Repair what is broken:

Digital specimens and sequence information

MSc Research Project Report, Biology, Leiden University

Student name: Niki Kyriakopoulou
Student number: s2315653
Master specialisation: Biodiversity and Sustainability
Name and address of the research institute: Naturalis Biodiversity Center, Darwinweg 2, 2333 CR Leiden, The Netherlands
Department: ISBI (International Collaboration on Biodiversity Infrastructures)
Responsible supervisor: Dr. Dimitris Koureas, Head of the ISBI department, Distributed System of Scientific Collections (DiSSCo) Coordinator
Daily supervisor: Dr. Niels Raes, NLBIF Node Manager, Senior Project Leader Internationalization, and DiSSCo-NL National Node Manager
Start - End date: 16/03/2020 - 18/08/2020
Total ECs: 30
Contact: niki.krp89@gmail.com, n.kyriakopoulou@umail.leidenuniv.nl
Format: Research Ideas and Outcomes

Abstract

The volume and diversity of information derived from Natural Science Collections (NSCs) has been exponentially increasing due to the emergence of new tools in Information Technology. As a result, recent world-class Research Infrastructures for natural science such as the Distributed System of Scientific Collections, aim to digitally unify all European collections under common curation and access policies and practices that will make the data easily Findable, Accessible, Interoperable and Reusable (FAIR). A holistic approach is now required, where open access to cross-linked information will be the basis of future research. This effort is currently severely limited by incomplete and/or broken links between major data sources about the natural world and by the lack of knowledge on the technological means to facilitate data exploitation and use. This research project aims to investigate the degree to which links between genetic sequence records deposited at data repositories such as ENA and GenBank and the physical specimens stored in NSCs from where the sequence data were derived are broken and analyse the underlying reasons. An R workflow was created that queried the ENA data portal through its Application Programming Interface (API). An in-depth analysis of the Darwin Core Triplet - formatted source identifiers of 1.4 million sequences indicated the extent to which sequence data resolved to the specimens from where they were extracted and highlighted the reasons why the linkages were fragmented. Three recently published papers on DNA analyses were studied to examine how the identifiers are used in practice by various scientific communities today. The results point to the lack of careful curation of identifiers before deposited in the data systems. Guidelines about how to construct them are

often not followed and the provision of source pointers is frequently overlooked, making it impossible to relink sequence data to their physical specimens. Recommendations for better practices were provided.

Keywords

Natural Science Collections (NSCs), Distributed System of Scientific Collections (DiSSCo), FAIR principles, identifiers, Darwin Core Triplets, specimens, R workflow, European Nucleotide Archive (ENA), NCBI GenBank

1. Introduction

1.1. Recent advancements

Throughout the past decades we observed a substantial shift of the traditional research applications in the field of natural sciences towards new modern techniques and automation. Species distribution modelling/ ecological niche modelling, remote sensing and high-throughput sequencing technologies in genomics represent some of the breakthroughs that contribute to address high-priority issues such as the spread of infectious diseases, predicting the effects of global climate and land use change, effective conservation planning, global sustainability, world food and health security, as well as conserving ecosystem services. These issues are ultimately based on how the ecosystem really works, with complex processes and biological interactions that take place over a wide range of time scales and hierarchical levels of organisation (from genes to ecosystems) (*Hardisty et al. 2013*).

The recent technological advances in genomics such as DNA barcoding *(Hebert and Gregory 2005*) coordinated by the International Barcode of Life (iBOL), whole genome sequencing, the emergence of proteomics and metabolomics, new imaging methods (e.g. Computer Tomography or CT) *(Hardisty et al. 2020*), have produced significant quantities of data which along with chemical, morphological and geo-spatial information, bring the traditional "species" term to a higher level of knowledge. This assemblage of biodiversity data has been held in the Natural Science Collections (NSC) that hold approximately 3 billion specimens globally (*Koureas and Addink 2017*). NSCs represent a unique resource for scientific research in multiple ways. They are interconnected creating a worldwide interdisciplinary network that has enabled the emergence of modern uses of biodiversity data and an ever-increasing number of new users from various scientific fields. Some examples are as follows: the contribution of frozen tissue collections in genomic studies (*Schindel and Cook 2018*), the digitisation and georeferencing of specimen records (*Heerlien et al. 2015*) as well as the data sharing through data portals such as the Global Biodiversity Information Facility (GBIF) and the Encyclopedia of Life (EOL), the use of georeferenced records for modelling species distributions with the purpose of selecting conservation areas and predicting changes in species ranges as a result of climate change and species invasions (*Hannah et al. 2020*), and the availability of online, public datasets containing digital images, traits and measurements of specimens stored in NSCs (GenBank for nucleotide sequences, GBIF and Morphbank for specimen-based images

etc.) that allows for cost-effective and relatively effortless analysis of biodiversity data (*Schindel and Cook 2018*).

On a continental level, the European NSCs collectively maintain an estimated 1.5 billion specimens that stand for 80% of the world's bio- and geo-diversity (*Addink et al. 2019*), a small fraction of which has been converted into digital form (about 10%) (https://icedig.eu/). Nevertheless, there is still much work to be done regarding the validation of scientific species names that constitute the primary identifiers of the specimens in NSCs, as well as connecting these with highly scattered data derived from NSCs and produced by contemporary technological advances in different scientific fields *(Koureas and Addink 2017*; *Addink et al. 2019*; *Hardisty et al. 2020*).

1.2. Linking specimen data

Towards this direction, a new Research Infrastructure (RI) that acts on a pan-European level has been formed, the Distributed System of Scientific Collections (DiSSCo). The primary goal of DiSSCo is to bring together NSCs held by more than 120 Natural History Museums (NHMs), botanical gardens and universities across 21 European Countries (https://www.dissco.eu/) so as to mobilise their NSCs to the digital domain and persistently link all derived information, under the umbrella of common digital practices and policies in curation, data sharing and open data access over different scientific fields (*Lannom et al. 2020*; *Hardisty et al. 2020*). Several EU-funded projects (ICEDIG[1], SYNTHESYS+[2], MOBILISE COST Action[3], ENVRI-FAIR[4], CoL+[5]) have been implemented. Their results are integrated in DiSSCo Prepare project, the Preparatory Phase of DiSSCo (Naturalis is a member of the DiSSCo RI and houses the DiSSCo Coordination and Support Office (CSO)). This phase is essential in order to achieve the desirable Implementation Readiness Level (IRL) of the DiSSCo RI in 2023 which is necessary to start constructing the DiSSCo RI (https://www.dissco.eu/).

At the epicentre of DiSSCo's mission, lies the concept of physical objects or specimens held by NHMs, how their quality will be safeguarded and the creation of a knowledge structure to support evidence-based bio- and geo-diversity scientific studies (https://www.dissco.eu/). Currently, researchers need either to borrow these specimens or visit corresponding collections physically or digitally, but the unavailability of linkages between relevant information derived from specimens (images, DNA sequence data, trait measurements, etc.) and their metadata (the description of the specimens) that establish the content of the deposition greatly obstructs research. These linkages are depicted as paths connecting biodiversity data objects in the biodiversity knowledge graph (**Fig. 1**) by *Page* (*2016*), shedding light on possible problems that might emerge. Tracing these paths can help answer many scientific questions for example, the construction of a phylogenetic

[1] https://icedig.eu/
[2] https://www.synthesys.info/
[3] https://www.mobilise-action.eu/
[4] https://envri.eu/home-envri-fair/
[5] https://www.catalogueoflife.org/

tree highlighting the evolutionary relationships among taxa by utilising nucleotide sequences extracted from vouchered specimens which are stored in an NSC.

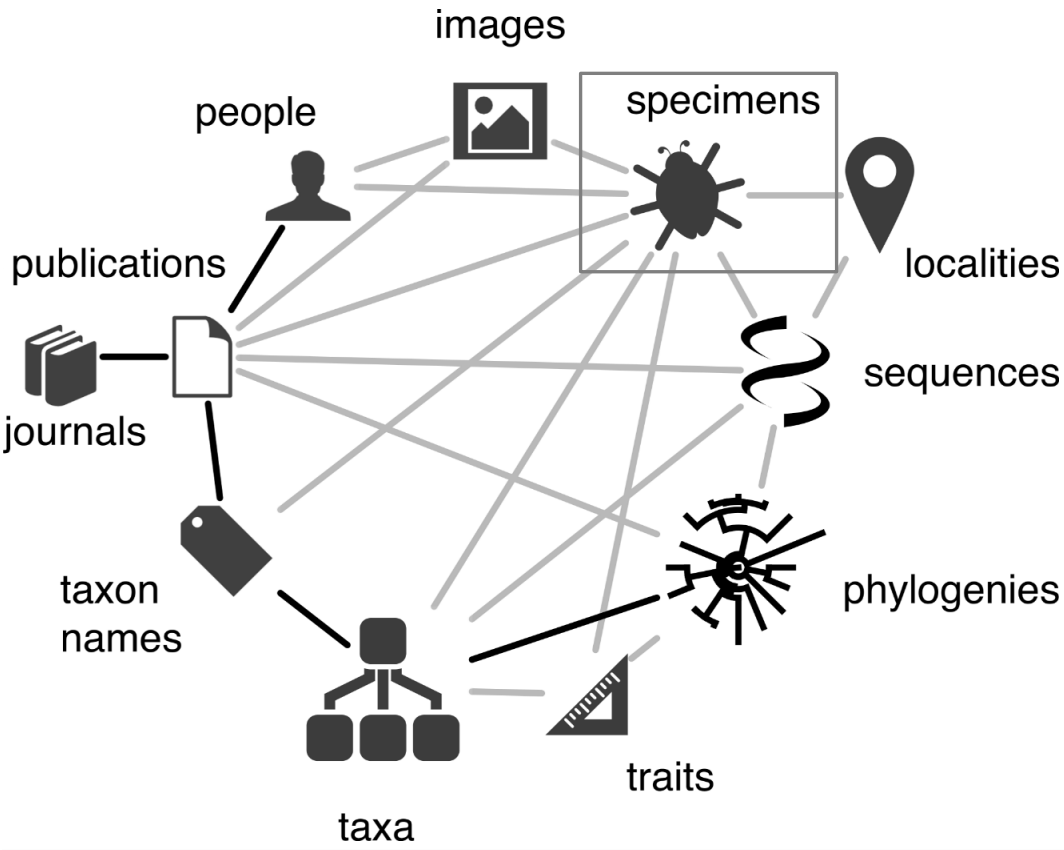

Figure 1.
Biodiversity knowledge graph (from *Page, 2016*). The position of the specimens is highlighted in the graph.

Towards the resolution of this matter, the Digital Object Architecture (DOA) concept "offers a way of grouping, managing and processing fragments of information relating to a natural science specimen" (*Hardisty et al. 2019*). A new approach discussed by *De Smedt et al.* (*2020*) is the FAIR Digital Object (FDO) that meets the FAIR Guiding Principles of Findable, Accessible, Interoperable and Reusable. FDOs are assemblages of data and metadata on specimens or collections, stored in repositories, that are findable and accessible via Globally Unique Identifiers (GUIDs). According to *De Smedt et al.* (*2020*), a GUID has the form of a Uniform Resource Identifier or URI, that is used within the Web, to redirect to the location of the web resource in demand. These identifiers are expected to be a) persistent; i.e., they always direct to a specific object and all the information linked to it, and b) actionable; i.e., they provide access to the identified object via mechanisms and services used by clients *(TDWG [Biodiversity Information Standards] 2013*). To sum up, Persistent Identifiers or PIDs offer fast access to data for various users, from researchers to policymakers, while satisfying the FAIR principles. The interaction with PIDs has been envisaged being

reliable and stable over long periods of time, regardless of continuous future technological advancements (*Hardisty et al. 2019*; *De Smedt et al. 2020*).

NSCs include pieces of heterogeneous data types such as occurrences, traits, DNA sequence data, images, research articles etc. scattered all over, acquired from the examination and analysis of physical objects. These pieces can theoretically be re-assembled within a digital vessel, for instance an FDO, in a robust and organised way (*Koureas et al. 2018*). Nevertheless, the type of identifiers used by the biodiversity data providers has not been consistent. Since the Darwin Core (DwC) Triplet (institutionCode:collectionCode:catalogNumber format as described in *Darwin Core and RDF/OWL Task Groups (2015)*) has become one of the typical means of making statements about resources in bio-collections, each physical object is associated with an institution code, a collection code and a catalog number delimited by a colon (":") (metadata) (*Guralnick et al. 2014; Guralnick et al. 2015*). This practice has shed light upon pitfalls that can be encountered when making specific queries in repositories and databases. For example, a triplet might be linked to different objects at different times as a result of re-allocation of catalogue numbers during reorganisation of a database. The lack of resolution regarding taxa names can also lead to single objects that may have several, different identifiers (*GBIF 2009*). Another type of limitation is the lack of adherence to the advised specifications during the construction of the identifiers such as the use of different types of delimiters. The distribution and curation of identifiers are rarely realised efficiently, often leading to semantic and syntactic errors and their subsequent inadequacy to relink data (*Guralnick et al. 2014*). All things considered; DwC Triplets do not have the characteristics of a proper Globally Unique Identifier (GUID) since they consist of a series of local identifiers that is susceptible to fragmentation as the metadata changes.

1.3. How databases contribute to tracing genetic data back to their specimens?

There have been numerous initiatives in public domain data sharing that strive for the provision of access to data and metadata obtained from nucleotide sequencing and DNA barcoding of specimens kept in NSCs. One of the most recognised collaborations has been the International Nucleotide Sequence Database Collaboration (INSDC) (www.insdc.org) that includes GenBank[6] (maintained by the National Center for Biotechnology Information or NCBI), the European Nucleotide Archive (ENA)[7] (developed and maintained at the European Molecular Biology Laboratory's European Bioinformatics Institute or EMBL-EBI) and the DNA Data Bank of Japan (Bioinformation and DDBJ Center)[8] (**Fig. 2**). Nucleotide sequence data, shared by these three providers, is accompanied with accession numbers and annotations, ensuring that the submission of data follows the FAIR principles for data sharing (*Karsch-Mizrachi et al. 2018*). ENA is also part of the European Life-Science Infrastructure (ELIXIR), an intergovernmental organisation that unites life science resources from across Europe. These resources refer to databases, software tools, training materi-

[6] https://www.ncbi.nlm.nih.gov/genbank/
[7] https://www.ebi.ac.uk/ena
[8] https://www.ddbj.nig.ac.jp/index-e.html

als, cloud storage and supercomputers (https://elixir-europe.org/). The aim of ELIXIR is to coordinate these resources and bring them together in a single infrastructure, an initiative that is in line with DiSSCo's cause. Additionally, the Barcode of Life Data System (BOLD) functions as the central informatics platform of the International Barcode of Life (iBOL) project. It is a DNA barcode library with more than a million public records, that are easily traceable and accessible through identifiers (http://www.boldsystems.org/).

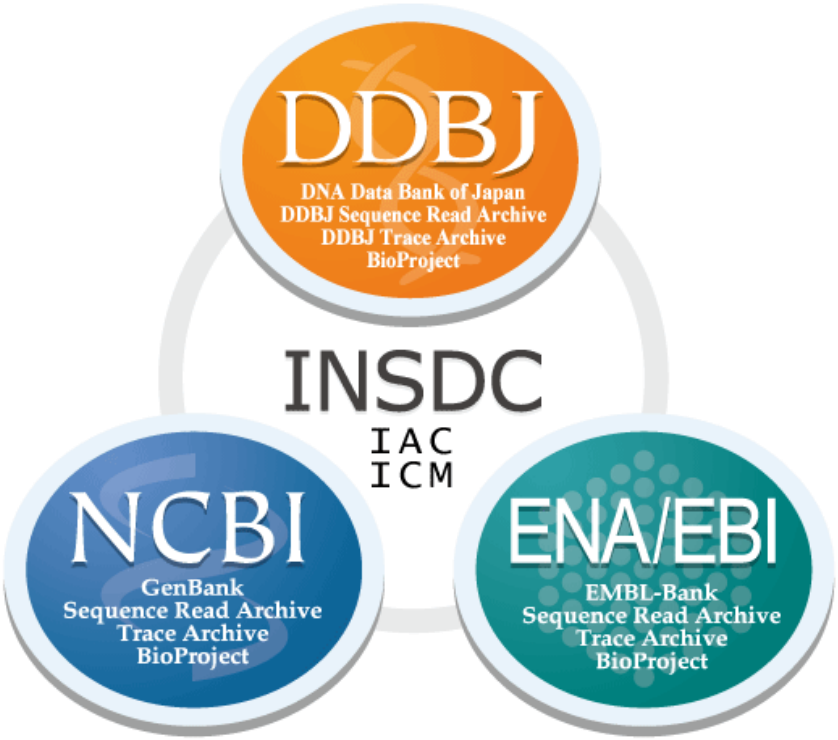

Figure 2.
The collaboration framework depicting the unified nucleotide sequence database INSD (International Nucleotide Sequence Database) (source: https://www.ddbj.nig.ac.jp/insdc-e.html).

The INSDC database, during the registration of nucleotide sequences, accepts several types of identifiers that indicate where the source material of the retrieved sequences is deposited (museum, herbarium or reference collection). Not all sequence records point to specimen vouchers or cultures stored in a collection such as those extracted from uncultured/unvouchered bacterial/archaeal, viral and eukaryotic sources (https://www.ncbi.nlm.nih.gov/books/NBK53701/). The source identifiers take the form of the qualifiers /bio_material, /culture_collection and /specimen_voucher, determined by the type of source material. The general format of the qualifiers follows the DwC Standard (See **Glossary of Acronyms and Terms**). Subsequently, the DwC Triplets can be subject to problems when linking back to the specimen and the lack of global uniqueness

and persistence, obstacles that were indicated above.

The process of linking source identifiers to their respective institutions and collections is enabled by databases such as: 1) the Research Organization Registry (ROR) (https://ror.org/), 2) the Global Research Identifier Database (GRID, https://grid.ac/), 3) the Global Biodiversity Information Facility (GBIF) Registry of Scientific Collections which has been built-upon the Global Registry of Scientific Collections (GRSciColl), a depository of information on collections originally established by the Consortium of the Barcode of Life (CBOL) (https://www.gbif.org/en/grscicoll) and 4) the Index Herbariorum, a worldwide catalogue of herbaria where plant specimen vouchers are stored, curated by The New York Botanical Garden (http://sweetgum.nybg.org/science/ih/).

1.4. Purpose of the study

The need for digitisation initiatives has been continuously increasing at global and regional scales, resulting in the establishment of DiSSCo, that will allow for an exponential rate of digital biodiversity records proliferation in the future. Accordingly, it has become an ultimate requirement to direct attention towards linking these records to their corresponding physical specimens and associated data from additional sources (*Nelson et al. 2018*). Since the DwC Triplet is the identifier for specimen records used by ENA and GenBank, the present study is set out to gain a deeper understanding of the degree to which the DwC Triplets or a part thereof are sufficient to link sequence records to their physical specimens. Additionally, we assessed how three recently published studies reported Natural History Collection (NHC) specimens linked to sequences stored in the sequence databases of ENA and GenBank.

1.5. Research questions

- What is the percentage of nucleotide sequence records aggregated by the ENA database associated with source material identifiers? We need to establish the extent to which links between sequences and NSC specimens are present or not.
- How often do these identifiers have the DwC Triplet format (constructed according to the INSDC guidelines)?
- How often do we come across formats such as institutionCode:catalogNumber; an identifier lacking the “collectionCode” part, or solely catalogNumber?
- How many sequence records contain source material information that is not constructed according to the INSDC standards, e.g. they contain dash (“-“) as delimiter?
- How many institutionCodes and collectionCodes resolve to none, one or more IDs when queried on organisation and institution databases, e.g. ROR and GRID?
- Are source identifiers linked to sequence records studied in the three use cases consistently formatted according to the INSDC guidelines? How researchers from different scientific groups perceive the importance of linking these sequences to the source material from where they were extracted?

The quantification of these issues along with the investigation of the primary causes of any broken links will help understand what kind of curatorial improvements are needed to better connect sequences with specimens and encourage the diverse scientific communities to contribute to it.

2. Methods

2.1. Data collection

We focused on the collation of a dataset that consists of nucleotide sequence records from the ENA database. Each new sequence submission from a biological sample in the ENA database requires the pre-registration of a study and a sample in order to trace the sequence data back to the physical objects from where they were derived. The registration of the biological sample is an essential part of the submission process and it should be ensured that it contains the minimum amount of metadata. There are several checklists of expected metadata depending on the type of the sample (Sample Checklists: https://www.ebi.ac.uk/ena/browser/checklists) which are produced by different scientific communities with various requirements (See **Table S1**). In this way, ENA strives to ensure that the database complies with the FAIR (Findable, Accessible, Interoperable and Reusable) data principles. From a total of 39 checklists, 25 include source identifiers (pointers to the physical material or source material identifiers). For example, the ENA default sample checklist (https://www.ebi.ac.uk/ena/browser/view/ERC000011), includes ***optional*** fields for sequence data extracted from specimens stored in museums and reference collections such as **bio_material**, **culture_collection**, **specimen_voucher** (See **Glossary of Acronyms and Terms**), conforming to the INSDC guidelines (and the DwC Triplet format: institutionCode:collectionCode:catalogNumber). Here, we used ENA data collected between 1/1/2015 and 31/12/2019 containing 1,404,517 records.

2.2. Data analysis

First, we assessed the percentage of records that have a reference to a source identifier. From those that have an identifier we establish whether their institutionCodes and collectionsCodes resolve against existing databases such as ROR, GIRD and GBIF GRSciColl.

A workflow was created in R (version 4.0.0 codename "Arbor Day") (*R Core Team 2020*), by using RESTful Application Programming Interface (API) queries to download nucleotide sequence data. The workflow along with the files used for this research were deposited in GitHub and they are publicly available via https://github.com/stafyli/DiSSCo-project. ENA offers access to public and pre-publication data in various formats via its RESTful API (https://www.ebi.ac.uk/ena). In the present study, the ENA portal API (https://www.ebi.ac.uk/ena/portal/api) was queried via the /search endpoint that performs "a search against a single data group (result) of the data available in ENA" (ENA Portal API documentation, available here: https://www.ebi.ac.uk/ena/portal/api/doc?format=pdf). The package "httr" (*Wickham 2019*) was employed in order to handle the URL. The following parameters (See **Glossary of Acronyms and Terms**) were used to construct the query:

- *result*=sequence_release
- *query*=collection_date>=2015-01-01 AND collection_date<=2019-12-31 (as a set of search conditions, two different collection dates were used, joined by the logical operator AND)
- *fields*=accession, country, location, description, scientific_name, bio_material, culture_collection, specimen_voucher, sample_accession, study_accession (a list of fields (comma separated) were added to the query so to be returned in the result)
- *limit*=0 (the maximum number of records to retrieve was set to 0 meaning that the full result was fetched)
- *format*=json (JSON formatted output)

A dataset of all the nucleotide sequences stored in the ENA database that fulfill the criteria set by the query parameters was returned. It was converted into a data frame, a two-dimensional structure used to store data tables, via the package "jsonlite" (*Ooms 2014*). "jsonlite" is a JSON parser that allows for working with JSON formatted responses and enables the interaction with the ENA Portal API. The data frame was stored as a .csv file for future use.

The data frame was further analysed by using the "tidyverse" set of packages (*Wickham et al. 2019*). The data frame rows were filtered in order to locate those where specific conditions were true. We looked at the columns "bio_material", "culture_collection" and "specimen_voucher" and extracted the rows that contain data in at least one of them. Then, we extracted the observations that do not contain data in any of the columns mentioned before. The numbers and percentages of the nucleotide sequence accessions that meet the criteria were calculated and the two sub - data frames were stored as .csv files for future use. The data frame containing the sequence accessions lacking information regarding the physical material from where they were extracted, was also filtered to assess the number of accessions associated with a sample accession ID.

The same data was examined by using the R Base package and the function "grepl" (*R Core Team 2020*) that searches for matches to a specific pattern. The column "description" was searched to locate rows that contain terms such as "virus", "bacterium", "plasmid" etc., to estimate the number and percentage of accessions that have been extracted from uncultured bacterial/archaeal, viral and eukaryotic sources for which no specimen voucher information is available.

The next step in our analysis was to examine the accessions that contain information regarding their source material and the package "tidyverse" was employed for this task. The function "str_detect" (*Wickham et al. 2019*) was used to detect the presence of semicolon (";") in the source qualifier columns. The semicolon indicates annotation with more than one bio_material, culture_collection or specimen_voucher identifier. The main data frame was filtered based on the different types of identifier annotation and split into smaller data frames for easier access to their elements. Each data frame was analysed separately. The source qualifier columns, depending on the structure of the identifier (absence or presence of one or two colons (":")) were split into

"institution_code", "collection_code" and "material_id"/"culture_id"/"specimen_id" (=catalogNumber). The numbers and percentages of accessions with DwC Triplet, institutionCode:catalogNumber and catalogNumber formatted qualifiers were calculated. We also detected the presence of the pattern "text-text-text" with dash ("-") as delimiter in the identifiers that contain only a catalog number and calculated the number and percentage of these accessions as a proportion of the total number of accessions with source material information.

Two comma separated character vectors of the institutionCodes and collectionCodes derived from the studied data frames were created. The former was used to create a list of queries against the following databases: 1) ROR API (https://github.com/ror-community/ror-api) using the *query* parameter (resource URL: */organizations?query=*) and 2) GRID (JSON format, Release 2020-03-15, DOI: 10.6084/m9.figshare.12022722). Both vectors were queried on the GBIF GRSciColl API (https://www.gbif.org/developer/registry) utilising the resource URLS */grscicoll/institution?q=* and */grscicoll/collection?q=* respectively. Both "httr" and "jsonlite" packages were employed to read and transform the responses in a user-friendly format for further analysis. Tables and pie charts were constructed highlighting the number and percentages of results for each query.

2.3. Use cases

We explored three different use cases that represent published scientific papers on the phylogeography, phylogeny and historical biogeography of taxa, that use nucleotide sequences which were deposited in repositories such as ENA and GenBank. These sequences were either extracted by the authors or were accessed via published sources in ENA and NCBI GenBank and are associated with vouchered specimens stored in herbaria or NHCs. These case studies report detailed information about the collection and storage of the specimens used as well as the sequences studied. This allowed us to assess if ENA/GenBank sequence accessions provide the necessary information to trace these specimens. The workflow along with the files used or created during this research on use cases were deposited in GitHub and are publicly available at https://github.com/stafyli/DiSSCo-project.

2.3.1. Use case 1: Bat coronavirus phylogeography in the Western Indian Ocean *(Joffrin et al. 2020)*

In this study, samples from 36 bat species (vouchered specimens from previous studies and swab samples taken during the present study) originating from islands of the Western Indian Ocean including Mozambique, Madagascar, Mauritius, Mayotte, Reunion Islands and Seychelles, were used. The DNA sequence data of coronaviruses (CoVs) used are stored in NCBI GenBank database "nuccore" (https://www.ncbi.nlm.nih.gov/nucleotide/) under accession numbers: MN183146 - MN183273 including 128 records.

In order to determine the links that connect the sequence accessions to their physical objects, we utilised the R package "rentrez" (*Winter 2017*) that provides access to data stored in GenBank through the NCBI REST API Entrez Utilities (E-utilities)

(https://www.ncbi.nlm.nih.gov/books/NBK25501/). A comma separated character vector of the accessions was created on which the function "entrez_search()" was applied to query the "nuccore" database and retrieve the records that match the accession numbers. The IDs were input as a vector in the function "entrez_summary()" to fetch a list of summarised data in a simple format, of the records matching those IDs. The function "extract_from_esummary()" was used to extract the elements "uid", "caption", "title", "subtype", "subname", "organism"[9].

The column "subname" was split into 2 new columns, "isolate" and "host" to acquire access to the information they contain regarding the viral source material (See **Glossary of Acronyms and Terms**, more information is available here: https://www.ncbi.nlm.nih.gov/books/NBK53701/). The field "host" was further split into two columns: "host.species" and "host.identifier". We examined the format of these two source qualifiers and determined the degree to which they can be used to trace back the vouchered specimens. The number and percentages of accessions linked to a specimen voucher were determined.

2.3.2. Use case 2: Museomics for reconstructing historical floristic exchanges: Divergence of stone oaks across Wallacea (*Strijk et al. 2020*)

According to the authors, material from 39 *Lithocarpus* species was collected from herbarium collections and through sampling of living collections from botanical gardens and forest areas in Asian continental mainland and islands. Data of 8 *Quercus* species was additionally collated as outgroup from a previous study to root the phylogenetic tree of the *Lithocarpus* ingroup (the original catalogue of accessions and voucher information is provided here: https://doi.org/10.1371/journal.pone.0232936.s001). The *Lithocarpus* raw reads generated in this study were deposited in the ENA database under project number PRJEB34850, which is currently private and not accessible to the public with a hold-until date of 21/10/2020 (after communicating with the ENA helpdesk). The data was obtained directly from the authors. The original dataset was split into two .csv files, one for the *Lithocarpus* species (55 sequence accessions) and one for the *Quercus* species (21 sequence accessions) (See **Table S2**). The methodology indicated below, follows the one described in the section **2.2. Data analysis** of the present study (R packages used are "httr", "jsonlite", "tidyverse" and the R Base package) and was applied on the *Quercus* dataset.

The *Quercus* accessions were stored as a comma separated character vector and were queried against the ENA portal API. The fields "accession", "scientific_name", "bio_material" and "specimen_voucher" were fetched and the response's content (JSON format) was converted into a data frame. The column "bio_material" was blank so we split the "specimen_voucher" column into "institution_code" and "specimen_id" while collection codes were not provided. The numbers

[9] The extracted elements represent the following data found in a typical GenBank formatted record; uid: GenBank ID, caption: accession, title: definition, subtype and subname: source qualifiers.

and percentages of accessions associated with qualifiers that correspond to the DwC Triplet or part of it, were calculated.

A character vector was produced, storing the institution codes of the *Lithocarpus* and *Quercus* vouchers listed in the supplementary file provided by the paper's authors, as well as those retrieved from the ENA search. The vector was used to construct a list of queries on ROR API, GRID database and GBIF GRSciColl API. Additionally, the institution codes were searched against Index Herbariorum API (https://github.com/nybgvh/IH-API/wiki/GET-Institutions) using the resource URL /v1/institutions/{code}. Pie charts were constructed indicating the percentages of results for each query.

2.3.3. Use case 3: A complete time-calibrated multi-gene phylogeny of the European butterflies (*Wiemers et al. 2020*)

In this study, 496 extant butterfly species geographically distributed across Europe were examined. The dataset was collated both from published sources in NCBI GenBank and new sequences generated for the purpose of this study which were eventually submitted to GenBank. Most of the taxa studied are represented by multiple sequence accessions. The full dataset listing taxonomical, specimen voucher and accession information of the sequences used to build the phylogeny of European butterfly species is available here: https://zookeys.pensoft.net/article/50878/download/suppl/31/. The methodology used in this use case was adopted from the sections **2.2. Data analysis** and **2.3.1. Use case 1** (R packages used are "httr", "rentrez", "jsonlite", "tidyverse" and the R Base package).

The dataset was imported in R environment and the column "Voucher" was searched for the presence of colon (":") with the function "str_detect" resulting in 100% of the vouchers (496 in total) having the INSDC /specimen voucher format =specimen_id. We searched the NCBI GenBank database "nuccore" for the studied vouchers via the NCBI REST API E-utilities and the R package "rentrez" (as described in section **2.3.1. Use case 1**). The query returned vouchers associated with 0, 1 or more, unique GenBank IDs (2,123 in total) which were subsequently queried again in order to fetch the summaries of the records linked to them. Specific elements were extracted from each summary ("uid", "caption", "title", "subtype", "subname") and stored in a data frame. The latter was subset, keeping only the accessions extracted from the butterfly genera found in the original dataset (column "Genus") (functions "filter" and "grepl" were used). The new data frame was exported as a .csv file.

We extracted the specimen vouchers and isolates stored in the column "subname" and added them in a new column "source_identifier". Using the function "str_detect" in this column, we assessed the presence of more than one source qualifiers in each cell. The qualifiers seem to have the INSDC formats =institution-code:specimen_id and =specimen_id. Depending on their structure (absence or presence of one or two colons (":")), the qualifiers were split into "institution_code" and "specimen_id/idA/idB". The numbers and percentages of accessions associated with qualifiers that correspond to the DwC Triplet or part of it, were calculated.

After communicating with the corresponding author of the paper Martin Wiemers, we were provided with tissue IDs of 14 specimen vouchers used in the study, originating from his personal collection (See **Table S3**). These identifiers resemble the DwC Triplet format, but they have dash ("-") instead of a colon (":") as a delimiter. The institution and collection codes were queried on ROR API, GRID database and GBIF GRSciColl API.

3. Results

3.1. Analysis of the main dataset

A total of 1,404,517 accessions from the EMBL-ENA sequence_release dataset covering the time-period between 1/1/2015 and 31/12/2019 (quarterly release containing all sequences that are public at that time) were downloaded in JSON format. Approximately 33% of the sequence accessions (461,712 records) have source identifier information in at least one of the fields bio_material, culture_collection or specimen_voucher (**Fig. 3**).

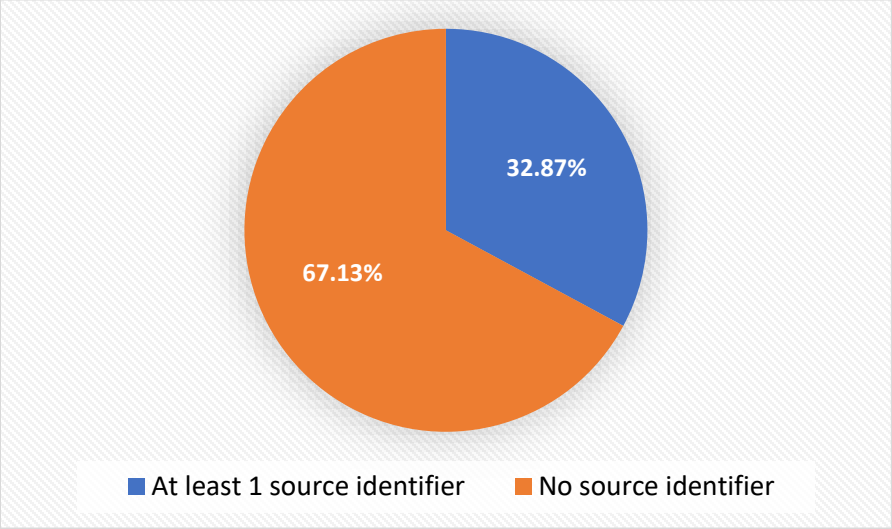

Figure 3.
Pie chart showing the percentages of 1.4 million sequence accessions stored in the ENA database that lack source identifier and those with at least one source identifier.

From the remaining ~67% (942,803 sequence records without source identifier information), only ~7% has an associated sample accession ID and 69.9% originate from microorganisms such as viruses and bacteria (**Fig. 4**). Two accessions were not assigned to either of the 2 groups and were excluded from the analysis.

Figure 4.
Pie charts showing the percentages of sequence accessions without a source identifier that have a sample accession ID (A) and those extracted from viral, bacterial etc., material (B).

The analysis indicated that 0.04% of the 461,712 sequence accessions were linked to more than one identifier of the same source qualifier. The results are shown in **Table 1**.

Table 1.
Annotation types of source qualifiers.

Annotation of source qualifier	Number of sequences	Percentage
>1 bio_material identifier	0	0
>1 culture_collection identifier	46	0.01
>1 specimen_voucher identifier	123	0.03
1 identifier in each source qualifier column	461,543	99.96
Total	**461,712**	**100**

The pie chart shown in **Figure 5** indicates the percentages of accessions linked to source identifiers with the DwC Triplet format (institutionCode:collectionCode:catalogNumber) or a part of it. Three point nine percent of the accessions with source information, had a source identifier in the form of a catalogNumber that contained the pattern “text-text-text” with dash (“-“) as a delimiter.

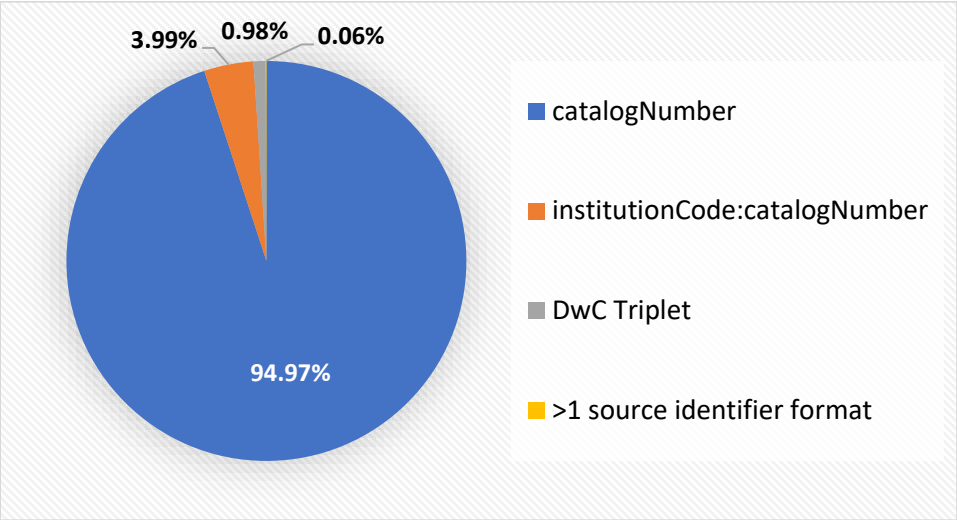

Figure 5.
Pie chart depicting the types of source identifier formats found in 461,712 sequence accessions studied (33%) with some kind of source identifier.

481 unique institution codes (**Table S4**) and 104 unique collection codes (**Table S5**) were extracted from the identifiers pointing to the physical material used to generate the nucleotide sequences studied. An overview of the results retrieved by querying the ROR API, GRID database and GBIF GRSciColl is shown in **Fig. 6**. The results per database are stored in https://github.com/stafyli/DiSSCo-project, in the respective .csv files. Nearly 62% of the institution codes queried on GRID were not resolved. GBIF was the data provider with the largest percentage of institution codes (31.6%) for which 1 search result was retrieved, among the 3 databases used. Almost 50% of the collection codes searched against the GBIF GRSciColl resulted in an error response from the API.

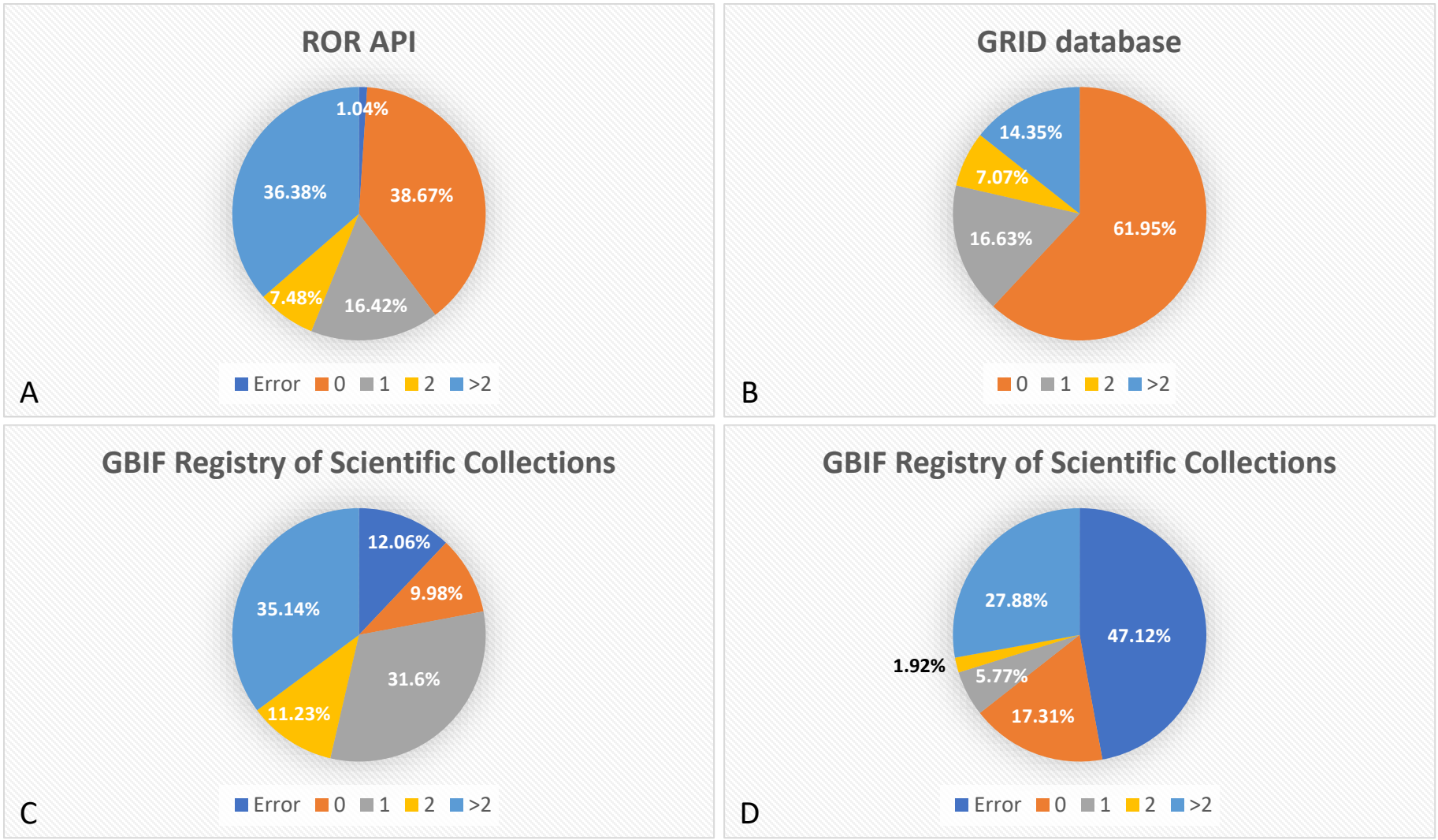

Figure 6.
Pie charts showing the percentages of institution (A-C) and collection codes (D) that had 0, 1, 2, >2 (or error response) query results.

3.2. Analysis of the use cases

3.2.1. Use case 1 (*Joffrin et al. 2020*)

Each of the 128 sequence accessions had one GenBank ID and all of them were linked to the same organism (bat coronavirus). GenBank IDs (uids) were not included in the paper but accessions could be easily traced in "nuccore" database.

The field "isolate" contained numerical information that could not be used to extract information regarding institution and/or reference collections. On the other hand, the field "host" contained information regarding the host species, sex and an identifier. From a total of 128 accessions, 114 (89.06%) had a host identifier in the format of an INSDC specimen voucher =specimen_id and corresponded to the viral sequences extracted from vouchered bat specimens. The remaining 10.94% (14 sequence accessions) did not have a host identifier, pinpointing to swab samples taken from bats.

3.2.2. Use case 2 (*Strijk et al. 2020*)

After a preliminary examination of the original dataset, we concluded that both *Lithocarpus* (55 in total) and *Quercus* (21 in total) sequence accessions were associated with specimen vouchers with the INSDC format =institution-code:specimen_id (institutionCode:catalogNumber).

After querying the ENA database, the records for the *Quercus* accessions were fetched (See **Table S6**). The chart below (**Fig. 7**) shows the percentages of accessions linked to source identifiers formatted as a part of the DwC Triplet. Sixty-seven percent of the *Quercus* accessions (14 in total) could be traced back to their physical specimens through institution codes.

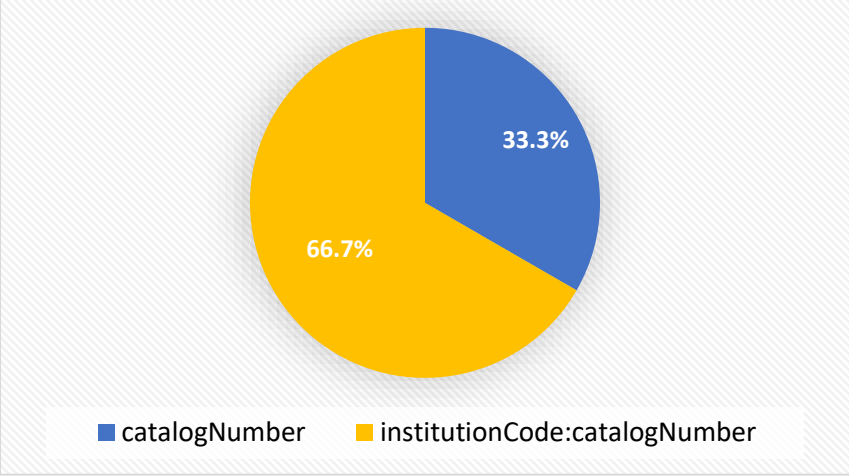

Figure 7.
Pie chart depicting the types of source identifier formats found in the 21 *Quercus* sequence accessions studied.

After examining and comparing **Table S2** and **Table S6**, differences between the specimen voucher information provided in the original dataset and the ENA database were evident. Such differences were found in both the institution codes ("KYO" in the paper and "FU<JPN>" in the ENA database) and the format of the specimen vouchers (7 accessions, with the prefix MF-, had the INSDC specimen voucher format =specimen_id or catalogNumber).

Eight unique institution codes ("BKF", "L", "KYO", "BGT", "KAG", "DLU", "FU", "FU<JPN>") were extracted from the *Lithocarpus* and *Quercus* specimen voucher identifiers included in the supplementary file and those obtained from the ENA database. An overview of the results retrieved by querying ROR API, GRID database, GBIF GRSciColl API and Index Herbariorum is shown in **Fig. S1** . Querying Index Herbariorum resolved 75% of the institution codes examined, while ROR and GBIF returned more than 2 IDs in 37.5% of the cases.

3.2.3. Use case 3 (*Wiemers et al. 2020*)

All specimen vouchers (496 in total) had the INSDC specimen voucher format =specimen_id so they could not be resolved via any institution and/or collection codes. The vouchers were queried in NCBI GenBank and no GenBank ID was retrieved for 18% of them. Eighty-two percent of the studied vouchers (406 unique vouchers) was associated with 1 or more GenBank IDs (a total of 2,123 IDs).

The summaries of 2,123 IDs were retrieved but 1,464 originated from the butterfly species studied (they were represented by 360 unique specimen vouchers and isolates from an expected number of 496). **Table 2** indicates the results of the search in the NCBI GenBank database "nuccore". The results are presented in more detail in the .csv file "butterflies.vouchers.analysis" (stored in https://github.com/stafyli/DiSSCo-project).

Table 2
Summary of the results retrieved by querying GenBank.

Query results	Number of vouchers	Percentage
Vouchers not linked to any GenBank IDs (sequences)	90	18.15
Vouchers linked to the species studied	360	72.58
Vouchers linked to other species	46	9.27
Total	**496**	**100**

The chart below (**Fig. 8**) shows the percentages of sequences linked to vouchers and isolates (as they were extracted from the GenBank summaries) formatted as a part of the DwC Triplet. The

largest number of sequences (1,348 IDs or 92.08%) had 1 or 2 identifiers with the format =specimen_id (catalogNumber).

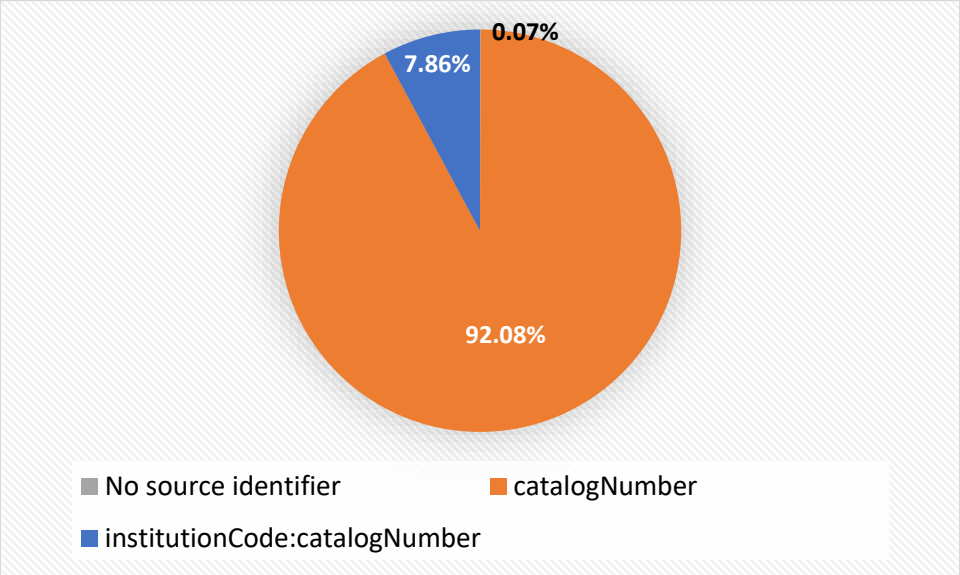

Figure 8.
Pie chart showing the types of source identifier formats associated with the 1,464 sequences linked to the vouchers studied.

The institution and collection codes queried were “ZMFK” and “MCZ”, and “TIS”, respectively. Via both ROR API and GRID, “ZMFK” resolved to “Zoological Research Museum Alexander Koenig” while this was not the case for “MCZ”. Both institution codes were resolved to their institutions through GBIF Registry of Scientific Collections API; 1 result: "Zoologisches Forschungsmuseum Alexander Koenig” and 2 results: "Museum of Comparative Zoology" "Museo Civico di Zoologia". As for the collection code “TIS”, the correct result had the 4th and 5th place among the results:

[1] "The NIST Biorepository"
[2] "Great Lakes Human Health Fish Fillet Tissue Study"
[3] "Great Lakes Human Health Fish Tissue Study"
[4] "ZFMK Tissue Bank (part of ZFMK Biobank)"
[5] "ZFMK Tissue Bank (part of ZFMK Biobank)"
[6] "Tissues" …

4. Discussion

The Darwin Core Standard has been the main provider of terms such as identifiers and definitions with the purpose of facilitating the sharing of biological diversity information (*Guralnick et al. 2014; Guralnick et al. 2015*). The DwC Triplet, a colon separated combination institutionCode:collectionCode:catalogNumber functions as a straightforward and user-friendly identifier that builds upon voucher specimens as they are documented by physical specimens or cultures stored in NSCs. The INSDC genetic data providers such as ENA and GenBank widely use the DwC Triplet to construct their source qualifiers. However, from the three qualifier fields only the third part of the DwC Triplet, namely the "catalogNumber" represented by "material_id", "culture_id" and "specimen_id" for the qualifiers bio_material, culture_collection and specimen_voucher respectively, is mandatory in all cases. A catalogue number by itself can summarise all required information that is needed to trace a voucher specimen but often it is difficult if not impossible to trace specimens or even the institutions where the material is stored. DiSSCo sets out to digitise the entire European NSCs and link all derived information from specimens to the digital specimens and construct the entire biodiversity knowledge graph around voucher specimens. DNA sequence data is a key element of the biodiversity knowledge graph and is a data source that has seen exponential growth over the past decades. Here we assessed which percentage of INSDC records has a source qualifier and the extent to which this allows tracing the specimen records from where the DNA material was obtained. The present study gives a first insight into what is needed to restore the links between specimen vouchers stored in NSCs and their derived DNA sequence data stored in INSDC databases, i.e. what is needed to repair what is broken?

We examined sequence data records deposited in the ENA database between 2015 and 2020 and determined how many records are linked to source material identifiers and whether the identifiers are constructed according to the recommended DwC Triplet syntax. Our findings indicated that about 1/3 of the sequences studied were associated with a source material (**Fig. 3**). Most of the remaining sequences (~70% - **Fig. 4**) originated from bacterial, viral etc. material and only 7% had a sample accession ID. We can assume that the instability of these source materials during the sequencing process can lead to the destruction of the sample and the subsequent inability of the researchers to provide source information during the sequence submission. The vast majority (~95%) of the sequence accessions with a source identifier only had a catalogNumber. Additionally, 3.9% of the source identifiers that contained only catalogNumbers were constructed with the use of dashes as delimiters, holding all parts of the DwC Triplet (it has been a common practice among many institutes to use the DwC Triplet as a recordID, some noteworthy examples were: ZFMK-DNA-FC19224938, PUZ-IMP-1129, IFR-SDO-S5, VU-ZOO-02 etc.). Merely 1% of the records was stored as canonical triplets in the ENA database (**Fig. 5**). These findings can be justified by the following: a) the data input is done by many different researchers with various perceptions on what should be included in a source identifier, b) mechanisms to ensure compliance to the INSDC guidelines (the use of colon (":") as a delimiter) are lacking and c) the filling in of source qualifier fields during submission of a sample (See **Table S1**) and the use of the DwC Triplet format are not mandatory (http://www.insdc.org/documents/feature-table#3.3.3).

The discrepancies among various communities regarding the importance of linking sequences to specimen vouchers using correct DwC Triplets became evident via the three use cases that were analyzed. No specimen voucher qualifiers were provided by researchers to GenBank for the bat coronavirus material and source identifiers were only provided in the field "host". The sequences extracted from swab samples did not contain any source information although the latter were preserved in medium both in the field and upon arrival at the laboratory *(Joffrin et al. 2020*). Regarding case study 2 on stone oaks, the *Lithocarpus* sequences studied are not yet accessible through the digital repository of ENA. Additionally, differences in vouchers of the *Quercus* accessions between the original dataset and the ENA database should not be overlooked, highlighting the lack of agreement and curation of identifiers. As for case study 3 on butterfly species, querying the vouchers in GenBank resulted in several GenBank IDs matching sequences that were extracted from non-butterfly species, while some vouchers could not be located at all. This indicated that, by solely using the vouchers, we were not able to reach the sequence records studied in the paper (one voucher resolved to multiple sequences). From the three case studies on recently published papers we can conclude that we were not able to trace the specimens in their entirety, based on data provided in the publications. This calls for further improvements in the procedure of submitting DNA sequence data derived from voucher specimens to INSDC data portals.

Databases such as ROR, GRID, GBIF GRSciColl and Index Herbariorum provide access to consistent lists of institution and collection acronyms (e.g. ZMFK = Zoological Research Museum Alexander Koenig). However, only ~16% of the institutionCodes from the ENA database sample used in our study were queried against ROR and GRID and resulted in a single identified institute. Nearly 62% of the institutionCodes queried on GRID resulted in zero IDs. Queries against GBIF GRSciColl resolved to 1 ID in 31.6% of the cases. The query success for the collectionCodes against GRSciColl was even lower. Some inconsistencies in the way that institution codes had been created impeded the query, e.g. iNAT and iNat, personal and Personal (referring to "personal collections that are privately owned and are associated with the virtual institutionCode "personal", according to *Schindel et al.* (*2016*)) etc. Index Herbariorum, being a worldwide index of herbaria, had the highest success in linking the institution codes for the *Lithocarpus* and *Quercus* accessions from the second use case, to a single institutional ID. More than half of the codes were not resolved through GRID while ROR and GBIF returned more than 2 IDs in 37.5% of the cases. The success rate of resolving institutional IDs will increase only if unique and unambiguous institution codes across all institutions are created and consistently used. Moreover, collection codes should be easily findable within an institution (*Schindel et al. 2016*). In all cases, the need for better organisation and maintenance of these databases with updated information on institution and collection codes emerges and should be emphasised before they are used on a large scale in sequence repositories such as ENA and GenBank.

The issues that can emerge from the use of DwC Triplets as identifiers in different data systems such as the lack of persistence and uniqueness and their susceptibility to breaks as the metadata changes, have been extensively discussed, primarily in the context of several studies such as those from *Guralnick et al.* (*2014*) and *Guralnick et al.* (*2015*). Although, sequence data repositories such

as ENA and GenBank are striving for the creation of a unified system of specifications for the construction of identifiers that conform to the FAIR principles, it is evident that it is not working in an efficient way. Our results from the ENA database sample and the 3 use cases showed that researchers from various communities, either face difficulties during the process of registration of the samples examined and the products of analysis and sequencing, or they have contradicting opinions on what type of information should be put in specimen voucher, culture collection and bio-material fields. The problems can be intensified by making the source qualifier fields optional or recommended and by the lack of supervision of the process followed in order to construct the DwC Triplet identifiers. Additionally, although the registration of a sample in the ENA database is mandatory before the submission of sequences, 93% of the sequences that were not linked to source material, did not have a sample accession ID (probably because the sample was destroyed after the analysis).

Limitations that could possibly affect the extent of applicability of our findings refer to time constraints regarding the completion of this MSc project and were identified as: 1) the use of sequences deposited in the sequence_release dataset of the ENA database and 2) the lack of further querying catalog numbers against databases such as the Life Science Identifier (LSID) resolver (http://www.lsid.info/resolver/) and ISNI (International Standard Name Identifier) (https://isni.org/page/search-database/) for cases where only these identifiers were available. However, the sequence accessions studied constitute a representative sample of the records stored in the ENA database and are a strong basis to make reliable inferences. In addition, the analysis of catalog numbers will be implemented in a future study by our research team in order to delve deeper into their use as concatenated identifiers.

Our results strongly point to the necessity of taking long-term measures regarding the curation of DwC Triplet identifiers. The significant value of a DwC Triplet system which brings together three terms: institutionCode, collectionCode and catalogNumber, and directly links to a specimen housed in a NSC, has long been identified but it also has been proven to be unstable and prone to changes even within the same institution. We recommend that the center of attention should be shifted towards complying to the guidelines provided by *Darwin Core and RDF/OWL Task Groups* (*2015*)) during the creation and validation of identifiers before they are submitted to any data system, either museum databases or sequence data repositories such as the INSDC members. In this way, local and global uniqueness and persistence along with the accordance to the FAIR principles will be safeguarded, elements that will ensure the successful tracking and management of specimens. New mechanisms that can supervise the registration of source information (sample identifiers, source qualifiers) in sequence repositories as well as a stricter approach to what should be considered optional or not, are matters that need re-evaluation. Specimens are not isolated entities but an irreplaceable part of a global interconnected data web, for instance, the biodiversity knowledge graph described by (*Page 2016*) and shown in **Fig. 1**. In addition, initiatives such as the DiSSCo RI will allow the creation of an ever-growing number of specimens, sequences, traits, occurrences and other digitised data. The demand for the resolution of linkages between them is currently, more urgent than ever in order to address the implications of challenging issues

namely the spread of pathogens e.g. coronaviruses, food security, invasive species, climate change-induced shifts in species distributions etc.

5. Acknowledgements

I would like to thank my supervisor, Dr. Dimitris Koureas for his invaluable insights throughout this project as well as my daily supervisor Dr. Niels Raes for his support and guidance in every step I took. Lastly, I stand thankful for the Onassis Foundation Scholarships Program for Greek students studying abroad which has been aiding me in achieving my academic aspirations from the very beginning to the end of this research master programme.

6. References

- Addink W, Koureas D, Rubio A (2019) DiSSCo as a New Regional Model for Scientific Collections in Europe. Biodiversity Information Science and Standards 3: e37502. https://doi.org/10.3897/biss.3.37502

- Darwin Core and RDF/OWL Task Groups (2015) Darwin Core RDF guide. Biodiversity Information Standards (TDWG). http://rs.tdwg.org/dwc/terms/guides/rdf/

- De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8 (2): 21. https://doi.org/10.3390/publications8020021

- GBIF (2009) Adoption of Persistent Identifiers for Biodiversity Informatics, a report from the GBIF LSID GUID Task Group (p. 23). Global Biodiversity Information Facility. http://www.gbif.org/orc/?doc_id=2956

- Guralnick R, Conlin T, Deck J, Stucky BJ, Cellinese N (2014) The Trouble with Triplets in Biodiversity Informatics: A Data-Driven Case against Current Identifier Practices. PLOS ONE 9 (12): e114069. https://doi.org/10.1371/journal.pone.0114069

- Guralnick RP, Cellinese N, Deck J, Pyle RL, Kunze J, Penev L, Walls R, Hagedorn G, Agosti D, Wieczorek J, Catapano T, Page R (2015) Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. ZooKeys 494: 133–154. https://doi.org/10.3897/zookeys.494.9352

- Hannah L, Roehrdanz PR, Marquet PA, Enquist BJ, Midgley G, Foden W, Lovett JC, Corlett RT, Corcoran D, Butchart SHM, Boyle B, Feng X, Maitner B, Fajardo J, McGill BJ, Merow C, Morueta-Holme N, Newman EA, Park DS, … Svenning J (2020) 30% land conservation and

climate action reduces tropical extinction risk by more than 50%. Ecography 43 (7): 943–953. https://doi.org/10.1111/ecog.05166

- Hardisty A, Ma K, Nelson G, Fortes J (2019) 'OpenDS' - A New Standard for Digital Specimens and Other Natural Science Digital Object Types. Biodiversity Information Science and Standards 3: e37033. https://doi.org/10.3897/biss.3.37033

- Hardisty A, Roberts D, The Biodiversity Informatics Community (2013) A decadal view of biodiversity informatics: Challenges and priorities. BMC Ecology 13 (16). https://doi.org/10.1186/1472-6785-13-16

- Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalga A, Paul D, Runnel V, Vermeersch X, van Walsum M, Willemse L (2020) Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1. Research Ideas and Outcomes 6: e54280. https://doi.org/10.3897/rio.6.e54280

- Hebert PDN, Gregory TR (2005) The Promise of DNA Barcoding for Taxonomy. Systematic Biology 54 (5): 852-859. https://doi.org/10.1080/10635150500354886

- Heerlien M, Van Leusen J, Schnörr S, De Jong-Kole S, Raes N, Van Hulsen K (2015) The Natural History Production Line: An Industrial Approach to the Digitization of Scientific Collections. Journal on Computing and Cultural Heritage 8 (1): 1-11. https://doi.org/10.1145/2644822

- Joffrin L, Goodman SM, Wilkinson DA, Ramasindrazana B, Lagadec E, Gomard Y, Le Minter G, Dos Santos A, Corrie Schoeman M, Sookhareea R, Tortosa P, Julienne S, Gudo ES, Mavingui P, Lebarbenchon C (2020) Bat coronavirus phylogeography in the Western Indian Ocean. Scientific Reports 10 (1): 6873. https://doi.org/10.1038/s41598-020-63799-7

- Karsch-Mizrachi I, Takagi T, Cochrane G, on behalf of the International Nucleotide Sequence Database Collaboration (2018) The international nucleotide sequence database collaboration. Nucleic Acids Research 46 (D1): D48-D51. https://doi.org/10.1093/nar/gkx1097

- Koureas D, Addink W (2017) Associating Occurrences with Genes, Phenotypes, and Environments through the Distributed System of Scientific Collections (DiSSCo). Proceedings of TDWG 1: e17010. https://doi.org/10.3897/tdwgproceedings.1.17010

- Koureas D, Addink W, Hardisty A (2018) Digital Object Cloud for linking natural science collections information; The case of DiSSCo. Biodiversity Information Science and Standards 2: e25474. https://doi.org/10.3897/biss.2.25474

- Lannom L, Koureas D, Hardisty AR (2020) FAIR Data and Services in Biodiversity Science and Geoscience. Data Intelligence 2 (1-2): 122-130. https://doi.org/10.1162/dint_a_00034

- Nelson G, Sweeney P, Gilbert E (2018) Use of globally unique identifiers (GUIDs) to link herbarium specimen records to physical specimens. Applications in Plant Sciences 6 (2): e1027. https://doi.org/10.1002/aps3.1027

- Ooms J (2014) The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. ArXiv: 1403.2805 [Cs, Stat]. http://arxiv.org/abs/1403.2805

- Page R (2016) Towards a biodiversity knowledge graph. Research Ideas and Outcomes 2: e8767. https://doi.org/10.3897/rio.2.e8767

- R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

- Schindel DE, Cook JA (2018) The next generation of natural history collections. PLOS Biology 16 (7): e2006125. https://doi.org/10.1371/journal.pbio.2006125

- Schindel D, Miller S, Trizna M, Graham E, Crane A (2016) The Global Registry of Biodiversity Repositories: A Call for Community Curation. Biodiversity Data Journal 4: e10293. https://doi.org/10.3897/BDJ.4.e10293

- Strijk JS, Binh HT, Ngoc NV, Pereira JT, Slik JWF, Sukri RS, Suyama Y, Tagane S, Wieringa JJ, Yahara T, Hinsinger DD (2020) Museomics for reconstructing historical floristic exchanges: Divergence of stone oaks across Wallacea. PLOS ONE 15 (5): e0232936. https://doi.org/10.1371/journal.pone.0232936

- TDWG [Biodiversity Information Standards] (2013) Globally Unique Identifiers (GUID) Wiki. http://wiki.tdwg.org/twiki/bin/view/GUID

- Wickham H (2019) httr: Tools for Working with URLs and HTTP. R package version 1.4.1. https://CRAN.R-project.org/package=httr

- Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, … Yutani H (2019) Welcome to the Tidyverse. Journal of Open Source Software 4 (43): 1686. https://doi.org/10.21105/joss.01686

- Wiemers M, Chazot N, Wheat C, Schweiger O, Wahlberg N (2020) A complete time-calibrated multi-gene phylogeny of the European butterflies. ZooKeys 938: 97-124. https://doi.org/10.3897/zookeys.938.50878

- Winter DJ (2017) rentrez: An R package for the NCBI eUtils API. The R Journal 9 (2): 520. https://doi.org/10.32614/RJ-2017-058

Web Pages (as they are found in the text)

- ICEDIG (2020) Innovation and Consolidation for Large Scale Digitisation of Natural Heritage. https://icedig.eu/
- DiSSCo (2020) Distributed System of Scientific Collections. https://www.dissco.eu/
- INSDC (2020) International Nucleotide Sequence Database Collaboration. www.insdc.org
- ELIXIR Europe (2020) European Life-Science Infrastructure. https://elixir-europe.org/
- BOLDSYSTEMS (2020) Barcode of Life Data System v4. http://www.boldsystems.org/
- DDBJ (2020) Bioinformation and DDBJ Center. https://www.ddbj.nig.ac.jp/insdc-e.html
- The GenBank Submissions Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US) (2011-) Providing Source Information in your Submission. https://www.ncbi.nlm.nih.gov/books/NBK53701/
- ROR (2020) Research Organization Registry. https://ror.org/
- GRID (2020) Global Research Identifier Database. https://grid.ac/
- GBIF GRSciColl (2020) Global Biodiversity Information Facility, Registry of Scientific Collections. https://www.gbif.org/en/grscicoll
- The New York Botanical Garden (2020) Index Herbariorum - The William & Lynda Steere Herbarium. http://sweetgum.nybg.org/science/ih/
- European Nucleotide Archive (ENA) (2020) Sample Checklists. https://www.ebi.ac.uk/ena/browser/checklists
- European Nucleotide Archive (ENA) (2020) Checklist: ERC000011. ENA default sample checklist. https://www.ebi.ac.uk/ena/browser/view/ERC000011
- European Nucleotide Archive (ENA) (2020) ENA Browser. https://www.ebi.ac.uk/ena
- European Nucleotide Archive (ENA) (2020) ENA Portal API - EMBL-EBI. https://www.ebi.ac.uk/ena/portal/api
- Research Organization Registry (ROR) (2020) ror-community/ror-api: ROR API - GitHub. https://github.com/ror-community/ror-api
- Global Biodiversity Information Facility (GBIF) (2020) GBIF registry API. https://www.gbif.org/developer/registry
- National Center for Biotechnology Information (NCBI) (2020) Home - Nucleotide - NCBI. https://www.ncbi.nlm.nih.gov/nucleotide/
- Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US) (2010-). https://www.ncbi.nlm.nih.gov/books/NBK25501/

- The New York Botanical Garden (2020) GET Institutions - nybgvh/IH-AOI Wiki - GitHub. https://github.com/nybgvh/IH-API/wiki/GET-Institutions
- International Nucleotide Sequence Database Collaboration (INSDC) (2019) The DDBJ/ENA/GenBank Feature Table Definition. http://www.insdc.org/documents/feature-table#3.3.3
- LSID (2020) Life Science Identifier resolver. http://www.lsid.info/resolver/
- ISNI (2020) International Standard Name Identifier. (https://isni.org/page/search-database/)

7. Supplementary material

Table S1
Sample checklists provided by the ENA database that guide the registration of each new sample as well as the minimum amount of information required depending on the type of sample (Available here: https://www.ebi.ac.uk/ena/browser/checklists). The field names that link to the physical material from where sequencing and analysis data are derived, and the level of requirement are indicated.

Accession	Name	Field Name	Requirement
ERC000012	GSC MIxS air	source material identifiers	**optional**
ERC000013	GSC MIxS host associated	source material identifiers	**optional**
ERC000014	GSC MIxS human associated	source material identifiers	**optional**
ERC000015	GSC MIxS human gut	source material identifiers	**optional**
ERC000016	GSC MIxS human oral	source material identifiers	**optional**
ERC000017	GSC MIxS human skin	source material identifiers	**optional**
ERC000018	GSC MIxS human vaginal	source material identifiers	**optional**
ERC000019	GSC MIxS microbial mat biolfilm	source material identifiers	**optional**
ERC000020	GSC MIxS plant associated	source material identifiers	**optional**
ERC000021	GSC MIxS sediment	source material identifiers	**optional**
ERC000022	GSC MIxS soil	source material identifiers	**optional**
ERC000023	GSC MIxS wastewater sludge	source material identifiers	**optional**
ERC000024	GSC MIxS water	source material identifiers	**optional**
ERC000025	GSC MIxS miscellaneous natural or artificial environment	source material identifiers	**optional**
ERC000027	ENA Micro B3	source material identifiers	**optional**
ERC000028	ENA prokaryotic pathogen minimal sample checklist	Pointer to physical material: bio_material, culture_collection, specimen_voucher	**optional**

Table S1
Sample checklists provided by the ENA database that guide the registration of each new sample as well as the minimum amount of information required depending on the type of sample (Available here: https://www.ebi.ac.uk/ena/browser/checklists). The field names that link to the physical material from where sequencing and analysis data are derived, and the level of requirement are indicated.

ERC000029	ENA Global Microbial Identifier reporting standard checklist GMI_MDM:1.1	Pointer to physical material: bio_material, culture_collection, specimen_voucher	**optional**
ERC000030	ENA Tara Oceans		
ERC000031	GSC MIxS built environment	source material identifiers	**optional**
ERC000032	ENA Influenza virus reporting standard checklist		
ERC000033	ENA virus pathogen reporting standard checklist		
ERC000034	ENA mutagenesis by carcinogen treatment checklist		
ERC000035	ENA Crop Plant sample enhanced annotation checklist		
ERC000036	ENA sewage checklist		
ERC000037	ENA Plant Sample Checklist	source material identifiers	**recommended**
ERC000038	ENA Shellfish Checklist		
ERC000039	ENA parasite sample checklist		
ERC000040	ENA UniEuk_EukBank Checklist		
ERC000041	ENA Global Microbial Identifier Proficiency Test (GMI PT) checklist		
ERC000043	ENA Marine Microalgae Checklist	Pointer to physical material: culture_collection	**optional**
ERC000044	COMPARE-ECDC-EFSA pilot human-associated reporting standard		
ERC000045	COMPARE-ECDC-EFSA pilot food-associated reporting standard		

Table S1
Sample checklists provided by the ENA database that guide the registration of each new sample as well as the minimum amount of information required depending on the type of sample (Available here: https://www.ebi.ac.uk/ena/browser/checklists). The field names that link to the physical material from where sequencing and analysis data are derived, and the level of requirement are indicated.

ERC000047	GSC MIMAGS	source material identifiers	**optional**
ERC000048	GSC MISAGS	source material identifiers	**optional**
ERC000049	GSC MIUVIGS	source material identifiers	**optional**
ERC000050	ENA binned metagenome	source material identifiers	**optional**
ERC000051	PDX Checklist		
ERC000052	HoloFood Checklist		
ERC000011	ENA default sample checklist	Pointer to physical material: bio_material, culture_collection, specimen_voucher	**optional**

Table S2
Part of the original dataset showing the *Quercus* accessions studied along with voucher information.

Species	BGT database # and source material	Voucher info	Institution code	Accession
***Quercus annulata* Sm.**	Silica	Tagane et al. V4730 (KYO), Vietnam	KYO	LC318796
***Quercus annulata* Sm.**	Silica	Tagane et al. V4730 (KYO), Vietnam	KYO	LC318516
***Quercus annulata* Sm.**	Silica	Tagane et al. V4730 (KYO), Vietnam	KYO	MF770291
***Quercus auricoma* A. Camus**	Silica	Ngoc et al. V3135 (KYO), Vietnam	KYO	LC318778
***Quercus auricoma* A. Camus**	Silica	Ngoc et al. V3135 (KYO), Vietnam	KYO	LC318498
***Quercus auricoma* A. Camus**	Silica	Ngoc et al. V3135 (KYO), Vietnam	KYO	MF770277
***Quercus austrocochinchinensis* Hickel & A. Camus**	Silica	Ngoc et al. V3129 (KYO), Vietnam	KYO	LC318777
***Quercus austrocochinchinensis* Hickel & A. Camus**	Silica	Ngoc et al. V3129 (KYO), Vietnam	KYO	LC318497

Table S2
Part of the original dataset showing the *Quercus* accessions studied along with voucher information.

***Quercus austrocochinchinensis* Hickel & A. Camus**	Silica	Ngoc et al. V3129 (KYO), Vietnam	KYO	MF770276
***Q. bambusifolia* Hance**	Silica	Ngoc et al. V3788 (KYO), Vietnam	KYO	LC318789
***Q. bambusifolia* Hance**	Silica	Ngoc et al. V3788 (KYO), Vietnam	KYO	LC318509
***Q. bambusifolia* Hance**	Silica	Ngoc et al. V3788 (KYO), Vietnam	KYO	MF770284
***Quercus macrocalyx* Hickel & A. Camus**	Silica	Tagane et al. V6457 (KYO), Vietnam	KYO	LC318800
***Quercus macrocalyx* Hickel & A. Camus**	Silica	Tagane et al. V6457 (KYO), Vietnam	KYO	LC318520
***Quercus macrocalyx* Hickel & A. Camus**	Silica	Tagane et al. V6457 (KYO), Vietnam	KYO	MF770294
***Quercus kerri* Craib**	Silica	Tagane et al. V6765 (KYO), Vietnam	KYO	LC318801
***Quercus kerri* Craib**	Silica	Tagane et al. V6765 (KYO), Vietnam	KYO	LC318521
***Quercus kerri* Craib**	Silica	Tagane et al. V6765 (KYO), Vietnam	KYO	MF770295
***Quercus* sp.**	Silica	Hoang et al. V5101 (DLU, FU), Vietnam	DLU, FU	
***Quercus poilanei* Hickel & A. Camus**	Silica	Yahara et al. V2986 (KYO), Vietnam	KYO	LC318774
***Quercus poilanei* Hickel & A. Camus**	Silica	Yahara et al. V2986 (KYO), Vietnam	KYO	LC318494
***Quercus poilanei* Hickel & A. Camus**	Silica	Yahara et al. V2986 (KYO), Vietnam	KYO	MF770273

Table S3
Tissue IDs for 14 specimen vouchers used in the study of *Wiemers et al.* (*2020)* provided by the corresponding author, Martin Wiemers.

Voucher	Tissue ID
MW99018	ZFMK-TIS-31171
MW06110	ZFMK-TIS-32408
JC08001	ZFMK-TIS-32393
MW01114	ZFMK-TIS-32115
MW00119	ZFMK-TIS-30226
MW09049	ZFMK-TIS-30681
MW99520	ZFMK-TIS-32013
MW06105	ZFMK-TIS-32403
MW06101	ZFMK-TIS-32399
JC00040	ZFMK-TIS-30491
BA09010	ZFMK-TIS-31288
JC00045	ZFMK-TIS-30496
MW01027	ZFMK-TIS-31643
MW02025	ZFMK-TIS-30549

Table S4
Institution codes queried in the present study and number of times they were encountered in the main dataset.

AA	AAC	ABRIICC	ABTC	ABTRI	ACAD	ACCC	AH	AK	ALCB	AM
10	2	4	36	1	5	3	9	4	2	50
AMH	AMNH	AMWH	ANFC	ANIC	ANWC	AQUA-MEB	ARAN	ARSEF	ATCC	AUMC <EGY>
20	17	15	51	6	1	3	1	7	103	2
B	BCCM/ LMG	BCCM/ MUCL	BCRC	BEI	BEOU	BILAS	BJFC	BJTC	BLAT	BM
133	2	4	73	3	54	3	45	14	1	2
BMNH	BNA<ESP>	BNHS	BNI	BP	BPBM	BPI	BR<BEL>	BRIP	BRNM	BSA
4	1	27	7	6	1	10	6	92	1	1
BUPL	BUQ	CAF	CAL	CANA	CAS	CAU	CAUP <CZH>	CBM	CBMAI	CBS
70	1	1	6	93	13	894	2	38	10	336
CCBAU	CCDB <BRA>	CCF	CCF<CZE>	CCGVCC	CCM	CCTCC	CCTU	CCUG	CECT	CFBH
5	3	111	4	1	2	16	20	13	3	22
CGMCC	CHR	CIB<CHN>	CIBIO	CICC	CIMAP	CIRAD	CMCC	CMFRI	CNC	CNCM
131	70	1	9	3	7	132	4	2	13	1
CNF	CNHM <CRO>	CNU <KOR-TJ>	CNU <KOR>	CONC	CONN	CORBIDI	CORD	CORT	CPC	CPCC

Table S4
Institution codes queried in the present study and number of times they were encountered in the main dataset.

6	1	194	5	42	68	13	23	2	193	1
CPH	CSCA	CSIRO	CTES	CU	CUH	DAOM	DAOMC	DB	DBCUCH	DBG
1	11	6	10	14	6	20	55	3	10	2
DBVPG	DEI	DMKU	DNA	DPC	DPIC	DSM	DU	DZUP	E	EAP
2	563	418	103	2	1	323	4	8	1	7
EBCC	EGE	ELM	EMCC	ESA	ESK	F	FACU	FAKU	FCME	FDA
1	245	23	1	22	2	1	20	31	4	156
FH	FLAS	FLOR	FMNH	FMR	FOF	FPM	FR	FRIM	FU	FU<JPM>
174	24	5	1	7	4	3	3	22	23	1
FU<JPN>	GCU	GDGM	GJO	GUBH	GUM <IRN>	GZAC	H	HAW	HBG	HHUF
176	1	22	2	4	1	10	5	49	2	52
HIB	HIRO	HKAS	HMAS	HMJAU	HNHM <HUN>	HUB	HUEFS	HUI	HUJ	HYO
6	33	76	35	23	9	6	36	4	1	61
IABHU <JPN>	IB	IBAR	IBH	IBIW	IBL<POL>	IBRC	IBSBF	IBSC	IBUNAM	ICEL
2	2	9	281	8	9	1	1	13	8	1
ICHUM	ICMP	IDB	IDEA	IFM	IFM<JPN>	IFO	IHB	InaCC	iNat	iNAT

Table S4
Institution codes queried in the present study and number of times they were encountered in the main dataset.

32	343	53	19	10	16	1	17	16	8	24
INHS	INIA<CHL>	INM<JPN>	INPA	INVEMAR	IPPAS	IRAN	ITCC<IND>	JCM	JEPS	JH
108	74	6	17	132	5	1	1	183	4	1
JPH	K	K(M)	KACC	KAS	KAUM	KBP	KCCM	KCOM	KCTC	KEIB
8	73	21	138	21	1	3	7	1	183	25
KFCC	KH	KIEL	KLU	KMNH	KNU	KNU16	KPABG	KPM	KR	KSU
1	4	2	18	3	5	4	1	30	31	3
KUHE	KUMA	KUMCC	KUN	KUN-HKAS	KUZ<JPN>	KWP	KYO	KYUM <JPN>	KZP	L
40	6	20	12	2	202	6	9	61	12	6
LACM	LCU	LDS	LE<RUS>	LG	LGMF	LMG	LUH	LY	LZ	M
71	2	1	7	12	16	63	16	2	2	170
MA	MAFF	MAK	Malicek	MCC	MCCC	MCVE	MCZ	MDC	MEDEL	MEXU
11	128	2	14	11	40	4	4	1	2	1
MFLU	MFLUCC	MH<IND>	MHNG	MIB	MICH	MIMB	MJG <DEU>	MLP	MNCN	MNHN
610	569	2	8	9	37	8	2	15	13	38
MNKNU	MNRJ	MO	MOR <USA-IL>	MPM <JPN>	MRAC	MRSN <USA>	MSB	MSC	MSNP	MTCC

Table S4
Institution codes queried in the present study and number of times they were encountered in the main dataset.

3	19	190	3	98	32	15	91	15	1	14
MTUF	MUB	MUBI	MUCC <JPN>	MUCL	MUGT	MUM	MUMH <JPN>	MUMNH Hir	MUSM	MVZ
1	4	5	5	27	1	35	20	18	3	1
MZB	MZH	MZUEL	MZUFBA	MZUSP	NBFGR	NBGB <BEL>	NBM	NBRC	NCCB	NCCP
6	19	12	8	84	8	4	2	53	1	18
NCCP <PAK>	NCHU	NCIMB	NCPF	NCPPB	NCSM	NCTC	NEHU	NFCCI	NGU	NHM
118	63	7	6	2	2	6	63	41	1	35
NHMC <GRC>	NHMUK	NHMW	NHRS	NIAES	NIES	NIWA	NK	NMBT	NMK	NMNZ
2	3	81	4	26	14	6	1	1	171	2
NMVF223094	NMW <GBR>	NN	NPCCAA	NRC<EGY>	NRM	NRRL	NSMT	NTM	NTNU	NTNU-VM
1	9	1	1	29	71	7	108	3	1	2
NTOU	NY	NYBG	NZFRI	NZFS	NZG	O	OMNZ	OSA	OSC	OSU
20	21	2	1	60	43	69	4	2	2	2
OSUC	PAMC	PC	PCC	PDD	PE	personal	Personal	PERTH	PGR	PM
9	5	1	6	178	4	1370	36	19	4	1
PMA	PMNH <PAK>	PPRI	PRC	PRM	PSU<THA>	PTBG	PU	PUCMF	PUL	QCAZ

Table S4
Institution codes queried in the present study and number of times they were encountered in the main dataset.

32	3	18	79	42	8	192	1	12	11	25
RB	RCC	RCPF	RGPG	RHK	RMRIMS	ROM	RU	S	SAMA	SARA
1	12	6	16	6	1	139	2	154	54	15
SASRI	SAV	SBC	SCUF	SD	SDAU	SDNCU	SDNHM	SDNU	SDSM	SERC <USA-MD>
2	68	1	1	3	6	70	1	3	86	19
SEV	SIO	SJRP	SMF <DEU>	SMNG	SMNH <ISR>	SMNH <SWE>	SMNHTAU <ISR>	SMNS	SPMN	STAR
102	111	10	4	17	9	14	1	1	1	46
STRI	STU	SY	SYSBM	SZMC	TAIF	TAMU	TAN	TASM	TAU<ISR>	TBRC
2	9	7	13	38	2	10	6	1	11	5
TENN	TFC	TFC<EST>	TIMM	TMI	TNM	TNS	TROM	TRPM	TRTE	TU<DEU>
505	41	19	4	12	92	157	1	26	1	1
TU<EST>	TUM	TUMH	TUR	TUS	TUZ	UAM	UAM <ESP>	UAMIZ	UARK	UAS
4	141	18	3	20	10	60	12	5	3	2
UAS-B	UASB	UBC	UBOCC	UC <USA-CA>	UCD	UCH	UCONN	UCR<CRI>	UCR <USA-CA>	UEFS
264	22	83	480	4	1	4	12	26	41	24
UF	UFMG <BRA>	UFRGS	UFRN	UGDA	UGM	UICC	UMMZ	UMSNH	UNAM	UNCW

Table S4
Institution codes queried in the present study and number of times they were encountered in the main dataset.

3	5	17	14	16	1	1	87	68	8	1
UND	UNN	UNSW	UNSW <AUS>	UO<CZE>	UOA/HCPF <GRC>	UPLB	UPS	URM	URM <BRA>	US
4	21	1	29	8	1	8	20	25	6	2350
USF	USGS	USM <MYS>	USMS	USNM	UTA	UW	UWM	UWM <POL>	UWO	VBCCA
5	9	160	24	2692	2	3	8	2	37	4
VIGG	VKM	Vondrak	VTCC	WAM	WIN	WNC	WTU	XJNU	ZFMK	ZLVG
6	20	7	3	1	15	2	1	1	32	6
ZMBN	ZMMU	ZMUB	ZMUC	ZMUD	ZRC	ZSI	ZUEC			
7	18	6	3	35	19	115	2			

Table S5
Collection codes queried in the present study and number of times they were encountered in the main dataset.

AAr	ADN	Ali Mansour	Amalendu Ghosh	ANN	ASC
24	4	3	1	18	8
Atsushi Kawakita	Atsushi Sugano	B2	BIC	Birds	CGR
4	83	1	111	292	3
Charoonluk	Chevonne Reynolds	CHN	chunlai liu	CNAN	CNNE
2	40	5	1	8	8
Cr	Damon Tighe	DNA	DNA Bank	Dominik Begerow culture collection	ENT
9	5	10	38	10	466
Ento	F	Feifei Li	Fish	FISH	Fungi
11	1	383	55	151	1
Fungos	Hedayat-Evright	Herbarium Berolinense	Herp	HERP	Herp Image
10	18	38	503	20	61
Herp Tissue	Hiroshi Masumoto	Hollinger	Huiqin Xing	I	I-L
1	2	2	33	1	2
I. Silas	IA	ICH	ICT	Institute of Agricultural Sciences-Spanish Natural Research Council	IZ

Table S5
Collection codes queried in the present study and number of times they were encountered in the main dataset.

2	1	18	14	5	903
Jiri Hulcr	Kevin Keegan	Kojiro Hara	Koukichi Maruyama	LEPID	Liliana Davalos
21	3	6	4	22	28
liuchunlai,liuliang	M. Hayder	Maduranga	MAMM	Mara DeMers	Marianne Nyegaard
3	1	84	132	229	3
Matt Kasson	Meijia Li	Ming-Hsun Chou	Mingliang and Jirimutu	Moll	Ole Lonnve
1	24	126	9	1	3
ORN	P	P7	Para	Ph4	PI
23	50	1	86	1	1
Pl	Prmanshayev Mamay	R.Nicoletti	Renli Zhang	Rosana Rocha	Rossanna
6	6	3	1	1	68
S Atherton	S9	Sakata	Shannon Adams	Shingo Akita	Sinthu
16	1	1	1	25	28
Susumu Yoshizawa	Sylvia Schaeffer	T. Aeman	T1	Tomoko Fukuda	UWRF
1	1	1	1	2	8
Valerio Monteiro Neto	Veli Vikberg	Watanuki	X.Yao	Xiaodong Jiang	xinmin li

Table S5
Collection codes queried in the present study and number of times they were encountered in the main dataset.

10	3	1	3	37	6
Ya Zhang	Yoshio Tani	Yume Imada	ZC	Zhang YanLong	Zhaohui Luo
1	3	5	4	3	3
ZOOL	ZPL				
51	9				

Table S6
Records of the studied *Quercus* accessions retrieved after querying the ENA database.

Accession	Scientific name	Specimen voucher	Institution code	Catalog number
LC318796	*Quercus annulata*	FU<JPN>:V4730	FU<JPN>	V4730
LC318516	*Quercus annulata*	FU<JPN>:V4730	FU<JPN>	V4730
MF770291	*Quercus annulata*	V4730		V4730
LC318778	*Quercus auricoma*	FU<JPN>:V3135	FU<JPN>	V3135
LC318498	*Quercus auricoma*	FU<JPN>:V3135	FU<JPN>	V3135
MF770277	*Quercus auricoma*	V3135		V3135
LC318777	*Quercus austrocochinchinensis*	FU<JPN>:V3129	FU<JPN>	V3129
LC318497	*Quercus austrocochinchinensis*	FU<JPN>:V3129	FU<JPN>	V3129
MF770276	*Quercus austrocochinchinensis*	V3129		V3129
LC318789	*Quercus neglecta*	FU<JPN>:V3788	FU<JPN>	V3788
LC318509	*Quercus neglecta*	FU<JPN>:V3788	FU<JPN>	V3788
MF770284	*Quercus neglecta*	V3788		V3788
LC318800	*Quercus macrocalyx*	FU<JPN>:V6457	FU<JPN>	V6457
LC318520	*Quercus macrocalyx*	FU<JPN>:V6457	FU<JPN>	V6457
MF770294	*Quercus macrocalyx*	V6457		V6457
LC318801	*Quercus kerrii*	FU<JPN>:V6765	FU<JPN>	V6765
LC318521	*Quercus kerrii*	FU<JPN>:V6765	FU<JPN>	V6765
MF770295	*Quercus kerrii*	V6765		V6765
	Quercus sp.			
LC318774	*Quercus poilanei*	FU<JPN>:V2986	FU<JPN>	V2986
LC318494	*Quercus poilanei*	FU<JPN>:V2986	FU<JPN>	V2986
MF770273	*Quercus poilanei*	V2986		V2986

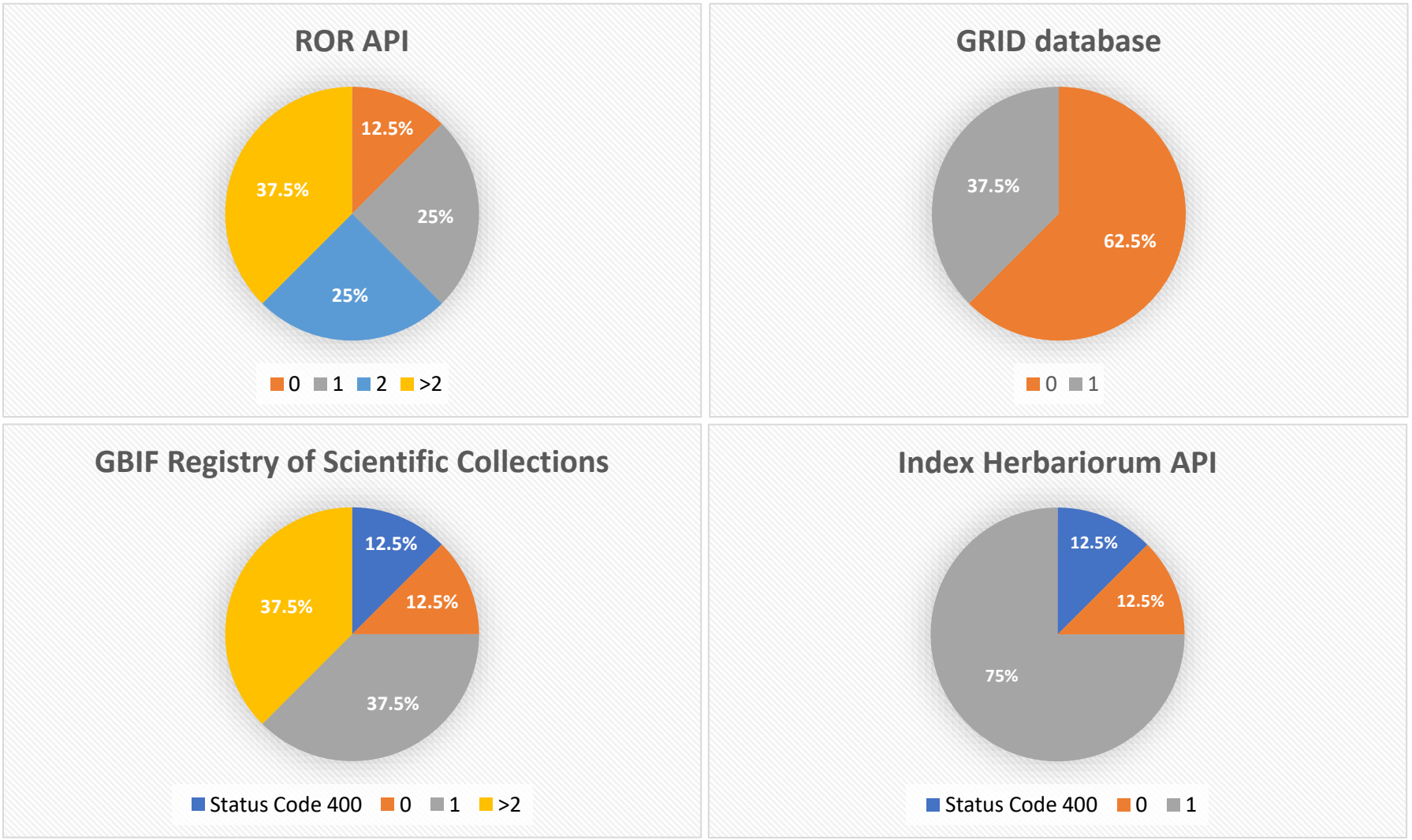

Figure S1.
Pie charts showing the percentages of institution codes extracted from the study of Strijk et al. (2020) having 0, 1, 2, >2 results (or Status Code 400 response) after querying ROR API, GRID database, GBIF Registry of Scientific Collections and Index Herbariorum API.

Glossary of Acronyms and Terms

Acronyms

API: Application Programming Interface
BOLD: Barcode of Life Data System
CBOL: Consortium of the Barcode of Life
CoL+: Catalogue of Life Plus
CoV: Coronavirus
DDBJ: DNA DataBank of Japan
DiSSCo: Distributed System of Scientific Collections
DiSSCo CSO: DiSSCo Coordination and Support Office
DOA: Digital Object Architecture
DwC: Darwin Core
ELIXIR: European Life-Science Infrastructure
EMBL-EBI: European Molecular Biology Laboratory-European Bioinformatics Institute
ENA: European Nucleotide Archive
EOL: Encyclopedia of Life
FAIR: Findable, Accessible, Interoperable and Reusable
FDO: FAIR Digital Object
GBIF: Global Biodiversity Information Facility
GRID: Global Research Identifier Database
GRSciColl: Global Registry of Scientific Collections
GUID: Globally Unique Identifier
iBOL: International Barcode of Life
INSDC: International Nucleotide Sequence Database Collaboration
IRL: Implementation Readiness Level
JSON: JavaScript Object Notation
NCBI: National Center for Biotechnology Information
NSC: Natural Science Collection
PID: Persistent Identifier
RDF/OWL: Resource Description Framework/Web Ontology Language
REST: Representational State Transfer
*R*I: Research Infrastructure
ROR: Research Organization Registry
TDWG: Taxonomic Databases Working Group (also known as Biodiversity Information Standards)
URI: Uniform Resource Identifier
URL: Uniform Resource Locator

Terms

ENA Portal API documentation, https://www.ebi.ac.uk/ena/portal/api/doc?format=pdf)

collection_date: Date that the specimen was collected

sequence_release: Quarterly release containing all sequences that are public at that time

accession: Accession number

country: Locality of sample isolation: country names, oceans or seas, followed by regions and localities

location: Geographic location of isolation of the sample. Latitude and longitude are given in decimal degrees. When using latitude and longitude in the geospatial search functions, positive and negative values should be given to represent direction. When returned in the results report, N/S and E/W are displayed with latitude and longitude

description: Brief sequence description

scientific_name: Scientific name of the organism from which the sample was derived

bio_material: Identifier for biological material including institute and collection code

culture_collection: Identifier for the sample culture including institute and collection code

specimen_voucher: Identifier for the sample culture including institute and collection code

sample_accession: Sample accession number

study_accession: Study accession name

source material identifiers (mixs:source mat id): A unique identifier assigned to a material sample (as defined by https://dwc.tdwg.org/terms/#dwc:materialSampleID, and as opposed to a particular digital record of a material sample) used for extracting nucleic acids, and subsequent sequencing. The identifier can refer either to the original material collected or to any derived subsamples. The INSDC qualifiers /specimen_voucher, /bio_material, or /culture_collection may or may not share the same value as the source_mat_id field. For instance, the /specimen_voucher qualifier and source_mat_id may both contain 'UAM:Herps:14' , referring to both the specimen voucher and sampled tissue with the same identifier. However, the /culture_collection qualifier may refer to a value from an initial culture (e.g. ATCC:11775) while source_mat_id would refer to an identifier from some derived culture from which the nucleic acids were extracted (e.g. xatc123 or ark:/2154/R2).

/bio_material= (http://www.insdc.org/documents/feature-table#3.3.3):

Definition: identifier for the biological material from which the nucleic acid sequenced was obtained, with optional institution code and collection code for the place where it is currently stored.

Value format: "[<institution-code>:[<collection-code>:]]<material_id>"

Example: /bio_material="CGC:CB3912" <- Caenorhabditis stock centre

Comment: the /bio_material qualifier should be used to annotate the identifiers of material in biological collections that are not appropriate to annotate as either /specimen_voucher or /culture_collection; these include zoos and aquaria, stock centres, seed banks, germplasm repositories and DNA banks; material_id is mandatory, institution_code and collection_code are optional; institution code is mandatory where collection code is present; institution code and collection code are taken from a controlled vocabulary maintained by the INSDC.

/culture_collection= (http://www.insdc.org/documents/feature-table#3.3.3):
Definition: institution code and identifier for the culture from which the nucleic acid sequenced was obtained, with optional collection code.
Value format: "<institution-code>:[<collection-code>:]<culture_id>"
Example: /culture_collection="ATCC:26370"
Comment: the /culture_collection qualifier should be used to annotate live microbial and viral cultures, and cell lines that have been deposited in curated culture collections; microbial cultures in personal or laboratory collections should be annotated in strain qualifiers; annotation with a culture_collection qualifier implies that the sequence was obtained from a sample retrieved (by the submitter or a collaborator) from the indicated culture collection, or that the sequence was obtained from a sample that was deposited (by the submitter or a collaborator) in the indicated culture collection; annotation with more than one culture_collection qualifier indicates that the sequence was obtained from a sample that was deposited (by the submitter or a collaborator) in more than one culture collection. culture_id and institution_code are mandatory, collection_code is optional; institution code and collection code are taken from a controlled vocabulary maintained by the INSDC.
http://www.insdc.org/controlled-vocabulary-culturecollection-qualifier

/host= (http://www.insdc.org/documents/feature-table#3.3.3):
Definition: natural (as opposed to laboratory) host to the organism from which sequenced molecule was obtained.
Value format: "text"
Example: /host="Homo sapiens"
/host="Homo sapiens 12 year old girl"
/host="Rhizobium NGR234"

/isolate= (http://www.insdc.org/documents/feature-table#3.3.3):
Definition: individual isolate from which the sequence was obtained
Value format: "text"
Example: /isolate="Patient #152"
/isolate="DGGE band PSBAC-13"

/specimen_voucher= (http://www.insdc.org/documents/feature-table#3.3.3):
Definition: identifier for the specimen from which the nucleic acid sequenced was obtained
Value format: "[<institution-code>:[<collection-code>:]]<specimen_id>"
Examples: /specimen_voucher="UAM:Mamm:52179"
/specimen_voucher="AMCC:101706"
/specimen_voucher="USNM:field series 8798"
/specimen_voucher="personal:Dan Janzen:99-SRNP-2003"
/specimen_voucher="99-SRNP-2003"
Comment: the /specimen_voucher qualifier is intended to annotate a reference to the physical specimen that remains after the sequence has been obtained; if the specimen was destroyed in

the process of sequencing, electronic images (e-vouchers) are an adequate substitute for a physical voucher specimen; ideally the specimens will be deposited in a curated museum, herbarium, or frozen tissue collection, but often they will remain in a personal or laboratory collection for some time before they are deposited in a curated collection; there are three forms of specimen_voucher qualifiers; if the text of the qualifier includes one or more colons it is a 'structured voucher'; structured vouchers include institution-codes (and optional collection-codes) taken from a controlled vocabulary maintained by the INSDC that denotes the museum or herbarium collection where the specimen resides.
http://www.insdc.org/controlled-vocabulary-specimenvoucher-qualifier