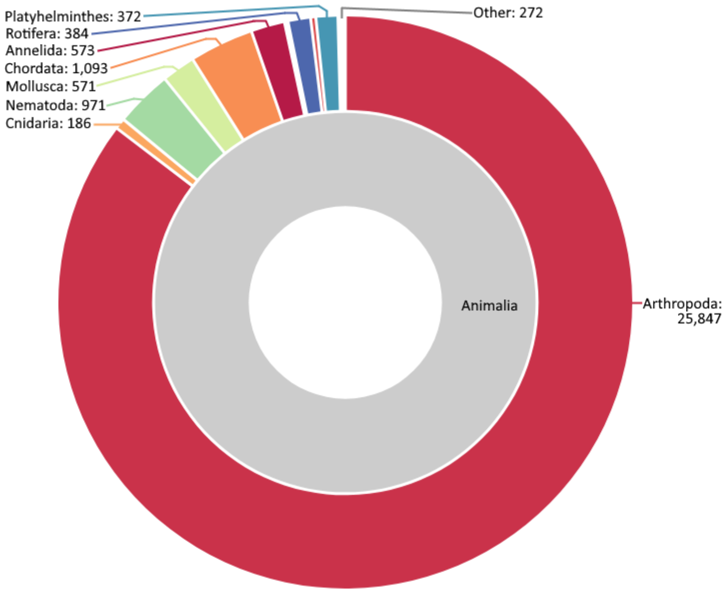
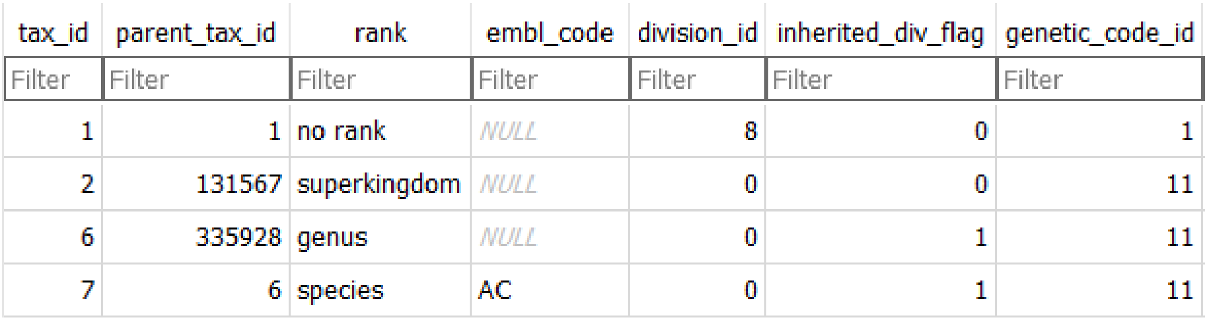


Figure 12: The number of species in the NSR export data categorized by higher taxon (phylum).

Diverting to secondary classification (NCBI Taxonomy)

GenBank’s full taxonomy, as used for annotating sequences, was available in the NCBI Taxonomy database. The taxizedb package facilitated access to the NCBI taxonomy and is alternatively available through the NCBI Taxonomy FTP[75] (File Transfer Protocol). The taxonomy database release consisted of three tables: hierarchy, names, and nodes. Each table was based on a node’s taxonomy ID and captured either the node’s ancestral relationships, taxonomic name, classification, or information about its parent. The nodes table linked the taxonomic ID to its parent ID and displayed the node’s taxonomic rank, as shown in Figure 13.



tax_id	parent_tax_id	rank	embl_code	division_id	inherited_div_flag	genetic_code_id
Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	1	no rank	NULL	8	0	1
2	131567	superkingdom	NULL	0	0	11
6	335928	genus	NULL	0	1	11
7	6	species	AC	0	1	11

Figure 13: Illustration of the nodes table, representing from left to right the node id in GenBank taxonomy database, the parent node id in GenBank taxonomy database, the rank of the node itself, the locus-name prefix, and several flags/IDs relating to inherited data from the parent.

Matching NSR species names to corresponding NCBI taxonomic IDs enabled the retrieval of all its parent taxa. The extraction of all species’ taxonomy levels yielded 18,822 taxonomic hierarchies, with 62% of the species recognized in the NCBI taxonomy. Differences in classification were observed between the two backbones (NSR vs NCBI), as shown by the unique number of taxa per taxonomic rank in Table 9.

Table 9: Overview of the classification of NCBI taxa. The first column, ‘taxonomic rank’, represents the taxon’s level in the Linnaean hierarchy; the second and third columns respectively portray the number of unique taxa for each given rank along with their most significant contributors. The second column additionally presents the respective number of missing/unique taxa when compared to the NSR.

Taxonomic rank	**Number of taxa** (Missing/Unique NCBI)	**Biggest contributors**
kingdom	1	Animalia (100%)
phylum	28 (3/6)	Arthropoda (83%)
class	75 (25/28)	Insecta (72%), Arachnida (6%)
order	302 (84/69)	Coleoptera (20%), Diptera (15%), Hymenoptera (13%)
family	1,821 (354/185)	Staphylinidae (4%), Ichneumonidae (2%)
genus	7,970 (2,450/177)	*Coleophora* (0.5%), *Megaselia* (0.5%)
species	18,822 (11,477/31)	

Public reference data (BOLD)

Following the public data retrieval pipeline, aggregation of genera for NSR's 30,322 species (incl. known synonyms) yielded a total of 2,580,759 records from the BOLD database. However, 1,996 genera (nearly one-fifth of all genera) did not return any records.

Records for 7,906 distinct species matched against the NSR checklist, resulting in a 26% coverage rate. The country field 'Netherlands' yielded 2,537 records, representing 783 species (2.6%), as shown in Table 10.

Table 10: Overview of the type of records retrieved from BOLD. The first column, 'type', represents the different categories of specimen records; the second and third columns respectively portray the number of unique taxa and number of records for each given category.

Type	Number of taxa	Number of records
Total	111,242	2,580,759
Matching Species	7,906	457,356
Dutch Species	783	2,537

Distribution of taxa

The distribution of the Dutch specimens revealed that each species had a decent number of sequence records. Classification of specimen data held the same key contributors as the NSR, as shown in Table 11. Most species belonged to the Arthropoda phylum and class Insecta.

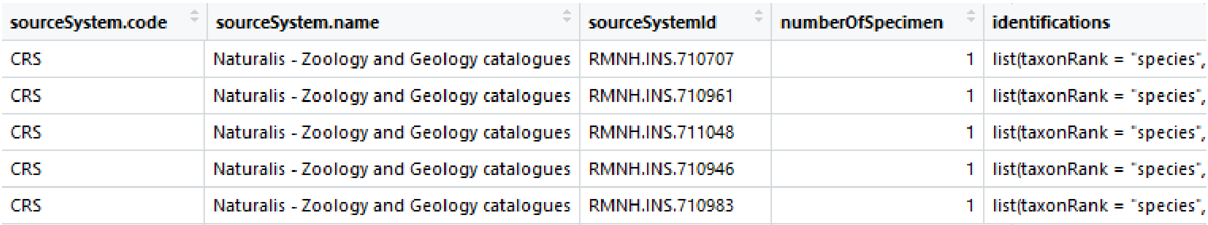
Table 11: Overview of BOLD's Dutch annotated specimen records. The first column, 'taxonomic rank', represents the taxon's level in the Linnaean hierarchy; the second, third and fourth columns respectively portray the number of unique taxa for each given rank along with their coverage percentage against the NSR and most significant contributors.

Taxonomic Rank	Number of taxa	Coverage (%)	Biggest contributors
kingdom	1	100%	Animalia (100%)
phylum	8	32%	Arthropoda (75%)
class	15	21%	Insecta (63%), Aves (18%), Arachnida (12%)
order	53	17%	Lepidoptera (52%), Araneae (10%), Passeriformes (9%)
family	184	9%	Noctuidae (15%)
genus	542	5%	*Autographa* (10%)
species	783	2.6%	

Internal reference data (Naturalis)

Naturalis’ specimen records were provided by the CRS, as accessed through the Netherlands Biodiversity API. These records acted as a reference to determine which obtained public reference data could complement the Naturalis database.

Naturalis’ digitized collection yielded 45,187 matching specimen records (illustrated in Figure 14) after aggregating and querying each species’ genera.



sourceSystem.code	sourceSystem.name	sourceSystemId	numberOfSpecimen	identifications
CRS	Naturalis - Zoology and Geology catalogues	RMNH.INS.710707	1	list(taxonRank = "species",
CRS	Naturalis - Zoology and Geology catalogues	RMNH.INS.710961	1	list(taxonRank = "species",
CRS	Naturalis - Zoology and Geology catalogues	RMNH.INS.711048	1	list(taxonRank = "species",
CRS	Naturalis - Zoology and Geology catalogues	RMNH.INS.710946	1	list(taxonRank = "species",
CRS	Naturalis - Zoology and Geology catalogues	RMNH.INS.710983	1	list(taxonRank = "species",

Figure 14: Illustration of the CRS specimen records, with from left to right the data source code, the data source name, the data source identifier of a specimen record, the number of specimens of the specified record, and a list of taxonomic identifications of the specimen.

There were 28,752 matching species with 7,003 distinct species names. The distribution of these records revealed that each species had a large number of specimens. Classification of specimen data held the same key contributors as the NSR, as shown in Table 12.

Table 12: Overview of the CRS specimen records. The first column, ‘taxonomic rank’, represents the taxon’s level in the Linnaean hierarchy; the second, third and fourth columns respectively portray the number of unique taxa for each given rank along with their coverage percentage against the NSR and most significant contributors.

Taxonomic Rank	Number of taxa	Coverage (%)	Biggest contributors
kingdom	1	100%	Animalia (100%)
phylum	17	68%	Arthropoda (69%)
class	62	86%	Insecta (49%), Arachnida (9%)
order	276	87%	Coleoptera (20%), Lepidoptera (15%)
family	1,469	74%	Noctuidae (4.2%), Aphididae (3.2%)
genus	4,714	46%	
species	7,003	23%	

Complementing data

A comparison between the (28,752) Naturalis specimen records and the (2,537) obtained Dutch records from BOLD showed an equal rate of overlap and discrepancies. Operations performed on the data sets included set intersections (AND), set unions (OR), and set differences, as shown in Table 13.

Table 13: Overview of the statistical analysis of Naturalis and BOLD specimen data. The first column, 'data set', represents the different types of analysis (A only, Intersection, B only, Union); the second column, 'number of species', the number of species for each given data set.

Data set	Number of species
CRS only	6,637
CRS ∩ BOLD (AND)	369
BOLD only	363
CRS ∪ BOLD (OR)	7,369

Analysis of specimen differences

In total, records for 363 species complemented the Naturalis data set. The majority of these records belonged to an insect order, primarily butterflies and moths, as shown by the unique number of taxa by taxonomic rank in Table 14.

Table 14: Overview of specimen records unique to BOLD. The first column, 'taxonomic rank', represents the taxon's level in the Linnaean hierarchy; the second and third columns respectively portray the number of unique taxa for each given rank along with their most significant contributors.

Taxonomic Rank	Number of taxa	Biggest contributors
kingdom	1	Animalia (100%)
phylum	6	Arthropoda (89%), Chordata (10%)
class	10	Insecta (75%), Arachnida (13%), Aves (9%)
order	28	Lepidoptera (56%), Araneae (12%), Diptera (10%)
family	98	Geometridae (13%), Gracillariidae (10%), Noctuidae (7%)
genus	239	*Phyllonorycter* (7%), *Coleophora* (4%)
species	363	

Aside from arthropods, only a few dozen species represented the remaining taxa. Most of these species had no direct common ancestor or lineal descendant. Figure 15 depicts the taxonomic hierarchy of these species (excl. arthropods).

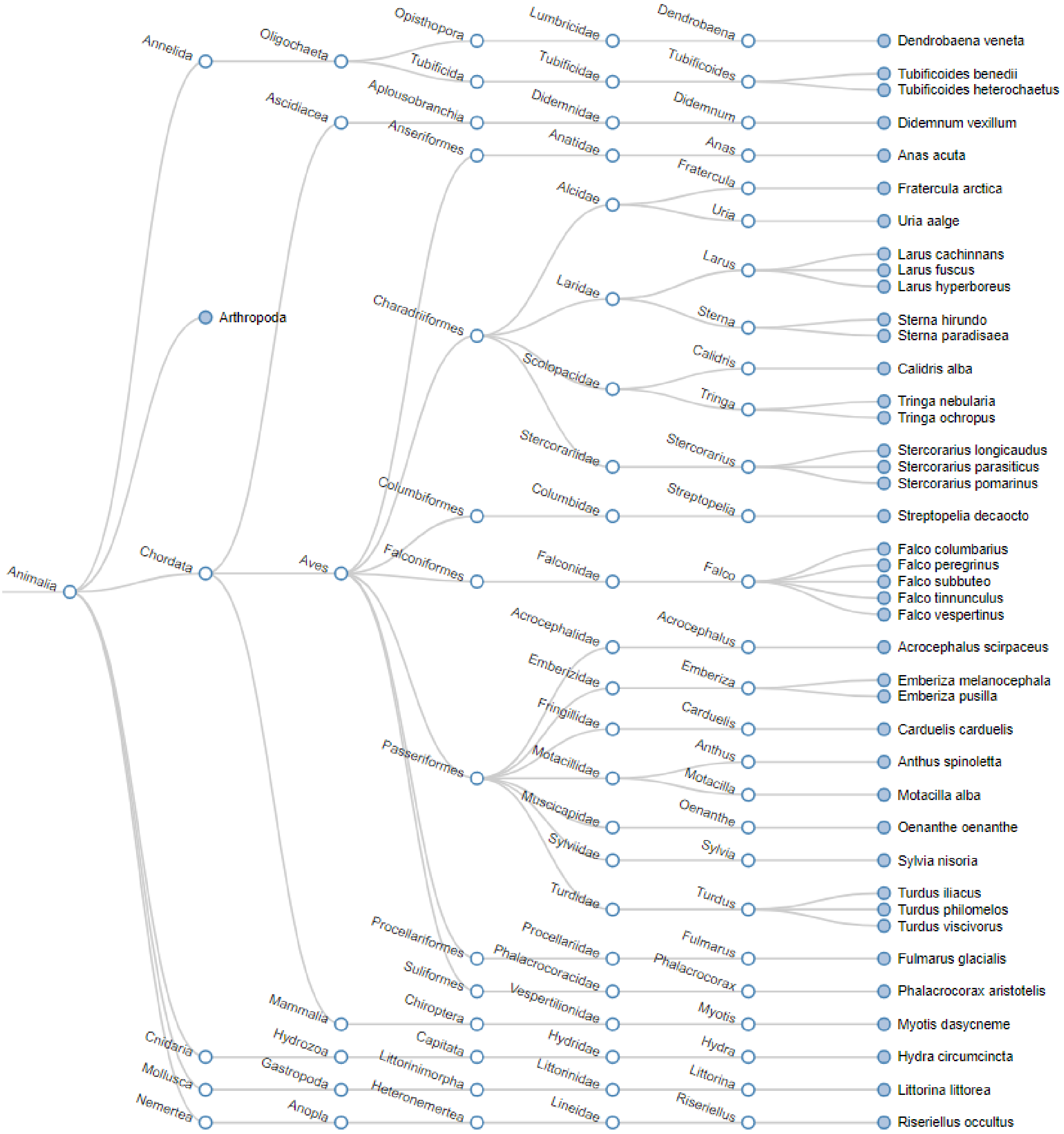


Figure 15: Tree diagram showing the taxonomic hierarchy of all complementary BOLD specimen data, excluding the phylum Arthropoda. Starting on the left, the animal kingdom is the root of all species. Branching out towards the right are all taxon's levels in the Linnaean hierarchy, ultimately branching out to a particular species name.

The majority of all records, aside from those labeled as being mined from GenBank, originated from 17 different institutions, including, but not limited to, the Canadian National Collection of Insects, Arachnids and Nematodes (35), University of Oulu (12), and the Animal and Plant Health Agency of the United Kingdom (10).

Assessment of the diversity in Dutch fauna

The union of obtained Naturalis and BOLD specimen records served as a snapshot of the overall diversity of Dutch freshwater and saltwater areas. This collection of specimen data included records for 7,369 distinct species. The coverage of this data set against the NSR checklist was 24%, as shown in Table 15 by taxonomic rank and in Figure 16 for the number of species by higher taxon.

Table 15: Overview of unioned Naturalis and BOLD specimen records. The first column, 'taxonomic rank', represents the taxon's level in the Linnaean hierarchy; the second, third and fourth columns respectively portray the number of unique taxa for each given rank along with their coverage percentage against the NSR and most significant contributors.

Taxonomic Rank	Number of taxa	Coverage (%)	Biggest contributors
kingdom	1	100%	Animalia (100%)
phylum	17	68%	Arthropoda (73%), Chordata (11%)
class	63	88%	Insecta (51%), Arachnida (13%)
order	277	87%	Lepidoptera (18%), Coleoptera (17%)
family	1,483	75%	Aphididae (3.2%), Staphylinidae (3.0%), Noctuidae (2.7%)
genus	4,847	47%	*Phyllonorycter* (0.4%), *Coleophora* (0.3%)
species	7,369	24%	

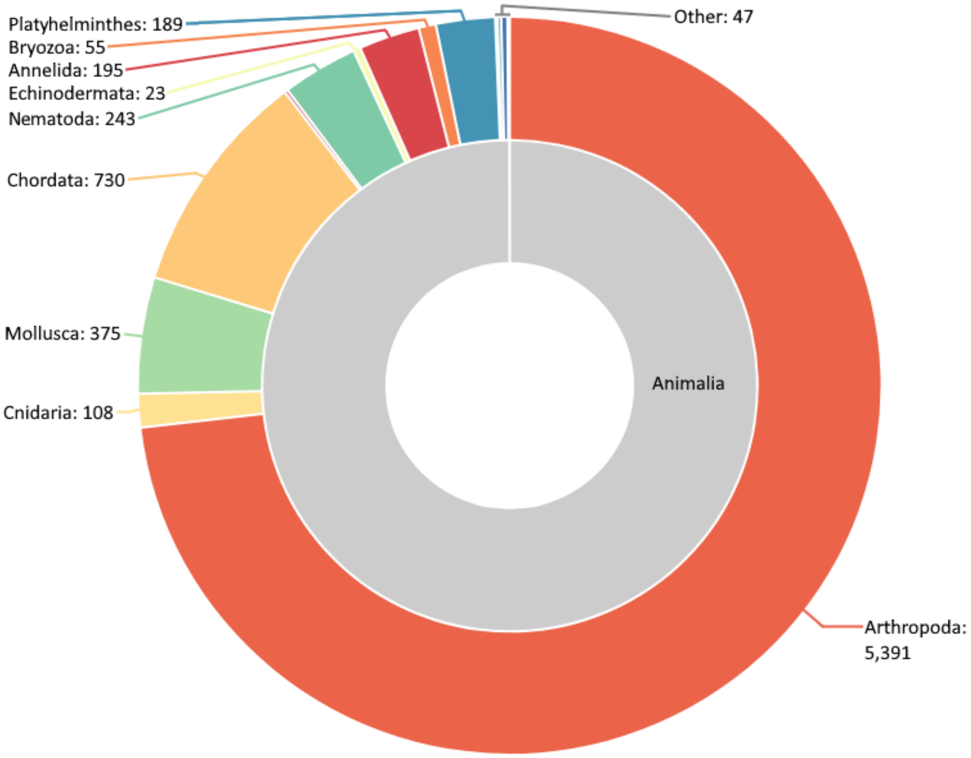


Figure 16: The number of species in the unioned data set categorized by higher taxon (phylum).

The visualization of the unioned data set's taxonomic backbone, as covered by the NSR, and the genetic clusters for each taxon on this backbone provided insight into the specimen data. The reactive Shiny filters in the tree diagram enabled the selection of taxa, which allowed for subsequent analysis of species' distribution (such as the number of specimens per database, number of records per taxonomic group, and the distribution of species within and between clades).

Taxa were assigned a coverage score following their representation against the NSR. If, for example, the genus *Abera* had three accepted species names in the checklist but only one of them was obtained through BOLD/Naturalis, the taxa were assigned a coverage score of 33%. The majority of taxa were represented entirely, as shown in Table 16. Coverage of species showed highest for Mollusca and Chordata (~66%), followed by Cnidaria and Platyhelminthes (~50%), and Annelida, Nematoda, and Arthropoda (20-35%).

Table 16: Overview of unioned Naturalis and BOLD species' coverage against the NSR checklist. The first column, 'coverage score', represents the levels of coverage intervals; the second and third columns respectively portray the number of taxa for each given level and the distribution percentage of species for class Insecta (one of the largest contributors of observed specimens).

Coverage score	**Number of taxa**	**Distribution Insecta**
100%	2,990	54.9%
90-99%	0	0%
80-89%	21	0.6%
70-80%	80	2.1%
60-69%	253	5.5%
50-59%	659	14.0%
40-49%	101	2.4%
30-39%	271	6.5%
20-29%	261	7.0%
10-19%	148	4.4%
<10%	63	2.3%

Quality system performance

Compiling the statistics for the quality system's performance used all (2,580,759) exported records from BOLD, regardless of whether they matched against the NSR checklist.

Incorrect assignments (due to using genera-level taxa to export specimen data) and a lack of information both qualified as a mismatch against the NSR. Records missing taxonomic identification (no species name) or missing reference identification (no authority and/or year of publication) were the two instances resulting in a mismatch caused by a lack of information. About half of all records (55%) lacked information, as shown in Table 17.

Table 17: Overview of the type of mismatch records retrieved from BOLD. The first column, 'type', represents the different categories of mismatch classification; the second column, 'number of records, the number of records for each given category.

Type	Number of records
Total lack of information	1,426,099
Lack of taxonomic identification	636,167
Lack of reference identification	789,932

Record annotation

Any record with a "Mined from GenBank, NCBI" label or lacking reference identification details was isolated, hereafter mentioned as a B-type record. All remaining specimens were listed as A-type records, regardless of whether they matched the NSR checklist. About a third of all records fell into the A-type category, as shown in Table 18.

Table 18: Overview of the type of specimen records, from left to right: All, A-type, and B-type records, with respectively their number of records per category compared against the NSR checklist.

Type of record	All	A-type	B-type
Match	457,356	327,662	129,694
Mismatch	2,123,403	498,192	1,569,728
Total	2,580,759	825,854	1,699,422

The overall completeness of record annotation was highest for A-type records, as shown in Figure 17. The percentage of fully identified records for B-type records was 63%. Almost every record had a good sequence length (500 bp+), less than 1% ambiguous base calls, a BIN assignment, and an annotation with either country or latitude-longitude information.

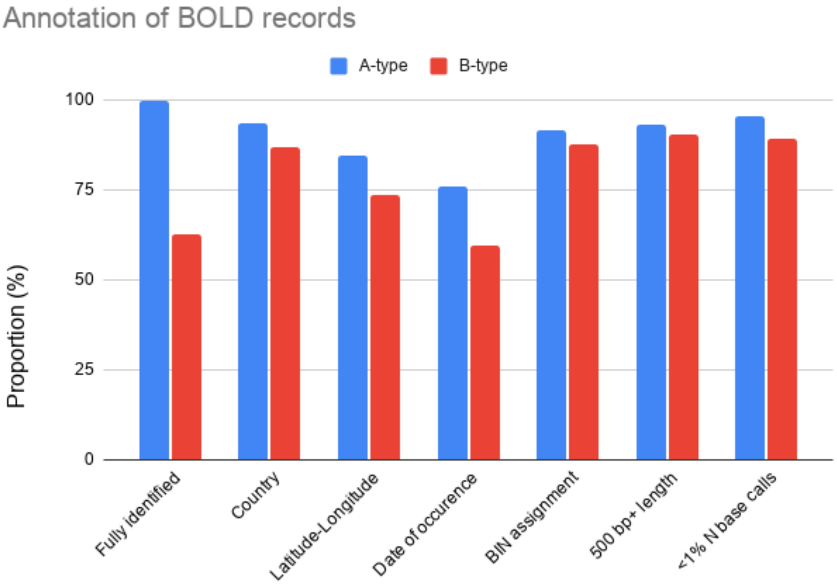


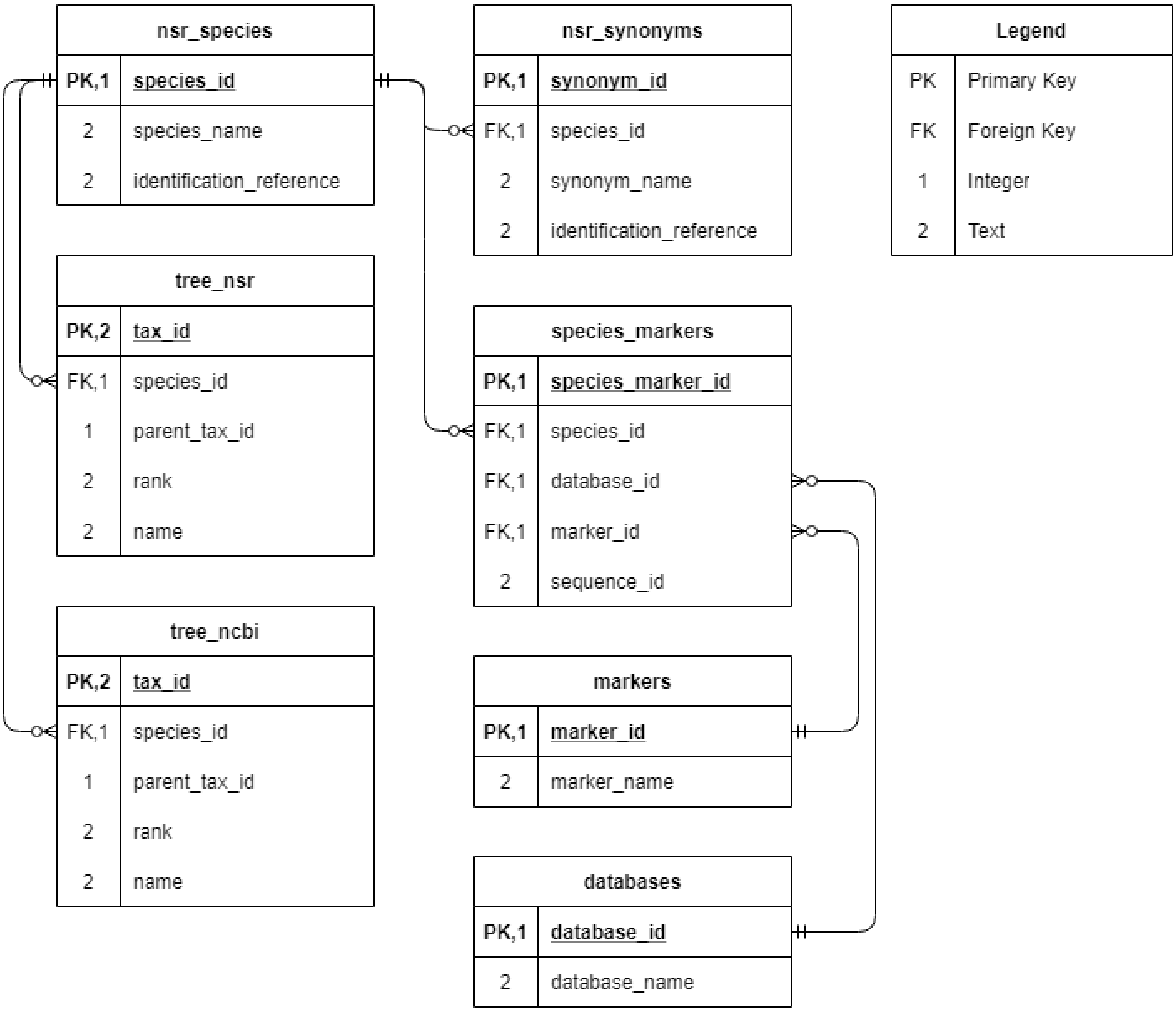
Figure 17: Illustration of BOLD record annotation. The first bars show the proportion of records that are fully identified. The subsequent four bars show the proportion of records with sufficient specimen data regarding country, latitude-longitude annotation, date of occurrence, and a BIN assignment. The final two bars show the proportion of sequence records with a good sequence length (500 bp+) and less than 1% ambiguous base calls.

Data structure

The data structure combined molecular data from various sources through NSR's accepted names and linked sequence information to their taxonomic data. Identifiers marked the relationships and linked data between data sets. This structure consisted of seven CSV files, which included the following data:

- nsr_species - List of species from the Dutch Species Register (NSR).
- nsr_synonyms - List of known synonyms for NSR species names.
- species_markers - Construct which linked sequence data to their corresponding NSR species names and held its relevant details (e.g. sequence ID and marker).
- markers - List of DNA barcoding markers.
- databases - List of databases reflecting sources of collected data.
- tree_nsr - Classification of higher taxa for species names (NSR).
- tree_ncbi - Classification of species names as present in the NCBI Taxonomy database.

Figure 18 depicts the model used to store and define the relationships between data sets.



nsr_species	
PK,1	species_id
2	species_name
2	identification_reference

nsr_synonyms	
PK,1	synonym_id
FK,1	species_id
2	synonym_name
2	identification_reference

Legend	
PK	Primary Key
FK	Foreign Key
1	Integer
2	Text

tree_nsr	
PK,2	tax_id
FK,1	species_id
1	parent_tax_id
2	rank
2	name

species_markers	
PK,1	species_marker_id
FK,1	species_id
FK,1	database_id
FK,1	marker_id
2	sequence_id

tree_ncbi	
PK,2	tax_id
FK,1	species_id
1	parent_tax_id
2	rank
2	name

markers	
PK,1	marker_id
2	marker_name

databases	
PK,1	database_id
2	database_name

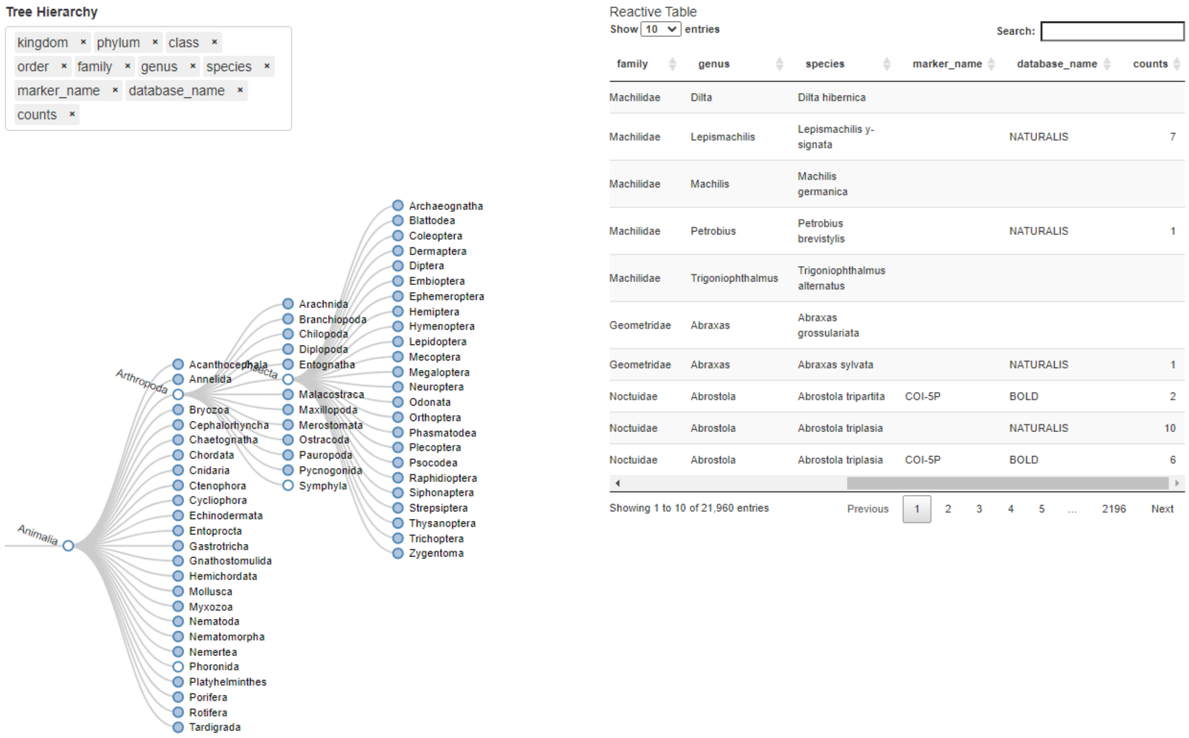
Figure 18: Entity-relationship model representing the flow of information between each of the data sets. Each table is composed of entity types, describing the configuration and type of information stored, with the arrows marking the relationships between those entities.

Reactive d3js visualization

A collapsible tree visualized the taxonomic backbone of NSR species. This diagram used the NSR's entire species classification and included information on specimen data from Naturalis and BOLD. Each species' metadata held its number of specimens along with their respective genetic marker and database.

The objects in the 'Tree Hierarchy' input box controlled the hierarchical order of the metadata. Dragging and dropping, for example, the marker name object to the front resulted in an immediate hierarchical division of records by marker.

The tree diagram's output was reflected in a reactive table for the last selected node, as illustrated in Figure 19 for taxa of the class Insecta (kingdom: Animalia, phylum: Arthropoda). The reactive table showed all taxa within the last selected node (Insecta), along with Naturalis and BOLD specimen data.

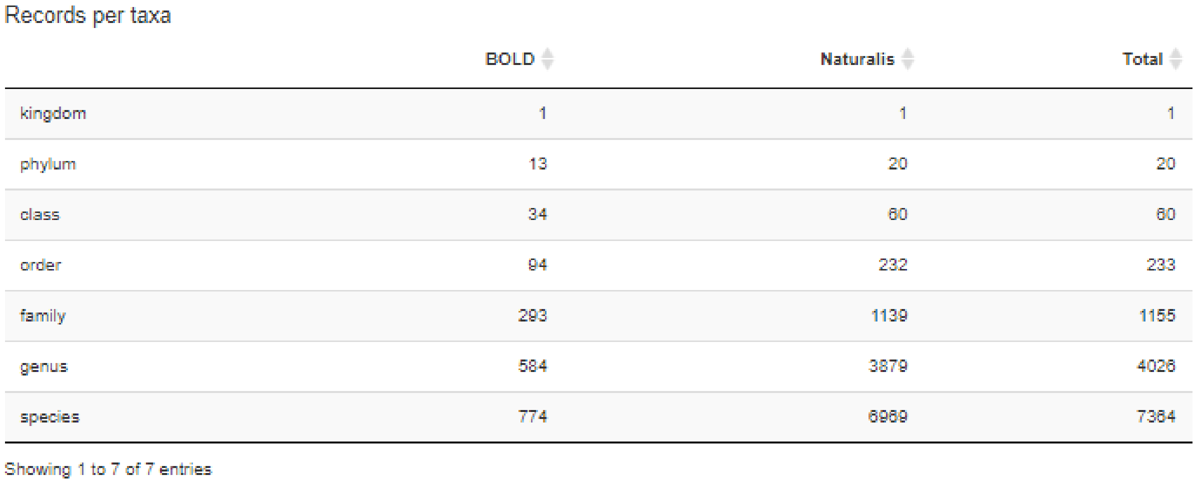


family	genus	species	marker_name	database_name	counts
Machilidae	Dilta	Dilta hibernica			
Machilidae	Lepismachilis	Lepismachilis y-signata		NATURALIS	7
Machilidae	Machilis	Machilis germanica			
Machilidae	Petrobius	Petrobius brevistylis		NATURALIS	1
Machilidae	Trigoniophthalmus	Trigoniophthalmus alternatus			
Geometridae	Abraxas	Abraxas grossulariata			
Geometridae	Abraxas	Abraxas sylvata		NATURALIS	1
Noctuidae	Abrostola	Abrostola tripartita	COI-5P	BOLD	2
Noctuidae	Abrostola	Abrostola triplasia		NATURALIS	10
Noctuidae	Abrostola	Abrostola triplasia	COI-5P	BOLD	6

Showing 1 to 10 of 21,960 entries

Figure 19: Visualization of taxa within the Insecta class for all NSR species and the presence of those species within Naturalis and BOLD. The collapsible tree (bottom left) shows their taxonomic classification in the Linnaean hierarchy, with the adjacent data table (right) its plain text output. The hierarchical input box (top left) controls the order of levels as displayed in the tree diagram.

Various charts visualized the number of occurrence records. Similar to the reactive table, all statistics related to the last selected nodes. The number of unique records per taxa was categorized by taxonomic rank for each database, as shown in Figure 20.



Records per taxa

	BOLD	Naturalis	Total
kingdom	1	1	1
phylum	13	20	20
class	34	60	60
order	94	232	233
family	293	1139	1155
genus	584	3879	4026
species	774	6969	7364

Showing 1 to 7 of 7 entries

Figure 20: Number of all unique records per taxa categorized by taxonomic rank, representing the taxon’s level in the Linnaean hierarchy, as available in each database and the union of databases.

The number of records per database subsequently determined its coverage against the NSR, as shown in Figure 21. A chart for the share of markers of sequence records accompanied this data.

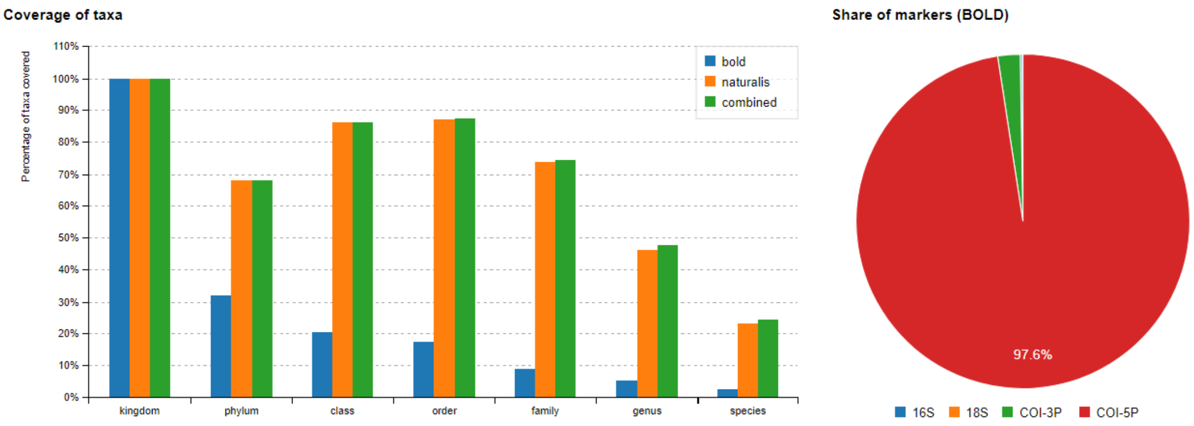
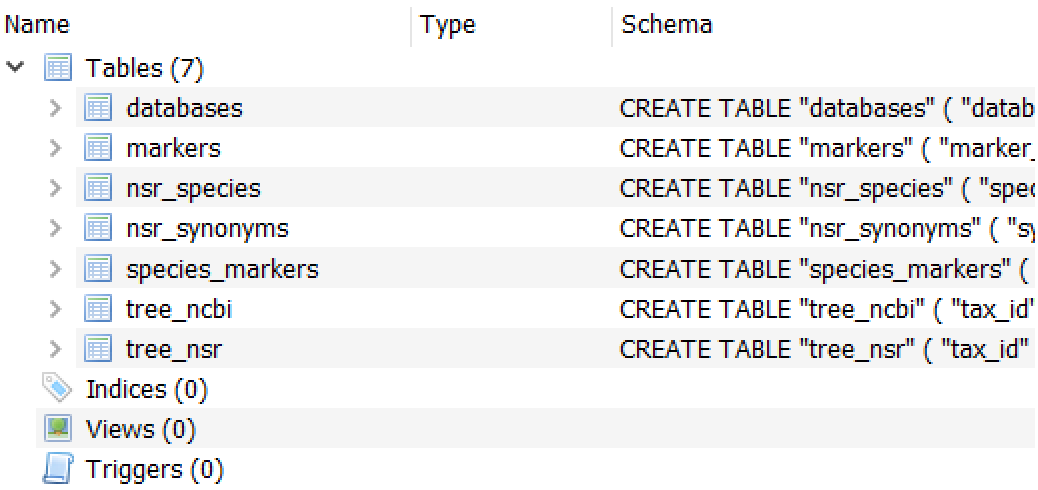


Figure 21: Representation of taxon coverage (left) and share of markers for all taxa (right). Coverage of taxa is categorized by taxonomic rank, representing the taxon’s level in the Linnaean hierarchy, and displays the percentage of each database (BOLD: blue, Naturalis: orange) and union of databases (Combined: green). The share of markers shows the frequency of markers as present in the BOLD records (16S: blue, 18S: orange, COI-3P: green, COI-5P: red).

SQLite Database

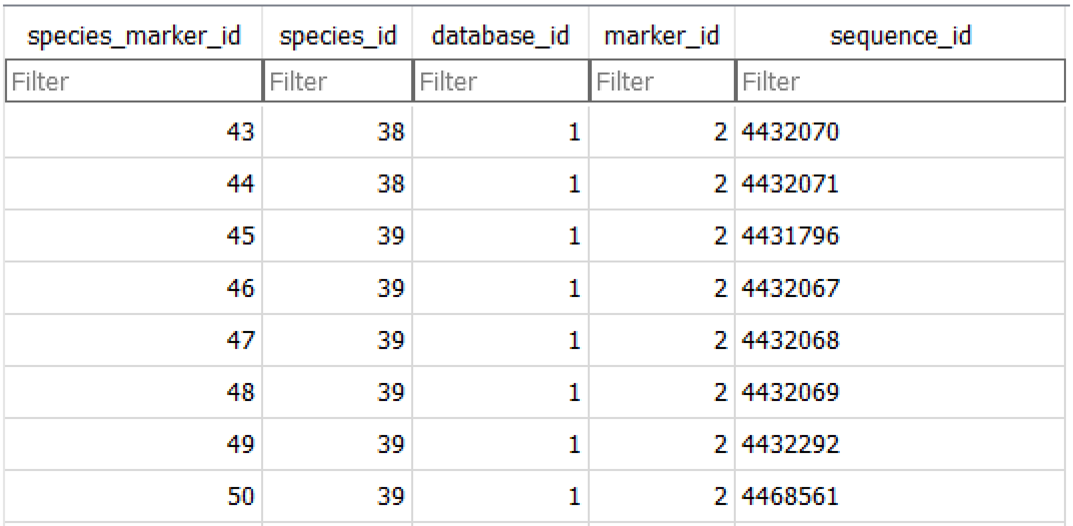
The database, which consisted of seven tables, as shown in Figure 18, allowed for (portable) access to all information. The creation and population of tables used data from the NSR, NCBI, BOLD, and Naturalis, as laid out in the attached SQL schema. Figure 22 depicts an overview of the created tables visualized with the SQLite Browser.



Name	Type	Schema
Tables (7)		
databases		CREATE TABLE "databases" ("datab
markers		CREATE TABLE "markers" ("marker_
nsr_species		CREATE TABLE "nsr_species" ("spec
nsr_synonyms		CREATE TABLE "nsr_synonyms" ("sy
species_markers		CREATE TABLE "species_markers" (
tree_ncbi		CREATE TABLE "tree_ncbi" ("tax_id"
tree_nsr		CREATE TABLE "tree_nsr" ("tax_id"
Indices (0)		
Views (0)		
Triggers (0)		

Figure 22: Overview of all data tables, representing each data set, as defined in the SQL schema, visualized with the SQLite browser.

Figure 23 depicts the species_markers table, which shows a species with multiple specimen records (each with a unique sequence ID) within the BOLD database. Each record's metadata, accessed via the sequence id, could, for example, show a different geographic site for each occurrence of a specimen.



species_marker_id	species_id	database_id	marker_id	sequence_id
Filter	Filter	Filter	Filter	Filter
43	38	1	2	4432070
44	38	1	2	4432071
45	39	1	2	4431796
46	39	1	2	4432067
47	39	1	2	4432068
48	39	1	2	4432069
49	39	1	2	4432292
50	39	1	2	4468561

Figure 23: Illustration of the species_markers table, representing from left to right the identifiers of a unique index (sm_id), the NSR species name (species_id), the database of origin (database_id), the marker name (marker_id), and the sequence ID corresponding to the respective record's metadata.

Discussion and Conclusion

Taxonomic representation

A submitted list of taxa or a reference database that reflects local species is essential for making accurate metabarcoding assignments. In this study, the Dutch Species Register (NSR) provided a custom export of Dutch taxonomy. A selection of taxa from this checklist allowed to retrieve the corresponding records as available internally and in public databases. If specimens do not match, the ability to divert to a different backbone provides additional insight into their classification outside the scope of the checklist or serves as a filter for specific taxa (e.g., marine: WoRMS[76]).

Before performing taxonomic assignments, it is essential to determine if existing databases reflect the targeted taxa. Understanding the composition of the current databases may help inform future work, for example, by targeted DNA barcoding of local species (e.g., ARISE), assessing statistical confidence for taxonomic assignments, and selecting the degree of recorded taxonomic resolution[77].

While assessing this study's local biodiversity, the provided presence status[78] of species was unused. All taxa were deemed retrievable in either internal or public records. A later examination of taxa for *Abdera*, a family of beetles, revealed that two species on the checklist were extinct. While incomplete information on species does not require their removal, the understanding of undetermined species improves accurate assessments. Including the presence status of species could therefore have helped to calculate and interpret assessment results.

Records during this research were retrieved for the Dutch freshwater and saltwater areas, as specified by the Dutch country field in public reference databases. However, a country annotation does not serve as an accurate representation of the North Sea area. Taxa from countries bordering the Netherlands may have also proved useful. The findings of this study do show that records for 363 species can complement the Naturalis data set. These records account for 46% of all Dutch annotated species retrieved from BOLD.

Exporting reference data

The large set of data retrieved from the BOLD database accounts for 26% (7,906) of the species on the NSR checklist. Although 20% of genera returned no records, indicating that species were not recognized, the poor quality of specimen data also ruled out many species for accurate identification.

In addition, the use of genera to retrieve specimen data resulted in the ambiguity of species names as observed by homonyms. Due to different authorities recording those names, homonyms exist between zoology and botany (e.g., *Clusia flava*). The marker parameter was not specified during data retrieval as all markers returned for a specimen matching the search string. Therefore, even if only COI was specified, a record with both COI and ITS resulted in sequence data for both markers. Although later distinguished by respective authorities for their binomial name, the API did not provide a solution.

Opposed to retrieving records via the BOLD Public Data Portal API in Python, data can also be retrieved with BOLD's R client[79]. Record retrieval in R, in addition to CRS and NSR data retrieval via the NBA, could result in a more streamlined pipeline.

Taxon matcher

Matching taxonomic data from various sources - such as BOLD specimen data and sequence records, Naturalis internal specimen records, names and phylogenetic lineages of the NCBI database, and taxonomic classification of species from the NSR - necessitated a system to link and connect corresponding information between data sets. The use of a checklist, whether provided by the NSR or derived from Naturalis data, serves as a foundation and reference for all data sets collected. The integrity of species names is preserved by linking all backbones and sequence records to an identifier of the corresponding taxa of the reference.

The respective sequence record identifiers in the custom database (see Figure 18; species-markers) allow for the extraction of sequences of specific taxa and barcoding regions by database. However, each database has its own identifier corresponding with a sequence record. Merging data sets requires the reformatting of data to make them compatible, with a unique identifier for each sequence. Following this process, sequence records can be extracted, reformatted, and dereplicated before building an indexed reference database with Naturalis data.

After performing a BLAST against the database, multiple taxon hierarchies are available for subsequent taxonomic and functional analysis. However, taxon IDs of the NSR are not available in this study's results. The Naturalis Document Store (which is queried by the NBA) for the Dutch Species Register only includes taxon documents on (sub)species level. Their higher taxonomy facilitates classification purposes and does not possess available taxon IDs to be retrieved via, for example, the NBA (and is subsequently not included in documents generated via the NBA such as the Darwin Core Archives[80]). Although live data should facilitate the taxon IDs, custom one-time/periodic NSR exports are available. Including taxon IDs in the custom database's structure would preserve index terms (when taxons merge) and serve as a more direct reference point for all respective data.

Diversity of specimen data

Analysis of data shows the Arthropoda phylum, which primarily included butterflies and moths, accounted for 89 percent of the unique records to BOLD. As one of the major phyla studied in taxonomy, representing over 80% of all described living animal species and the most significant contributor to the NSR taxa, numerous Arthropoda records were expected. However, Mollusca/Chordata and Nematoda/Arthropoda were the overall best and worst represented groups, respectively. Accounting for 10% of taxa, the coverage difference between chordates and arthropods is enormous, though it follows recent research findings[33].

The biodiversity assessment revealed a wide range of species, corresponding to the distribution of taxa within the NSR checklist, despite only a 24% coverage rate. Additionally, the subsequent assignment of coverage scores to individual species revealed a complete representation for the majority of retrieved taxa.

Concerns about public reference data

All segments of BOLD's specimen data submission protocol[81], as well as sequence record information[82], are available for quality control analysis. The examination of retrieved records revealed that ~73% hold identification to the species rank. However, with 30% of records missing reference information, more than half (55%) of all records lack information on taxonomic or reference identification. While corresponding to recent research findings[30], it is worth noting many GenBank mined records lack a complete identification.

A total of 221,539 (8.6%) sequence records are of insufficient length, with 93,209 (3.6%) records containing more than 1% ambiguous base calls. The imposed criteria for sequence quality are developed based on previous public research[20,30,34,71] and adjust for the use of various genetic markers (e.g., COI-3P and COI-5P). In comparison, databases such as SILVA apply a 2 percent threshold to sequence quality metrics on nucleotides and omit sequences with less than 300 bp. The availability of marker codes in combination with sequencing primers and directions would enable in-depth annotation of sequence data in future works. Additionally, detecting vector contamination would necessitate a sequence similarity search against a vector sequence database or the search for restriction sites.

While adhering to specified sampling areas, the sole use of a country annotation does not always cover the required data and poses difficulty to standardize across data sets. Regions like the North Sea cover multiple countries and are hence regularly classified in the country field. Additional geographic fields (such as province states, regions, sectors, and exact sites) may prove functional for specific areas (which in many records includes North Sea areas). However, they rely on the consistency of their classification. Latitude-longitude data may provide more resolution of record distribution within and between countries and are easier to combine across data sets, but they are more frequently lacking.

Following quality control, records should be compiled based on the assigned grades for species identification, incorporated into the species-marker table (see Figure 18). Opposed to filtering records before creating a custom reference database, it allows the user to download the reference data set with the option of including or excluding grades. Similar to the ambiguous base calls, the collectors, institutions storing specimen data, and sequencing centers can also be flagged if meeting a threshold of low-quality records. These flags provide the exclusion of data for subsequent retrieval iterations.

Visualization, reproducibility, and future works

Aside from providing a reference database, this study presents a structure capable of computing highly visual representations while maintaining the integrity of combined resources for future works. The development of a small web application (e.g., Django[83]) using the provided database schema would allow interactions with the database through object-relational mappings in Python. The software to compile the database is influenced by Galaxy[84], a web-based scientific analyses platform. Galaxy uses SQLite by default, which is compatible with Python. Integration of the database file into Galaxy could make metadata for BLAST hits interactive without leaving the Galaxy environment. Alternative software like MySQL is not supported by Galaxy, though PostgreSQL facilities an actual database server if needed.

Further research into the data set can also be performed by hand using R (RStudio). The RSQLite[85] / dbplyr[86] packages would equally allow (local) interaction with the database. The functionality for retrieving the collection and analysis of data in R has been documented (using R Markdown) and is compatible with an R shiny app. Shiny generates the visualizations from the Rmarkdown file and hosts a standalone app on a (local) webserver.

Prior data reformatting would allow the underlying system to work with records from multiple databases. The sequence ID currently refers to a specific BOLD or Naturalis record and could incorporate, for example, the accession numbers of NCBI records. However, with more taxonomic data and subsequent large phylogenies, improvements to the tree structure indexing system are recommended. Using an algorithm[87] to retrieve all parents for a query 1:1 as opposed to the direct parent taxa for each rank would significantly improve its efficiency.

While the Python script allows a user to retrieve public data on different occasions to reflect changes in public records, additional filters are required. Although imposed rules differ per stakeholder, it is essential that the versions are compared. All matching sequences and taxonomy between iterations do not require changes, but quality control will be required for new or changes to existing records, along with the (manual) inspection/verification of changed and deleted data. As retrieving public reference data for millions of sequences can take a long time, data can alternatively be added gradually for specific taxa or from reliable institutions (with the introduction of stored exclusion/inclusion flags). As both GenBank and BOLD share data on, for example, COI records, a comparison for the discrepancies between the data sets would lead to the most reliable taxonomic coverage. SILVA and UNITE data packages can directly be attached to the pipeline. The public records that complement the Naturalis data can then be combined to create a new reference data set. It is recommended to update the local database before new research (or roughly every few months).

References

1. Rapport DJ, et al. Eco-cultural health, global health, and sustainability. *Ecol. Res*. 2011; Vol. 26: pages 1039–1049. DOI: 10.1007/s11284-010-0703-5.

2. Helfrich AL. Sustaining America’s Aquatic Biodiversity - Why Is Aquatic Biodiversity Declining? *Virginia Tech*. Publication 420-521.
https://pubs.ext.vt.edu/420/420-521/420-521-digital-version.html

3. Borgwardt F, et al. Exploring variability in environmental impact risk from human activities across aquatic ecosystems. *Sci. Tot. Env*. 2019; Vol. 652: pages 1396-1408.
doi: 10.1016/j.scitotenv.2018.10.339.

4. River basin management. European Commission.
https://ec.europa.eu/environment/water/water-framework/index_en.html
Last updated: 04/08/2020. Accessed November 9th, 2020.

5. EU Coastal and Marine Policy. European Commission.
https://ec.europa.eu/environment/marine/eu-coast-and-marine-policy/marine-strategy-framework-directive/index_en.htm
Last updated: 02/07/2020. Accessed November 9th, 2020.

6. Weigand H, et al. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Sci. Tot. Env*. 2019; Vol. 678: pages 499-524. DOI: 10.1016/j.scitotenv.2019.04.247.

7. Carstensen J, et al. Confidence in ecological indicators: a framework for quantifying uncertainty components from monitoring data. *Ecol. Indic.* 2016; Vol. 67: pages 306-317.
doi: 10.1016/j.ecolind.2016.03.002.

8. Hebert PD, et al. Biological identifications through DNA barcodes. *Proc Biol Sci.* 2003; Vol. 270: pages 313–321. DOI: 10.1098/rspb.2002.2218.

9. Bruns TD, et al. Fungal molecular systematics. *Annu. Rev. Ecol. Syst.* 1991; Vol. 22: pages 525–564. DOI: 10.1146/annurev.es.22.110191.002521.

10. Woese CR. Bacterial evolution. *Micro. Rev.* 1987; Vol. 51(2): pages 221-271.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC373105/pdf/microrev00049-0051.pdf

11. Purty RS, et al. DNA Barcoding: An Effective Technique in Molecular Taxonomy. *Austin J. Biotechnol. Bioeng.* 2016; Vol. 3(1): page 1059.
https://austinpublishinggroup.com/biotechnology-bioengineering/v3-i1.php

12. Pavan-Kumar A, et al. DNA Metabarcoding: A New Approach for Rapid Biodiversity Assessment. *J. Cell Sci. Mol. Bio*. 2015; Vol. 2(1): page 111.
https://www.researchgate.net/publication/279885595_DNA_Metabarcoding_A_New_Approach_for_Rapid_Biodiversity_Assessment

13. Porter TM, et al. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Mol. Ecol*. 2018; Vol. 27: pages 313– 338. DOI: 10.1111/mec.14478.

14. Callahan BJ, et al. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*. 2017; Vol. 11: pages 2639–2643.
doi: 10.1038/ismej.2017.119.

15. Altschul SF, et al. Basic local alignment search tool. *J. Mol. Biol*. 1990; Vol. 215(3): pages 403-410. DOI: 10.1016/S0022-2836(05)80360-2.

16. Clark K, et al. GenBank. *Nucleic. Acids. Res.* 2016; Vol. 44(D1): pages 67-72.
doi: 10.1093/nar/gkv1276.

17. Crocetta F, et al. Does GenBank provide a reliable DNA barcode reference to identify small alien oysters invading the Mediterranean Sea? *J. Mar. Biol. Assoc. U. K.* 2015; Vol. 95: pages 111–122. DOI: 10.1017/S0025315414001027.

18. Smith BE, et al. From GenBank to GBIF: phylogeny-based predictive niche modeling tests accuracy of taxonomic identifications in large occurrence data repositories. *PLoS ONE*. 2015; Vol. 11(3). DOI: 10.1371/journal.pone.0151232.

19. Balakirev E, et al. Complete mitochondrial genomes of the Cherskii's sculpin *Cottus czerskii* and Siberian taimen *Hucho taimen* reveal GenBank entry errors: incorrect species identification and recombinant mitochondrial genome. *Evol. Bioinformatics*. 2017; Vol. 13. DOI: 10.1177/1176934317726783.

20. Ratnasingham S, & Hebert PD. BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Resour.* 2007; Vol. 7: pages 355–364.
doi: 10.1111/j.1471-8286.2007.01678.x.

21. Pentinsaari M, et al. BOLD and GenBank revisited – Do identification errors arise in the lab or in the sequence libraries?. *PLOS ONE*. 2020; Vol. 15(4): e0231814.
doi: 10.1371/journal.pone.0231814.

22. Quast C, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic. Acids. Research*. 2013; Vol. 41: pages D590-D596. DOI: 10.1093/nar/gks1219.

23. Kõljalg U, et al. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist.* 2005; Vol. 166 (3): pages 1063-1068. DOI: 10.1111/j.1469-8137.2005.01376.x.

24. GBIF: The Global Biodiversity Information Facility (2021). What is GBIF?.
https://www.gbif.org/what-is-gbif
Last updated: January 13th, 2020. Accessed March 21st, 2021.

25. Kingdoms of Life Being Barcoded. BOLD Systems.
http://www.boldsystems.org/index.php/TaxBrowser_Home
Last updated: -. Accessed March 20th, 2021.

26. UNITE - statistics. UNITE community. https://unite.ut.ee/statistics.php
Last updated: January 15th, 2020. Accessed November 9th, 2020.

27. SILVA. Release 138.1 https://www.arb-silva.de/documentation/release-1381/
Last updated: August 27th, 2020. Accessed February 13th, 2021.

28. Sonet G, et al. Utility of GenBank and the Barcode of Life Data Systems (BOLD) for the identification of forensically important Diptera from Belgium and France. *ZooKeys*. 2013; Vol. 365: pages 307-328. DOI: 10.3897/zookeys.365.6027.

29. Meiklejohn KA, et al. Assessment of BOLD and GenBank - Their accuracy and reliability for the identification of biological materials. *PLoS One*. 2019; Vol. 14(6): e0217084.
doi: 10.1371/journal.pone.0217084.

30. Porter TM, et al. Over 2.5 million COI sequences in GenBank and growing. *PLoS One*. 2018; Vol. 13(9): e0200177. DOI: 10.1371/journal.pone.0200177.

31. Lesack K, et al. Nomenclature Errors in Public 16S rRNA Gene Reference Databases. *bioRxiv*. 2018: 441576. DOI: 10.1101/441576.

32. Nilsson RH, et al. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS One*. 2006; Vol. 1: e59. DOI: 10.1371/journal.pone.0000059.

33. Curry CJ, et al. Identifying North American freshwater invertebrates using DNA barcodes: are existing COI sequence libraries fit for purpose? *Freshw. Sci*. 2018; Vol. 37(1): pages 178–189. DOI: 10.1086/696613.

34. Bridge PD, Roberts PJ, Spooner BM, et al. On the unreliability of published DNA sequences. *New Phytol.* 2003; Vol. 160: pages 43–48.
doi: 10.1046/j.1469-8137.2003.00861.x.

35. Ashelford KE, et al. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol*. 2005; Vol. 71: pages 7724–7736. DOI: 10.1128/AEM.71.12.7724-7736.2005.

36. Crocetta F, et al. Does GenBank provide a reliable DNA barcode reference to identify small alien oysters invading the Mediterranean Sea? *J. Mar. Biol. Assoc. U.K.* 2015; Vol. 95: pages 111–122. DOI: 10.1017/S0025315414001027.

37. Seah YG, et al. Levels of COI divergence in Family Leiognathidae using sequences available in GenBank and BOLD Systems: A review on the accuracy of public databases. *Aquac. Aquar. Conserv. Legis. Int. J. Bioflux Soc.* 2017; Vol. 10: pages 391–401.
http://www.bioflux.com.ro/docs/2017.391-401.pdf

38. Pentinsaar M, et al. BOLD and GenBank revisited – Do identification errors arise in the lab or in the sequence libraries? *PLoS One*. 2020: Vol. 15(4); e0231814.
doi: 10.1371/journal.pone.0231814.

39. Vilgalys R. Taxonomic misidentification in public DNA databases. *New Phytol*. 2003; Vol. 160: pages 4–5. DOI: 10.1046/j.1469-8137.2003.00894.x.

40. Contamination in Sequence Databases. NCBI.
https://www.ncbi.nlm.nih.gov/tools/vecscreen/contam/
Last updated: -. Accessed March 31st, 2021.

41. Hazkani-Covo E, et al. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet*. 2010; Vol. 6(2): e1000834.
doi: 10.1371/journal.pgen.1000834.

42. Hassanin A, et al. Comparisons between mitochondrial genomes of domestic goat (Capra hircus) reveal the presence of numts and multiple sequencing errors. *Mitochondrial DNA.* 2010; Vol. 21(3-4): pages 68-76. DOI: 10.3109/19401736.2010.490583.

43. Grau ET, et al. Survey of mitochondrial sequences integrated into the bovine nuclear genome. *Sci Rep*. 2020; Vol. 10: 2077. DOI: 10.1038/s41598-020-59155-4.

44. Wolff JN, et al. Selective enrichment and sequencing of whole mitochondrial genomes in the presence of nuclear encoded mitochondrial pseudogenes (numts). *PLoS One*. 2012; Vol. 7(5): e37142. DOI: 10.1371/journal.pone.0037142.

45. Song H, et al. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc. Natl. Acad. Sci. USA*. 2008; Vol. 105: pages 13486–13491. DOI: 10.1073/pnas.0803076105.

46. Identification Engine API. BOLD Systems.
http://www.boldsystems.org/index.php/resources/api?type=idengine
Last updated: -. Accessed November 9th, 2020.

47. BLAST: Basic Local Alignment Search Tool. National Center for Biotechnology Information. https://blast.ncbi.nlm.nih.gov/Blast.cgi
Last updated: -. Accessed November 9th, 2020.

48. Bartolo AG, et al. The current state of DNA barcoding of macroalgae in the Mediterranean Sea: presently lacking but urgently required. *Botanica. Marina*. 2020: Vol. 63(3); pages 253-272. DOI: 10.1515/bot-2019-0041.

49. The collections of Naturalis. Naturalis Biodiversity Center.
https://www.naturalis.nl/en/collections-of-naturalis.
Last updated: -. Accessed March 21st, 2021.

50. About | BioPortal. Naturalis Biodiversity Center.
https://bioportal.naturalis.nl/about?language=en&back.
Last updated: -. Accessed March 21st, 2021.

51. About | Arise Biodiversity | The Netherlands. Arise Biodiversity.
https://www.arise-biodiversity.nl/about Published: April 29th, 2020. Accessed April 4th, 2021.

52. Overview of biodiversity of the Netherlands. Nederlands Soortenregister.
https://www.nederlandsesoorten.nl/node/374
Last updated: -. Accessed November 9th, 2020.

53. EIS-Kenniscentrum Insecten. Eis-Nederland.
https://www.eis-nederland.nl/
Last updated: -. Accessed November 9th, 2020.

54. Hebert PDN, et al. Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 2003; Vol. 270: pages 313–321. DOI: 10.1098/rspb.2002.2218.

55. Python Software Foundation. Python Language Reference, version 3.8.
http://www.python.org
Last updated: -. Accessed November 9th, 2020.

56. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/
Last updated: -. Accessed November 9th, 2020.

57. Roche A. (2018). Taxon parser: A python library to parse taxon names into elementary components. https://github.com/aroche/taxon_parser. Accessed November 9th, 2020.

58. GBIF: The Global Biodiversity Information Facility (2021). The core GBIF scientific name parser library. https://github.com/gbif/name-parser. https://www.gbif.org/tools/name-parser. Accessed November 9th, 2020.

59. Public Data Portal API. BOLD Systems.
http://boldsystems.org/index.php/resources/api?type=webservices
Last updated: -. Accessed November 9th, 2020.

60. Netherlands Biodiversity Data services. https://docs.biodiversitydata.nl/en/latest/.
Last updated: -. Accessed February 23rd, 2021.

61. Hannes Hettling and Rutger Vos (2021). nbaR: R Package Client for the Netherlands Biodiversity API. R package version 0.1.0. https://docs.ropensci.org/nbaR, https://github.com/ropensci/nbaR. Accessed February 23rd, 2021.

62. Scott Chamberlain and Zebulun Arendsee (2021). taxizedb: Tools for Working with 'Taxonomic' Databases. R package version 0.3.0.
https://CRAN.R-project.org/package=taxizedb

63. Hadley Wickham, Maximilian Girlich, and Edgar Ruiz (2021). dbplyr: A ‘dplyr’ Back End for Databases. R package version 2.1.0. https://CRAN.R-project.org/package=dbplyr

64. Reingold EM and Tilford JS. Tidier Drawings of Trees. *IEEE Trans. Softw. Eng.* 1979; Vol. SE-5(5): pages 514-520. DOI: 10.1109/TSE.1979.234212.

65. Jonathan Sidi (2020). d3Tree: Create Interactive Collapsible Trees with the JavaScript’ D3’ Library. R package version 0.2.2. https://CRAN.R-project.org/package=d3Tree

66. Winston Chang, Joe Cheng, JJ Allaire, et al (2020). shiny: Web Application Framework for R. R package version 1.5.0. https://CRAN.R-project.org/package=shiny

67. Yihui Xie, Joe Cheng, and Xianying Tan (2021). DT: A Wrapper of the JavaScript Library’ DataTables’. R package version 0.17. https://CRAN.R-project.org/package=DT

68. Victor Perrier and Fanny Meyer (2020). billboarder: Create Interactive Chart with the JavaScript ‘Billboard’ Library. R package version 0.2.8. https://CRAN.R-project.org/package=billboarder

69. Hipp, R. D. (2020). SQLite. https://www.sqlite.org/index.html
Last updated: -. Accessed February 13th, 2021.

70. DB Browser for SQLite. SQLite Browser. https://sqlitebrowser.org/
Last updated: -. Accessed February 13th, 2021.

71. Nilsson RH, et al. Approaching the taxonomic affiliation of unidentified sequences in public databases–an example from the mycorrhizal fungi. *BMC Bioinformatics*. 2005; Vol. 6: 178. DOI: 10.1186/1471-2105-6-178.

72. Costa FO, et al. A Ranking System for Reference Libraries of DNA Barcodes: Application to Marine Fish Species from Portugal. *PLoS One*. 2012: Vol. 7(4); e35858. doi: 10.1371/journal.pone.0035858.

73. Oliveira LM, et al. Assembling and auditing a comprehensive DNA barcode reference library for European marine fishes. *J Fish Biology*. 2016: Vol. 89(6); pages 2741–2754. doi: 10.1111/jfb.13169.

74. Fontes JT, et al. BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries. *Mol. Ecol. Resour*. 2021: Vol 21; pages 573– 583. DOI: 10.1111/1755-0998.13262.

75. NCBI Taxonomy Site Guide. NCBI. https://www.ncbi.nlm.nih.gov/guide/taxonomy/
Last updated: -. Accessed January 31st, 2021.

76. WoRMS Editorial Board (2020). World Register of Marine Species. http://www.marinespecies.org. Last updated: -. Accessed April 8th, 2021.

77. Porter TM, Hajibabaei M. Automated high throughput animal CO1 metabarcode classification. *Sci Rep*. 2018; Vol. 8: 4226. DOI: 10.1038/s41598-018-22505-4.

78. Statuscodes voorkomen in Nederland. Nederlands Soortenregister. https://www.nederlandsesoorten.nl/content/statuscodes-voorkomen-nederland.
Last updated: -. Accessed April 8th, 2021.

79. Chamberlain, S. (2019). bold: Interface to Bold Systems API. R package version 1.1.0. https://cran.r-project.org/web/packages/bold/index.html

80. Wieczorek J, et al. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*. 2012: Vol. 7(1): e29715. DOI: 10.1371/journal.pone.0029715.

81. Specimen Data Submissions Protocol. BOLD Systems. https://v3.boldsystems.org/index.php/resources/handbook?chapter=3_submissions.html&section=data_submissions. Last updated: -. Accessed March 29th, 2021.

82. Sequence Submission Protocol. BOLD Systems. https://v3.boldsystems.org/index.php/resources/handbook?chapter=3_submissions.html&section=sequence_uploads. Last updated: -. Accessed March 29th, 2021.

83. Django (Version 3.2) (2021). https://djangoproject.com.
Last updated: -. Accessed April 8th, 2021.

84. Production Environments. Galaxy Project. https://docs.galaxyproject.org/en/master/admin/production.html
Last updated: -. Accessed April 8th, 2021.

85. Kirill Müller, Hadley Wickham, David AJ, et al (2021). RSQLite: 'SQLite' Interface for R. R package version 2.2.3. https://CRAN.R-project.org/package=RSQLite

86. Hadley Wickham, Maximilian Girlich, and Edgar Ruiz (2021). dbplyr: A 'dplyr' Back End for Databases. R package version 2.1.1. https://CRAN.R-project.org/package=dbplyr.

87. Vos RA, DBTree: Very large phylogenies in portable databases. *Methods in Ecology and Evolution*. 2020: Vol. 11(3); pages 457-463. DOI: 10.1111/2041-210X.13337.