MISSIONE 4
ISTRUZIONE
RICERCA

ITINERIS

Deliverable D6.4 - Activity 6.5, CNR-IBBR-BA

Shared protocols and best practices for acquisition, organization, standardization and long-term maintenance of (meta)data related to NSC; set-up of the centralized IT platform including a bioinformatic toolbox with web-GIS and modeling facilities

Authors: Gabriele Bucci, Massimo Ianigro, Eleonora Fornaro, Marina Tumolo, Domenico De Paola, Giovanni Giuseppe Vendramin

August 31, 2025

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

Deliverable number:	D6.4
Work package:	WP6 – Coordination and management
Intermediate Objective:	IO6.5
Deliverable type:	X Document, report
	☐ Websites, patent filings, videos, etc.
	☐ Other: please specify .....
Dissemination level:	X Public
	☐ Restricted
Estimated delivery (bimester):	B12 (2024-10-31)
Actual delivery date:	B17 (2025-08-31)
Author(s) (Partner-OU):	Gabriele Bucci, Massimo Ianigro, Eleonora Fornaro, Marina Tumolo, Domenico De Paola, Giovanni Giuseppe Vendramin (CNR-IBBR-BA)
Reviewed by:	ITINERIS Executive Board
Note:	

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System - CUP B53C22002150006 (D.D. n. 130/2022)

Funded by EU - Next Generation EU

Mission 4 “Education and Research” - Component 2: “From research to business” -

Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infra-structures”

Table of Contents

1. INTRODUCTION

1.1 Purpose of the document

This document describes the main outcomes of Activity 6.5 of the ITINERIS project and the procedures adopted to establish one of the Italian nodes of the European research infrastructure DiSSCo-RI ("Distributed System of Scientific Collections"). This node will be interconnected with the ITINERIS Hub, the WP6 Coordination Hub, the WP8 Virtual Research Environment, and several other Research Infrastructures (RIs) participating in the project.

The deliverable D6.4 was prepared by the operating unit (OU) CNR-IBBR-BA as part of Activity 6.5 ("Mining and mapping the functional biodiversity of in vivo and ex-situ research collections"), in collaboration with Activities 6.4 and 6.6, which are also involved in deploying and enhancing the mentioned RI for Italy.

DiSSCo-RI (https://www.dissco.eu) is a European initiative aimed at promoting the complete digitization and digital unification of all Natural Science Collections (NSC) at the European level. It provides for the development of shared activities and standard practices aimed at the retrieval, accessibility, interoperability, and reuse of the information associated with the individual specimens stored in collections. DiSSCo-RI is currently in the implementation phase and will be fully operational in 2026.

The Italian NSCs, including natural history museums, botanical gardens, and research collections, hold relevant information for several thousand species (plants, animals, fungi, microorganisms, etc.), both terrestrial, marine, and from internal waters, spread across the Italian peninsula. For many of them, the site and date of specimens' collection are available, and their main ecological performances/functions/guilds are known. Based on this data, the network of NSCs developed by Activities 6.4, 6.5, and 6.6 of ITINERIS will provide relevant information on Italian biodiversity, which will contribute to assess Essential Biodiversity Variables (EBV) across different regions and time periods for many species or groups of species. Moreover, the availability of fine-scale data from climatic series through downscaling tools will enable modeling the distribution of various species under future climate scenarios and mapping Italian regions at risk of biodiversity loss due to global change. Finally, the availability of a large amount of taxonomic, genetic, geographical, ecological, and physiological information on plant, animal, and microbial genetic resources will pave the way to mining and mapping functional biodiversity in both *in vivo* and *ex-situ* collections, with the ultimate goal of a better understanding of the specific mechanisms of adaptation to climate change.

In the above context, Activity 6.5 has focused on creating a data repository that consolidates all information related to the natural science collections of the Department of BioAgriFood Sciences at the National Research Council of Italy (CNR-DiSBA), serving as the first step toward establishing the Italian node of DiSSCo-RI. This task has demanded skills and expertise from various scientific and technological fields, including taxonomy, data management, semantics, information technology, and more. All the tools, software, protocols, workflows, procedures, and experiences for completing this challenging task are summarized in this document. Furthermore, Activity 6.5 aimed to develop the digital facilities and infrastructures to support the future integration of collections and biodiversity datasets from prospective data providers (natural history museums, botanical gardens, academic institutes, etc.) into a common "data lake", regardless of their current involvement in the ITINERIS project.

1.2 Definitions, acronyms, and abbreviations

- CED: *Centro Elaborazione Dati* (Datacenter)
- CNR: *Consiglio Nazionale delle Ricerche* (National Research Council)
- CNR-IBBA: Institute of Agricultural Biology and Biotechnology
- CNR-IBBR-BA: Institute of Biosciences and Bioresources in Bari
- CNR-IBE: Institute of BioEconomy
- CNR-IPSP: Institute for Sustainable Protection of Plants
- CNR-IRSA-VB: Water Research Institute in Verbania
- CNR-ISA: Institute of Food Science
- CNR-ISAFOM: Institute for Agricultural and Forest Systems in the Mediterranean
- CNR-ISB: Institute for Biological Systems
- CNR-ISMAR-VE: Institute of Marine Sciences in Venice
- CNR-ISPA: Institute of Sciences of Food Production
- CNR-ISPAAM: Institute for Animal Production System in Mediterranean Environment
- CNR-DiSBA: Department of BioAgriFood Sciences of the National Research Council of Italy
- CNR-DSSTTA: Department of Earth System Sciences and Environmental Technologies of the National Research Council of Italy
- CoL: Catalog of Life
- DiSSCo: Distributed System of Scientific Collections
- DMP: Data Management Plan
- DOI: Digital Object Identifier
- EBV: Essential Biodiversity Variables
- EMBL: European Molecular Biology Laboratory
- EML/XML: Ecological Markup Language
- ENA: European Nucleotide Archive
- EURISCO: European Search Catalogue for Plant Genetic Resources
- EVA: European Variation Archive
- FAIR: Findable, Accessible, Interoperable, and Reusable
- FDO: FAIR Digital Object
- FIP: FAIR Implementation Profile
- GBIF: Global Biodiversity Information Facility
- GCM: General Circulation Models
- GENESYS PGR: Aggregated database for Plant Genetic Resources for Food and Agriculture
- GRSciColl: Global Registry of Scientific Collections
- ICT: Information and Communication Technology
- IPT: Integrated Publishing Tool
- IT: Information Technology
- ITINERIS: Italian Integrated Environmental Research Infrastructure System
- NCBI: National Center for Biotechnology Information, Bethesda, MD, USA
- NSC: Natural Science Collections
- ORCID: Open Researcher and Contributor ID
- OU: Operating Unit
- PNIR 2021-2027: *Piano Nazionale delle Infrastrutture di Ricerca* (National Plan for Research Infrastructures)
- PUID: Permanent Unique Identifiers
- RDF: Resource Description Framework
- RI: Research Infrastructure
- ROR: Research Organization Registry
- SSR: Simple Sequence Repeats (DNA microsatellite)
- TDWG; Taxonomic Database Working Group
- UNIFI-SMA: *Sistema Museale di Ateneo* – University of Firenze
- UUID: Universally Unique Identifiers
- VM: Virtual Machine.

2. OVERVIEW

2.1 Objectives

DiSSCo-RI is a European Research Infrastructure included in the ESFRI roadmap since 2018, aimed at promoting the complete digitization and digital unification of Natural Science Collections (NSC) across Europe. Through the development of shared activities and practices, DiSSCo-RI will make NSC data and metadata "FAIR" (Findable, Accessible, Interoperable, and Reusable), with the ultimate goal of promoting knowledge in biodiversity, geodiversity, and scientific research innovation.

DiSSCo-RI will include all types of natural science collections to create a pan-European system capable of providing quality scientific data by integrating them with other information sources about species, genomes, phenotypes, geography, geology, and ecology, thus providing innovative tools for environmental research.

DiSSCo-RI closely collaborates with GBIF (Global Biodiversity Information Facility[1]), which is one of the largest biodiversity data repositories in the world. DiSSCo-RI will leverage this repository to develop its services for end users and support the digital unification of NSC in Europe. The ITINERIS operating unit (OU) CNR-IBBR-BA was officially recognized as a GBIF publisher and received endorsement from DiSSCo on February 4, 2023[2].

One of the main objectives of Activity 6.5 was to create a "digital twin" for all organisms stored in the CNR-DiSBA research collections and to make it accessible via the GBIF repository in a machine-readable, FAIR-compliant format, to be included in the DiSSCo-RI European catalog. To achieve this, the GBIF publisher has recently been renamed "*Consiglio Nazionale delle Ricerche, Istituti del Dipartimento di Scienze BioAgroAlimentari*"[3] to host the datasets from the CNR-DiSBA collections.

The OU CNR-IBBR-BA was also involved in deploying the facilities and infrastructures necessary to establish one of the Italian nodes of the European Research Infrastructure, in collaboration with other partners in the ITINERIS project, including CNR-ISMAR-VE and CNR-IRSA-VB, which are responsible for Activity 6.4, and SMA-UNIFI, which is accountable for Activity 6.6.

2.2 Preparatory projects

Activity 6.5 of the ITINERIS project partly leverages the NSC meta-information collected during the BioMemory project[4] by increasing the number of datasets included, enriching their meta-information, extending their usability and compliance with FAIR principles, as well as supporting the functionalities that are needed by the use cases of the ITINERIS central hub.

BioMemory is a one-year project promoted by the CNR-DiSBA aimed at inventorying the research collections maintained at its 9 Institutes (i.e., CNR-IBBA, CNR-IBBR, CNR-IBE, CNR-IPSP, CNR-ISA, CNR-ISAFOM, CNR-ISB, CNR-ISPA, CNR-ISPAAM) in view of the CNR participation in DiSSCo-RI. The project has been included in the PNIR 2021-2027 ("*Piano Nazionale Infrastrutture di Ricerca*") with medium priority and funded by CNR only for the first year (Oct 2021-Sep 2022). The project aimed to support the creation of a network of NSCs, institutes, and people involved in biobanking and analysis of the genetic resources of different kinds of organisms

[1] Global Biodiversity Information Facility (GBIF – https://www.gbif.org)
[2] GBIF Associate "Distributed System of Scientific Collections" (https://www.gbif.org/participant/418).
[3] CNR-DiSBA publisher page in GBIF (https://www.gbif.org/publisher/6563e0ba-fab7-431c-b897-b6bf364f4f1e)
[4] BioMemory project (https://biomemory.cnr.it/).

(plants, animals, fungi, bacteria, viruses, etc.).

BioMemory can be regarded as a "preparatory project" for ITINERIS Activity 6.5, since meta-information was preliminarily inventoried for 56 NSCs by 153 researchers and technicians working in the 27 local divisions of the nine institutes mentioned above. In many cases, these collections have been established over the years through participation in national and international research projects. However, most NSC information was collected using different protocols and formats, due to the diversity of organisms in the collections (from plant viruses to forest trees and soil nematodes), the various objectives of collections and projects (from studying biopathogens to improving crop productivity), and the different periods in which data has been compiled (from paper records in the 1960s to electronic spreadsheets in recent years), making their harmonization and interoperability a challenging task.

Within Activity 6.5 of the ITINERIS project, the available meta-information has been collected, enriched, reorganized, standardized, harmonized, ensured to comply with FAIR principles, and published online for end users. For the reasons mentioned above, the datasets originating from the BioMemory initiative have also been annotated with acknowledgments to this project. Moreover, the BioMemory platform (https://biomemory.cnr.it), which was largely incomplete at the end of 2022, has been further developed, enhanced, and integrated during the ITINERIS project to host all the functions needed for machine-to-machine data exchange (API, endpoints, etc.) and support the integration with other services.

2.3 Summary of outcomes

The primary outcomes achieved by the OU CNR-IBBR-BA within Activity 6.5 of the ITINERIS project can be outlined as follows:

- A data management plan (DMP) has been prepared in collaboration with the ITINERIS OUs UNIFI-SMA (Activity 6.4) and CNR-ISMAR (Activity 6.6), as detailed in **Chapter 3**.
- Best practices, protocols, and workflows have been developed, along with specific recommendations for using the facilities available at the CNR-IBBR-BA data center (see **Chapter 4** and **Appendix 1**).
- Some 23 CNR-DiSBA collections have been published on the Global Registry of Scientific Collections (GRSciColl[5]), and 52 related datasets have been fully digitized (totaling 55,073 specimens/accessions), of which 35 occurrence datasets are already registered and indexed in the GBIF platform (see **Chapter 5**).
- A shared data portal (https://dissco-itineris.it) encompassing the metadata catalog for all collections and the datasets generated by Activities 6.4, 6.5, and 6.6 has been successfully implemented (see **Chapter 6**).
- Several services tailored to the needs of data providers and end-users have been developed, including:
  - The IPT platform (Integrated Publishing Tool – https://ipt.ibbr.cnr.it/) for publishing biodiversity data in GBIF (see **Chapter 7.6**).
  - A data repository called GeneRAP ("Genetic Resources APplication") dedicated to managing biodiversity data by data curators, with endpoints aimed at connecting the digital resources with national and international aggregators of biodiversity data and research infrastructures (to be finalized, still offline) (see **Chapter 7.4** and **Appendix 3**).

[5] GRSciColl - The Global Registry of Scientific Collections (https://scientific-collections.gbif.org/)

- A mini-cloud based on a open-source object store (MinIO) (https://datalake.ibbr.cnr.it:8080/login) aimed to host large volumes of images, documents (PDFs), papers, protocols, and other digital objects related to NSC specimens (see **Chapter 7.5**).
- A web portal (Climate-DT – https://climatedt.org) for downscaling past and future climate data at specific geographic locations, corrected by the elevation of the requested sites (see **Chapter 7.9**).
- A series of suitable endpoints for recovering NSC data and metadata (as RSS channels and JSON objects) from the CNR-IBBR-BA repository through APIs (Application Programming Interface) (**Chapter 7.7**).

All the facilities and services listed above are housed in a computing facility built with hardware funded by ITINERIS, located at the CED of the *Area della Ricerca* CNR in Bari, Italy.

3. DATA MANAGEMENT PLAN (DMP)

The data management plan (DMP) encompasses a general description of the types of data contained in the CNR-IBBR-BA repository, the data model adopted, the data policy and licensing, and the actions taken to ensure the FAIRness of the data. The current DMP was developed in tight collaboration with other ITINERIS OUs involved in the construction of one of the Italian nodes of DiSSCo-RI. However, the final version of the DMP shall be integrated with the forthcoming data management plan of DiSSCo-RI, which is expected to be released shortly.

3.1 NSC data types

Each dataset managed by the OU CNR-IBBR-BA includes a list of living biological specimens of the CNR-DiSBA research collections and their biological and ecological characteristics. For each specimen, different kinds of information are available according to the diverse organisms maintained in the collections:

- **Identifiers**: each specimen is tagged with multiple permanent identifiers (both universal and local), which help identify fragmented or redundant information of specimens currently spread across different repositories (e.g., EURISCO[6], GENESYS-PGR[7], SUSMIRRI[8], GBIF, etc.).
- **Taxonomic identification** to the lowest taxa (from kingdom to species, subspecies, variety, cultivar, etc.) and reference (URI) to the taxonomic backbone adopted. Currently, 2,265 taxa from 197 families of 5 kingdoms (Plantae, Fungi, Bacteria, Animalia, Viruses) are represented.
- **Geographic information** of the collection site (continent, country, region, city or village, geographic coordinates and their precision, datum, etc.). Currently, 147 different countries from all continents, except Antarctica, are represented.
- **Dates** (collection date, acquisition date, preparation date, etc.) and **data provenance / agents** (e.g., information of the people involved in sampling, identification, recording, specimen preparation, data curation, digital recording, etc.) are associated with most specimen records. The chronological coverage of the collections varies from 1952 (fungi and bacteria) to present. In some cases, only the year of collection/recording is known.

[6] EC-PGR EURISCO (https://eurisco.ipk-gatersleben.de/apex/eurisco_ws_dev/r/eurisco/home)
[7] GENESYS-PGR (https://www.genesys-pgr.org/)
[8] SUS-MIRRI (https://susmirri-mbrc.di.unito.it/)

- **High-resolution images** and thumbnails are available for part of the specimens included in the datasets, which have been uploaded to suitable repositories and coded into the specimen record as the corresponding URL(s). Each image is identified by a set of metadata, and a PUID has been assigned.
- **Preservation**: for each specimen/accession, information is available about the nature of the preserved material (e.g., living specimens, preserved specimens, material samples, etc.), the conservation site (e.g., institute, collection), and the type of preservation (such as cryopreservation, maintenance in field orchards, storage of dried samples in vacuum bags, freeze-dried DNA in test tubes, etc.). Additionally, the life stage of the conserved specimens is recorded (adults, seeds, spores, juveniles, embryos, etc.).
- **References**: DOIs and URLs of the bibliographic references (papers in journals, online documentation, etc.) describing/analyzing the specimens are included in the datasets, whenever available.
- **Genetic information**: reference (URLs) to one or more DNA sequences (genes, ITS, etc.) stored at remote repositories (e.g., NCBI/GenBank[9], EMBL/ENA[10], etc.) are included in the datasets for part of the specimens.
- **Genetic markers**: the genetic profile of specimens based on molecular markers (e.g., nuclear and organelle Simple Sequence Repeats - SSRs) is available for a subset of collections/datasets.
- **Phenotypic data**: several datasets include information on the varietal characterization (diagnostic morphological traits) of crop species, while other datasets include data on resistance to pathogens, pathogenicity to humans, production of metabolites in culture, etc.
- **Climatic data**: for each specimen with a known collection date and site, climatic parameters (min, max, mean monthly temperature; mean monthly rainfall) and derived indexes (BIO1-BIO12, SPEI, etc.) have been assessed using the downscaling tool ClimateDT (Marchi et al. 2024) based on past series and future climate scenarios (CRU-TS, CHELSA, UKCP18) and corrected for the elevation at the sampling site.

3.2 Data model

The high heterogeneity of living organisms maintained in current NSCs at the CNR-DiSBA (varying from viruses to forest trees) involves different kinds of information for different datasets (see **Chapter 5**), making the choice of a unique data model particularly challenging.

The DiSSCo data model is based on a specific standard called "openDS"[11] (Hardisty et al. 2019), partly derived from and entirely consistent with standard biodiversity ontologies, such as ABCD and Darwin Core[12] (DwC - Wieczorek et al. 2012). However, it could be further extended to include many other data and metadata standards. The structure of the DiSSCo Digital Specimen object is based on the GBIF UM[13] ("Unified Model"), which borrows a large number of classes and properties from the Darwin Core TDWG (Taxonomic Databases Working Group[14]) standard. As of

[9] NCBI/GenBank (https://www.ncbi.nlm.nih.gov/genbank/)
[10] EMBL/ENA European Nucleotide Archive (https://www. ebi.ac.uk/ena/browser/home)
[11] Open Digital Specimen (https://terms.dissco.tech/digital-specimen-guide, https://github.com/DiSSCo/openDS)
[12] Darwin Core Quick Reference Guide (https://dwc.tdwg.org/terms/)
[13] GBIF GitHub (https://github.com/gbif/model-material)
[14] Biodiversity Information Standards – TDWG (https://www.tdwg.org/)

now, the DiSSCo data model is still provisional (version 0.4.0), and its final version has not been released yet.

3.2.1 Darwin Core (DwC)

After discussion with the other ITINERIS partners involved in the DiSSCo-related activities, the Darwin Core TDWG ontology (including vocabularies and thesauri) has been adopted for data and metadata collections in ITINERIS to ensure full compliance with the final DiSSCo data model and the current GBIF platform, where biodiversity data are indexed based on DwC schemas. Moreover, the large number of extensions available[15] for the DwC standards, along with the related thesauri and vocabularies, adequately covers the diverse information available for CNR-DiSBA NSC accessions and favors the full interoperability with and/or the conversion to many other ontologies commonly used for biodiversity data/metadata (e.g., BCO[16], MIXS-GSC[17], MCPD-Bioversity[18], etc.). Furthermore, the DwC standards are derived from the Dublin Core metadata schema, which is widely used through RDF in the semantic web, thus enhancing the correct indexing of biodiversity archives on major search engines.

3.2.2 DwC extensions

As mentioned above, specific extensions of the DwC ontology allow for an accurate description of trait measurements, DNA-derived data, amplification and cloning data, media, literature references, etc. Using such a large set of extensions, datasets generated from experiments carried out either in the field or in the lab can be accurately described and associated to the relative specimens in NSC through the use of PUIDs, as well as observational datasets for data collected in nature.

The DwC primary standards and all the extensions used for CNR-DiSBA NSC data description have been submitted to the "FAIRness" working group of ITINERIS WP2 and included in the list of semantic artifacts used to ensure full interoperability with the ITINERIS Central Hub. Following is the list of extensions which have been selected for inclusion:

- Identifiers (to host multiple identifiers for each specimen)
  https://rs.gbif.org/extension/gbif/1.0/identifier.xml
- Multimedia (to host accessory information on images, documents, etc.)
  https://rs.gbif.org/extension/gbif/1.0/multimedia.xml
  https://rs.gbif.org/extension/eol/media_extension.xml
- DNA-derived data (to host DNA sequences, SSR markers, etc.)
  https://rs.gbif.org/extension/gbif/1.0/dna_derived_data_2024-07-11.xml
- Amplification/cloning (to host methodological information on DNA amplification by PCR or laboratory cloning of collection organisms)
  https://rs.gbif.org/extension/ggbn/amplification.xml https://rs.gbif.org/extension/ggbn/cloning.xml
- Identification history (to host information on provenance data/agents involved in changes of nomenclature/taxonomic status through time)

[15] GBIF DwC registered extensions (https://rs.gbif.org/extensions.html)
[16] BCO - Biological Collections Ontology (http://obofoundry.org/ontology/bco.html)
[17] Minimum Information about any (x) Sequence (MIxS) standard (https://genomicsstandardsconsortium.github.io/mixs/)
[18] FAO/IPGRI Multi-Crop Passport Descriptors [MCPD] (https://alliancebioversityciat.org/publications-data/faoipgri-multi-crop-passport-descriptors-mcpd)

https://rs.gbif.org/extension/identification_history_2024-02-19.xml

- Measurements (to host information on multiple trait assessment of specimens, varietal characterization, and the corresponding trials)
  https://rs.gbif.org/extension/obis/extended_measurement_or_fact_2023-08-28.xml
  https://rs.gbif.org/extension/germplasm/MeasurementScore.xml
  https://rs.gbif.org/extension/germplasm/MeasurementTrait.xml
  https://rs.gbif.org/extension/germplasm/MeasurementTrial.xml
- Material sample (to host information on preserved material sampled from living organisms)
  https://rs.gbif.org/extension/ggbn/materialsample.xml
- Loans (to host information on loans/borrows of specimens between museums)
  https://rs.gbif.org/extension/ggbn/loan.xml
- Chronometric age (to host metadata of possible paleontological specimens)
  https://rs.gbif.org/extension/dwc/ChronometricAge_2024-03-11.xml
- Germplasm Accessions (to host information related to the management of genebanks/seedbanks)
  https://rs.gbif.org/extension/germplasm/GermplasmAccession.xml
- Preparation / Preservation (to host information on the preparation of the conserved material and its preservation conditions)
  https://rs.gbif.org/extension/ggbn/preparation.xml
  https://rs.gbif.org/extension/ggbn/preservation.xml

3.3 Data policy

According to the provisional version of the DiSSCo Data Policy (Lymer et al. 2024), we followed the guiding principle that data should be "as open as possible, as closed as legally necessary", thereby promoting maximum dissemination of the stored information through the internet.

All data, metadata, and datasets available at the CNR-IBBR-BA data center are open-access and freely reusable by default, with exceptions for legal, regulatory, or sensitivity reasons. The data suppliers and contributors are responsible for ensuring data accuracy and compliance with legal requirements. Transparency, data integrity, sustainability, and reproducibility are assured through the deployment of suitable IT tools and adequate technological solutions.

The right to use the stored information is generally granted unless otherwise specified. Most data are provided under the Creative Commons Attribution-Non-Commercial 4.0 International (CC BY-NC 4.0) license[19], or in some cases under the CC0 1.0 Universal license[20], except for a few sets of records/datasets referring to biological material released under a different license (as clearly stated in the dataset metadata).

All the available information complies with current European ethical and legal frameworks and aligns with international, national, and institutional rules, ensuring respect for privacy and ethical standards. Personal information and sensitive collection details are protected under the EU GDPR 2016/679.

[19] CC BY-NC 4.0 International License (https://creativecommons.org/licenses/by-nc/4.0/)
[20] CC0 1.0 Universal (https://creativecommons.org/publicdomain/zero/1.0/)

3.4 Data FAIRness

3.4.1 FAIRification and FAIRness of NSC data

All specimen data and datasets disclosed by the CNR-IBBR-BA data center have been made findable, accessible, interoperable, and reusable in accordance with the FAIR principles (Wilkinson et al. 2016, Lannom et al. 2020 - see also **Chapter 5.4**). Considerable efforts have been made to reorganize, standardize, harmonize, and publish the digital resources hosted in the repository according to the above principles.

The indexing of the NSC meta-information by the leading biodiversity data aggregators on the web (GBIF, DiSSCo, etc) and the deposit of DOIs associated with enriched metadata in DataCite[21] make the digital resources available at the CNR-IBBR-BA data center easily "findable" to end users and the public, as well as to biodiversity data harvesters and search engines. Moreover, these resources can be easily recognized as products from the ITINERIS project through appropriate information, PUID, and links included in the exposed digital objects.

Access to data, metadata, and datasets hosted at CNR-IBBR-BA data center is granted to anyone (including crawlers and web bots), and the availability of metadata catalogs (see **Chapter 6**) and specific APIs for data retrieval (**Chapter 7.7**) make these resources fully "accessible" and machine-readable. The only exceptions are datasets subjected to embargo before the publication of the related papers in scientific literature. Appropriate controls based on timestamps have been established to allow these datasets to become freely available after the embargo period.

The choice of a data or metadata schema for biodiversity data, such as the DwC and its extensions, which has been widely used for a long time, makes the digital objects fully "interoperable" in their semantic meaning with all the national and international aggregators or platforms of biodiversity data. Further, in collaboration with the IT team of WP2 in the ITINERIS project, DwC and its extensions have been added to the list of ITINERIS semantic artifacts (see **Chapter 3.2**) aimed at mapping the different ontologies to a common standard (EOSC v. 4.00) adopted by the central hub. This inclusion ensures the full interoperability between the CNR-IBBR-BA data center and the ITINERIS hub.

Finally, the data policy permits everyone to reuse the hosted digital resources for non-commercial purposes, as long as the data authors are credited, either directly or via the references or bibliographic citations linked to the resources. Additionally, the NSC datasets are released as DwC-A zip archives (see **Chapter 7.6.4**), which include a tab-separated .csv file, enabling easy reuse of the NSC data when imported into various statistical software or spreadsheets.

3.4.2 FAIR implementation profile (FIP)

The FAIR implementation profile (FIP) of a data repository includes information on metadata schema and linking mechanisms, structured vocabularies, communication protocols, identifier types, data usage licensing, authorization and authentication techniques, etc., for hosted digital resources.

In collaboration with the other project partners involved in DiSSCo-related activities, we prepared a FAIR Implementation Profile for the hosting repository[22]. The FIP has been submitted to the IT team of the ITINERIS WP2 to be published among the services of the central Hub. Particular atten-

[21] DataCite (https://datacite.org/)

[22] DiSSCo-ITINERIS FAIR Implementation Profile (https://cnrsc.sharepoint.com/:x:/r/sites/ISMAR-DiSSCo2/_layouts/15/Doc.aspx?sourcedoc=%7B2063B26A-2B05-4432-86F8-8C1F4B47C1B8%7D&file=FIP%20mini-questionnaire.xlsx&action=default&mobileredirect=true)

tion has been paid to the FIP of DiSSCo-RI (last updated in 2022) to ensure the consistency of the FAIR enabling resources of the repository with those of the European research infrastructure. Finally, the FIP will be updated in the future (whenever necessary) to include (possible) additional resources or modify the existing ones.

4. NSC BEST PRACTICES, PROTOCOLS, AND WORKFLOWS

4.1 Shared protocols for NSC maintenance

Maintaining natural science collections over the long term involves developing standard workflows, appropriate procedures, and best practices for preservation, propagation, characterization, digitization, and documentation of experimental materials. Making this information accessible online is essential for end-users interested in taxonomy, ecology, phytogeography, and the phylogenesis of genetic resources. Additionally, thorough documentation of these procedures can encourage an open-science approach across various disciplines.

The wide range of organisms preserved at the CNR-DiSBA research collections, along with their diverse research goals, requires the use of specific protocols. Most protocols focus on pathogenic or parasitic microorganisms such as fungi, bacteria, viruses, and nematodes that are relevant to agriculture. In some cases, these organisms are host-dependent and pose challenges for preservation in collections. For these reasons, the availability of detailed protocols for propagating these genetic resources can greatly assist end users such as NSC curators and researchers.

A set of 24 simplified protocols/workflows used at the CNR-DiSBA collections for the conservation of different types of organisms has been prepared by the partners of the former BioMemory project (see **Chapter 2.2**) in PDF format. Within the ITINERIS Activity 6.5, these protocols have been collected, endowed with metadata, and included in a metadata catalog currently available in JSON and XML formats[23]. Finally, each protocol has been assigned a DOI, and the relative metadata has been deposited on DataCite Fabrica[24] to improve its online discoverability and reusability by end users. The complete list of protocols for CNR-DiSBA NSC collected during Activity 6.5 is reported in **Tab. 1**.

Tab. 1 *- List of NSC protocols for specimen preservation at the CNR-DiSBA collections that have been endowed with metadata and assigned a DOI.*

Institute	Authors	Title	DOI
CNR-IBBA	Pizzi F, Turri F	CNR-IBBA-FANCRYOBANK protocols: Sanitary rules and blood collection	10.71780/n0aq-1337
CNR-IBBA	Pizzi F, Turri F	CNR-IBBA-FANCRYOBANK protocols: Semen collection and cryopreservation (Bulls, Bucks, Rams, and Boars)	10.71780/ftmb-9y04
CNR-IBBA	Pizzi F, Turri F	CNR-IBBA-FANCRYOBANK protocols: Semen Quality Assessment on Pre-Freeze and Post-Thaw Sample	10.71780/b375-ev96
CNR-IBE	Palanti S	CNR-IBE-CCWF Protocols: Maintenance of wood fungi at IBE-CNR	10.71780/bcdb-b206
CNR-IBE	Benelli C, De Carlo A	CNR-IBE-CRYOBANK protocols: Conservation of Malus and Prunus spp. in liquid nitrogen (-196 °C) by the dormant bud technique	10.71780/asy2-te18
CNR-IBE	Benelli C, De Carlo A	CNR-IBE-CRYOBANK protocols: Conservation of Citrus polyembryonic seeds in liquid nitrogen (-196 °C)	10.71780/hbqk-nf82

[23] XML: https://biomemory.cnr.it/api/protocol/xml/; JSON: https://biomemory.cnr.it/api/protocol/json/.
[24] DataCite Fabrica (https://doi.datacite.org/)

Institute	Authors	Title	DOI
CNR-IPSP	Raio A	CNR-IPSP-BAC protocols: Long-term storage of bacterial strains at -80 °C	10.71780/p2h8-b312
CNR-IPSP	Santini A, Luchi N	CNR-IPSP-FPOFUAOPWS Protocols: Maintenance of woody fungi and oomycetes	10.71780/cfed-c440
CNR-IPSP	Danti R, Della Rocca G	CNR-IPSP-FTFC Protocols: Long-term storage protocol of pathogenic fungi of forest trees (mycelium or spore)	10.71780/70f0-d789
CNR-IPSP	Danti R, Della Rocca G	CNR-IPSP-FTFC Protocols: *Seiridium* sp. - Conservation as spores in fruiting bodies on plant material	10.71780/4de7-8f48
CNR-IPSP	Gambino G, Pagliarani C, Perrone I	CNR-IPSP-GRINZANE protocols: Preserving grapevine in *in vitro* conditions and grapevine management in the open field	10.71780/by70-3d63
CNR-IPSP	Ciancio A, Troccoli A	CNR-IPSP-NEMACOLL protocols: EntomoPathogenic Nematodes (EPNs) cultures	10.71780/1cr7-hr92
CNR-IPSP	Ciancio A, Troccoli A	CNR-IPSP-NEMACOLL protocols: Plant parasitic nematode propagation	10.71780/3x57-e975
CNR-IPSP	Ciancio A	CNR-IPSP-NEMAMIC protocols: Conservation of fungi and bacteria parasitizing plant parasitic nematodes	10.71780/k3m3-mk36
CNR-IPSP	Accotto G, Ciuffo M	CNR-IPSP-PLAVIT Protocols: Maintenance and storage of plant viruses	10.71780/0422-73a3
CNR-IPSP	Bubici GN, Boscia D	CNR-IPSP-PPBM Protocols: Long-term storage of bacterial and fungal strains at -80 °C	10.71780/c9f1-5b3d
CNR-IPSP	Bianciotto C, Lumini E, Mello A	CNR-IPSP-SYMBIOSIS Protocols: Preserving Arbuscular Mycorrhiza Fungi (AMF) and their Bacterial Endosymbionts	10.71780/d451-f880
CNR-ISA-FOM	Bufacchi M, Mencuccini M	CNR-ISAFOM-CEO protocols: Preparation of Olive Endocarps	10.71780/k88z-hj97
CNR-ISA-FOM	Puglia GD, Toscano M, Calderaro P	CNR-ISAFOM-LS protocols: Seeds harvesting for plant biodiversity conservation	10.71780/cgte-bj42
CNR-ISB	Mariani F, Testone F	CNR-ISB-MGH Protocols: Sampling, postharvest management, and storage of seeds	10.71780/49nw-4g06
CNR-ISPA	Perrone G, Moretti A	CNR-ISPA-ITEM Protocols: Preserving filamentous fungi at -150 °C	10.71780/emgh-1k10
CNR-ISPA	Perrone G, Moretti A	CNR-ISPA-ITEM Protocols: Preserving lactic acid bacteria (LAB) at -150 °C	10.71780/ktkj-bc60
CNR-ISPA	Perrone G, Moretti A	CNR-ISPA-ITEM Protocols: Preserving *Xylella fastidiosa* cultures at -80 °C	10.71780/f2vx-sd02
CNR-ISPA	Perrone G, Moretti A	CNR-ISPA-ITEM Protocols: Preserving yeasts at -80 °C or -150 °C	10.71780/9d25-f6a6

4.2 Digitization, data curation, and integration

Numerous project-based, institutional, national, and regional databases exist, but many are often siloed, fragmented, and inconsistent regarding data structures, metadata standards, vocabularies, file formats, and curation practices, as in the case of plant genetic resources (Weise et al. 2020). These inconsistencies extend to using different taxonomic classifications, trait descriptors, and varying methods for recording passport, image, phenotypic, molecular phenotypic, and genotypic data (Blätke et al. 2021).

One of DiSSCo-RI's aims is to promote the systematic adoption of high-quality practices for the curation, management, and long-term accessibility of NSC-associated data across data repositories. This will ensure data integrity, efficient workflows, and full compliance with FAIR principles while addressing technical and operational barriers to data integration and sustainability.

To this end, detailed best practices for digitization and data management[25] have been compiled during the DiSSCo-RI preparatory projects, such as ICEDIG, SYNTHESIS+, DiSSCo Prepare, DiSSCo Transition, etc., and we recommend their adoption (with few exceptions, see below) for optimizing the processes of data acquisition and management throughout the workflow. Exceptions refer to persistent identifiers (DiSSCo-RI suggests CETAF identifiers) and the digital resources associated with specimens, for which a simplified procedure has been followed to help data curators in uploading/registering their resources.

The digitization process of the CNR-DiSBA collections has strictly followed the above recommendations, guidelines, protocols, and workflows, with some exceptions. In particular, bacterial and fungal collections have been poorly documented through images, as microscopy images are difficult and costly to obtain for each cultured strain.

4.3 MIDS (Minimum Information about a Digital Specimen)

Following DiSSCo recommendations, the minimum information about digital specimens must be considered before data publication. Details about how MIDS are calculated can be found in the DiSSCo GitHub repository[26]. A similar evaluation has been done for specimens/datasets hosted in the CNR-IBBR-BA repository (see **Chapter 5.4**) to provide end users and the public with an assessment of the completeness and quality of the data.

According to the ITINERIS objectives, MIDS are recalculated for each digital record upon adding new information or updating the existing one. As for datasets, the MIDS are calculated as the average MIDS over all specimens included in the datasets.

Following is the list of the MIDS levels redrawn from the DiSSCo GitHub, with the main characteristics of the digital objects considered for each level. The main mandatory fields for each level are listed in the openDS data schema (ods:propertyName). At the same time, the corresponding DwC term (or DC term) is reported in parentheses (dwc:propertyName / dcterm:propertyName).

- MIDS-0 (Bare records): a bare or skeletal record making the association between an identifier of a physical specimen and its digital representation, allowing for unambiguous attachment of all other information. This includes a specimen identifier (e.g., barcode), an entry in a catalog or database, and the institution where the specimen is held. However, the specimen's image(s) can be generated and referenced at this early stage. The following fields are mandatory:
  - ods:physicalSpecimenId (dwc:occurrenceID)
  - ods:organizationId (dwc:institutionID - e.g., ROR[27] or Wikidata identifier of the organization).
- MIDS-1 (Basic records): a basic record of specimen information that enables basic discoverability and how the user can use the data. In addition to the information reported above (MIDS-0), MIDS-1 includes the scope of the physical specimen, the kind of object, the scientific name of the specimen, the last record update, and the user permissions/information usage. The following fields are mandatory:
  - dcterms:license (dcterm:license)
  - ods:modified (dcterm:modified)
  - ods:objectType = ods:livingOrPreserved (dwc:basisOfRecord)

[25] DiSSCO Digitization Best Practices (https://dissco.github.io/DataManagement/Data.html)
[26] DiSSCo GitHub (https://github.com/DiSSCo/openDS/blob/master/mids-calculation/intro.md)
[27] Research Organization Registry (https://ror.org/)

- ods:type (the DOI to the FDO type/schema of the object)
- ods:specimenName (dwc:scientificName).

- MIDS-2 (Regular records): a regular level of information, including data that has been agreed over time to be essential for most scientific purposes. It includes the latitude/longitude coordinates, the collector name, the collecting number, and the collection date. For geological and paleontological specimens, the inclusion of geological age is a priority. The following fields are mandatory:
  - ods:collectingNumber (dwc:fieldNumber)
  - ods:collectorName (dwc:identifiedBy)
  - ods:dateCollected (dwc:eventDate)
  - dwc:typeStatus
  - ods:hasMedia (dwc:associatedMedia)
  - dwc:decimalLatitude
  - dwc:decimalLongitude
- MIDS-3 (Extended records): an extended level of information about a specimen, including identifiers enabling connections to be made to other data present or known about the specimen. These can include information about the collection site (location, habitat, etc.), links to associated genetic information (DNA sequences or markers), links to related references (published papers), specimens' preservation conditions, preparation methods, data provenance (information of the people involved in sampling, identification, recording, preparation of specimens, etc.), and other accessory information.

4.4 Permanent unique identifiers

DiSSCo-RI recommends assigning NSC resources a permanent unique identifier (PUID) to ensure the traceability of different objects related to the same organism/material across different repositories[28]. Examples of PUIDs include DOI for datasets/specimens, ORCID for individuals, ROR for institutions, and NCBI TaxonID for taxa, among others. PUID assignment to digital resources improves data findability across the internet and minimizes the fragmentation of information for the same resource across different repositories (Wu et al. 2024). This also facilitates the citation of digital resources in publications and allows for discovering and retrieving the (meta)information in other documents spread across the Internet (Cortea 2025).

PUIDs must be assigned to each digital resource (e.g., specimens, datasets, collections, images, documents, etc.) upon their inclusion in the master database to ensure data traceability within the terrestrial domain of the ITINERIS project. After discussion with the project partners involved in DiSSCo-related activities, the adoption of DOI (Digital Object Identifier) has been chosen. The assignment of DOI to each digital object of the CNR-DiSBA collections is still ongoing (see **Chapter 5.7**), taking advantage of the membership of the CNR to the DataCite™ consortium headquartered in Hamburg, Germany, which hosts the DOI and redirects to the corresponding landing page.

DOI minting at the DataCite Fabrica, which serves as the "back office" of DataCite, also facilitates the inclusion of metadata of the digital resource. For metadata deposition purposes, several fields are required, including DOI, landing page, publisher, publication year, and resource type, whereas others are optional. It is recommended here that the information associated with the DOI deposit should include the following metadata:

[28] DiSSCo Persistent Identifier Best Practices (https://dissco.github.io/DataManagement/IdentifierRecs.html)

- creators/contributors (name, affiliation, ROR code of the corresponding institute, personal ORCID, contributor type, etc.);
- identifiers (multiple PID/URLs for resources duplicated elsewhere);
- rights and licensing (rights holder, license type, etc.);
- related identifiers: a set of fields that indicate the relationship of the digital object whose DOI is being deposited with other related digital resources (e.g., the parent dataset and collection of a specimen, the metadata repository of a dataset, etc.), including the relation type, the identifier (PUID/URLs), and the related metadata schema.

The availability of extensive and user-friendly documentation on the DataCite platform[29] helps develop specific APIs to create/update new DOI deposits as JSON objects. It is recommended to register the DOI in the local database only after receiving a successful response from the DataCite server to avoid registration failures due to possible duplicate suffixes, which lead to records missing the assigned identifier. Moreover, updating the meta-information associated with the DOI is recommended whenever new or updated (meta)data becomes available for each digital resource. Finally, developers can refer to the section "Best Practices for DOI Registration"[30] of the DataCite online help to obtain more detailed information on the optimal registration procedures, which are being strictly followed during the assignment of a DOI to the digital resource of the CNR-IBBR-BA repository.

4.5 Associated digital resources

As for the documentary material (images, PDF, etc.) associated with specimens/datasets/collections, a specific repository based on user-friendly open-source object store (MinIO) has been implemented (see **Chapter 7.5**). This will facilitate partners in mobilizing digital resources that are not already available online, e.g., which are currently kept on local/offline hard disks. Each uploaded file/resource should be supplied with rich metadata that accurately describes the content based on the Audubon Media/Audiovisual Core (AC) TDWG standards and vocabularies. The assignment of DOI to each digital resource associated with specimens/datasets is ongoing.

A backup copy of the digital object repository will follow the approach outlined for data, metadata, and datasets (see **Chapter 7.2**), ensuring their safety, proper maintenance, and quick recovery in case of failure.

4.6 Recommendation for using the IPT platform

The IPT (Integrated Protocol Toolkit) is a tool developed by the ICT team at GBIF, designed to assist collections' curators in making their NSC data accessible online in full compliance with the FAIR principles. It also allows the registration of biodiversity data on the GBIF platform, making them fully discoverable to a large public and indexed by dedicated harvesting/crawling engines. The IPT software requires a remote server equipped with a Linux operating system (Ubuntu 22.04 LTS in our case), Apache Tomcat 9.0+ servlets, and a Java EE platform. The IPT platform is easy to use, though it is relatively challenging to install and fine-tune.

IPT has been installed on the CNR-IBBR-BA server to host the CNR-DiSBA datasets, as well as datasets from other data providers (Natural History Museums, Botanical Gardens, etc.) obtained either within national and international projects dedicated to biodiversity or from other PNRR projects, such as Agritech, National Center of Biodiversity, etc. The large capacity of data storage available at the CNR-IBBR-BA data center allows for hosting a large amount of data/metadata,

[29] Creating DOIs with the REST API (https://support.datacite.org/docs/api-create-dois)
[30] Best Practices for DOI Registration (https://support.datacite.org/docs/best-practices-for-datacite-members)

along with images, PDFs, other documents, etc. Indeed, this resource represents a service available to the entire scientific community committed to biodiversity data publishing.

Upon documented request, curators and researchers interested in sharing biodiversity data can be assigned a dedicated account on the IPT platform hosted on the CNR-IBBR-BA platform. This account enables them to independently manage the publication of their datasets on GBIF, exploiting a set of IPT tools designed to facilitate sharing biodiversity information online.

Below is a list of recommendations to be followed before, during, and after the data publication workflow through the IPT. These best practices have been included in this document to make it easier for anyone to use the IPT platform available at the CNR-IBBR-BA data center. For more detailed instructions, please refer to the IPT user manual[31] available online.

4.6.1 Data preparation

The NSC data can be easily managed using any popular spreadsheet (Microsoft Excel, LibreOffice Calc, Google Spreadsheet, etc.) and exported as a plain text file ("occurrence.txt") where rows are separated by carriage returns (\r\n) and columns are separated by tabs (\t). The first column of the spreadsheet should contain a permanent unique identifier (PUID) of the specimen (DOI, local identifier, etc.). We chose a 36-character alphanumeric string (e.g., 865c2da5-660d-3499-a47b-d80d4f000ac7) randomly generated by the MD5 algorithm as a PUID of specimens, but the URL of the landing page of the record in the repository could also be used. One should take care that the permanent identifier is unique across different datasets and permanent through time.

It is strongly recommended to include the largest amount of information available for each specimen at this stage, including the PUID of the parent collection on GRSciColl (see below), the PUID of the parent dataset on GBIF, and the PUID of the institution hosting the parent collection (e.g., ROR) to increase machine-to-machine interoperability. Further, the inclusion of the ORCID for individuals who have identified, recorded, georeferenced, or prepared each specimen is essential to acknowledge their contributions to the conservation of genetic resources through time. The use of PUID improves data findability across the internet and can help minimize the fragmentation of information for the same resource across different remote repositories (Wu et al. 2024, Cortea 2025).

To help data managers in the data curation process, a list of 72 DwC descriptors (out of the 900+ available) has been selected, and each term has been described in **Appendix 1**). The use of most such descriptors can effectively convey the key information of NSC specimens and enhance the data richness of the digital specimens.

It is also important to emphasize the need to include links to references for each preserved specimen, such as published papers describing it or its characterization, along with links to any DNA sequence information obtained from the specimen and stored in specialized repositories like NCBI GenBank[32], EMBL/ENA-EVA[33], and others. The availability of such information can facilitate cross-linking digital resources related to the same specimen or accession that are stored across different repositories.

4.6.2 Registration of institutions/collections

As the PUID of the collection and its owner institution are also mandatory in the occurrence data file, both identifiers should be obtained in advance by registration through the interface of the GBIF registry (https://registry.gbif.org/institution/search, https://registry.gbif.org/collection/

[31] GBIF Integrated Publishing Toolkit (IPT) User Manual (https://ipt.gbif.org/manual/en/ipt/latest/)
[32] NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/)
[33] EMBL/ENA European Nucleotide Archive (https://www. ebi.ac.uk/ena/browser/home)

search). To this end, a simple online form must be filled with the relevant collection metadata, which also includes the description, the people involved, the physical site where the collection is maintained, etc. The registration allows the collection to be included in the Global Registry of Scientific Collections (GRSciColl), which also exposes the relative metadata (JSON objects) based on the Latimer core ontology[34], a standard for collections derived from the DwC ontology. More details on the registration process are reported on the GBIF platform and supported by training videos.

4.6.3 Data uploading

Before data publication, a conformity test should be performed on the prepared data using the online "Data validator"[35], where the data to be registered and indexed on the GBIF platform can be tested against suitable schemas. Upon uploading the DwC-A archive, a detailed report is generated, which is particularly useful for correcting data inconsistencies or addressing any data gaps. The report provides a comprehensive overview of all the steps involved in ensuring data compliance with DwC vocabularies and thesauri, including issues such as incorrect nomenclature assigned to specimens, inconsistencies between the geographic coordinates of the location, country, and continent, as well as the consistency of the indexes assigned to datasets, collections, and institutions. By following this approach, the FAIRness of the NSC data to be published through the IPT and registered on the GBIF platform can be significantly improved.

4.6.4 Data mapping to DwC

Once the occurrence data has been uploaded onto the IPT platform, it is necessary to align one-to-one the descriptors (i.e., the column headers of the data file) with the Darwin Core (DwC) standards using the IPT mapping tool. Indeed, the mapping process starts with selecting the appropriate extension (see **Chapter 3.2.2**) that has fields matching the ones to map in the source data, and then pairing each field in the source data to the suitable DwC term through a dropdown menu.

When the column headers in the first row of the "occurrence.txt" data file are labeled with the appropriate DwC terms (see **Appendix 1**), the above process is straightforward, as the software automatically maps the fields to the correct DwC descriptors. However, when column headers use different labels, the mapping process can become challenging and time-wasting, due to the extensive list of terms available in the dropdown menus to be paired with each header. Therefore, we recommend labeling each column in the file with the corresponding DwC term in advance to simplify this step.

More detailed information on the mapping process and how to use the IPT mapping tool can be found in the online manual prepared by the ICT team of GBIF[36].

4.6.5 Dataset publication and DOI assignment

Using the publication functionality of the IPT platform, users can easily make available their uploaded datasets to the public and data crawlers/harvester engines. Moreover, the publication allows assigning a DOI to the uploaded dataset through GBIF, which is automatically registered at DataCite, along with its metadata.

The first step is to set the visibility of the resource on the IPT. The visibility of any dataset on the IPT determines who can view it. Options include: (i) Private: only accessible to managers; (ii) Pub-

[34] Latimer Core (https://ltc.tdwg.org/quick-reference/)
[35] GBIF Data Validator (https://www.gbif.org/tools/data-validator)
[36] Map your data to Darwin Core (https://manual.obis.org/ipt.html#map-your-data-to-darwin-core)

lic: available to everyone; (iii) Registered: globally discoverable via the GBIF website; (iv) Deleted but Public: no longer active but still available to everyone because it was assigned a DOI.

By default, each dataset is visible only to the user who created it and to other users with the "admin" privileges on the IPT where the resource was created.

When publishing a new dataset or a new version of a dataset, a new zip archive (DwC-A files) is created on the IPT, and a specific endpoint is created for metadata in EML/XML format. The best practice is to enter a change summary for each newly published version to help end users understand which version of the data to use.

Registering the dataset on the GBIF platform via IPT is straightforward. Upon registration, GBIF crawlers will queue to ingest the new dataset version, which will be processed and made publicly available on the remote platform shortly, depending on server traffic. Any issues in the dataset that could prevent its publication can be checked by visiting the GBIF registry's dedicated interface, section "Datasets" (https://registry.gbif.org/dataset/{UUID}/ingestion-history, where UUID is the unique identifier assigned to the dataset).

5. DIGITIZATION OF CNR-DISBA COLLECTIONS

5.1 Overview

As mentioned above, the NSC information collected through the ITINERIS project was highly heterogeneous in terms of nature, scope, and consistency, requiring significant efforts and extensive work for data standardization and harmonization across the different datasets. The current status of the digitization of the CNR-DiSBA collections carried out in the frame of Activity 6.5 is summarized in **Fig. 1**.

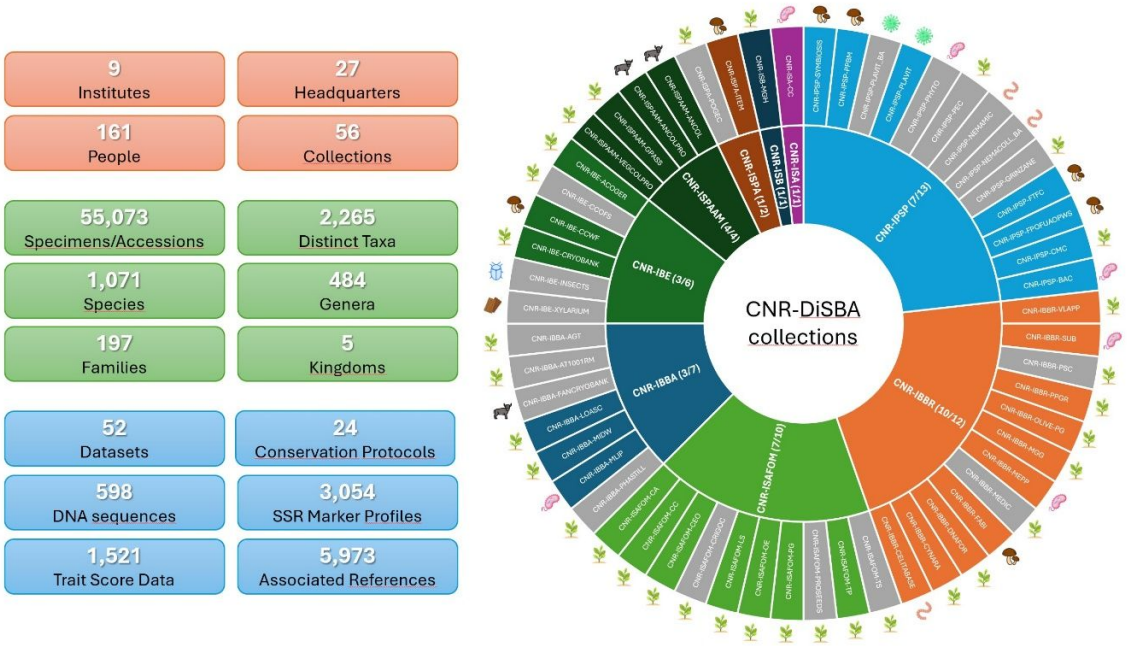


***Fig. 1** – Graphic summary of Activity 6.5 outcomes about the digitization of the CNR-DiSBA collection. The collections in grey in the pie chart on the right are still incomplete, embargoed, or not yet inventoried. Of the 52 datasets available at the CNR-IBBR-BA repository, 35 occurrence datasets have already been published and indexed on the GBIF platform.*

To date, 55,073 specimen records and 56 CNR-DiSBA collections of living organisms are available at the CNR-IBBR-BA data center, of which 23 have already been registered in the Global Registry of Scientific Collections (GRSciColl). Only collections with at least one published dataset have been registered (**Tab. 2**), totaling 50,176 digital specimens hosted on GRSciColl. The remaining collections are still incomplete, under embargo (until the results are published), not yet inventoried, or lacking the curators' consensus for their publication on GRSciColl.

***Tab. 2** - List of the 23 research collections of the CNR-DiSBA registered in the Global Registry of Scientific Collections (GRSciColl), including the number of digital records (n = 50,176).*

Collection label and title	Collection Type	Kingdom	No. Records	PUID (Click to open)
CNR-IBBA-MIDW - IBBA Duckweed Collection	In vitro culture of living organisms	Plantae	521	24fdf8b9-2165-44a5-9e30-019a2190331d
CNR-IBBA-MLIP - Microbiology Lab IBBA Pisa Collection (MLIP)	Cryopreserved collection	Fungi	262	85bb62f8-dc53-41d0-b0b0-980acdaa461d
CNR-IBBR-CELITABASE - C.elegans Italian collection	Cryopreserved collection	Animalia	417	1d38cf8c-78a6-411e-b574-533dc840b43c
CNR-IBBR-DNAFOR - DNA and samples for genetic analyses of European forest tree species	Cryopreserved collection, Material samples	Plantae	4,285	90a41989-9906-41d4-aff9-0dc8420c7a45
CNR-IBBR-FABI - Fungi of Agro-Biotechnological Interest	Cryopreserved collection	Fungi	684	2b6b8e4d-3b03-4d0b-845f-b8276e2ce9ae
CNR-IBBR-MEPP Mediterranean Plant Endophyte and Pathogens Culture Collection	Cryopreserved collection, in vitro culture	Bacteria	63	9d08fa57-d04f-4b1c-835c-f96fb8633aa8
CNR-IBBR-MGG - Mediterranean Germplasm Genebank	Cryopreserved collection, Seed collection	Plantae	26,686	15ae67ae-82f3-449a-9e78-c336989b8f90
CNR-IBBR-OLIVE-PG -Olive Varietal Field Collection	Field collection	Plantae	1,703	0ccffdaa-420d-4d6a-a862-439537602f08
CNR-IBBR-PPGR - Perennial Plants Germplasm Repository	Field collection	Plantae	438	f835e198-2dbf-4978-8cef-c0a81a6030fd
CNR-IBBR-SUB - Rhizobacteria promoting plant growth and bacteria degrading toxic pollutants	Cryopreserved collection	Bacteria	84	bf6ef975-8cde-47ea-9037-9ca01d8d61ea
CNR-IBBR-VLAPP - Annual Legumes / Perennial Grass Collection	Seed collection	Plantae	122	4b49e8c2-f2a9-439a-af96-c4ba7899f890
CNR-IBE-ACOGER - Autochthonous Certified Olive Germplasm of Emilia Romagna	Field collection, Seed collection (Plants)	Plantae	29	c2a6e0de-5334-4757-9ee3-224bdfe6782f
CNR-IBE-CCWF - Clonal Collection of Wood Fungi (CCWF)	In vitro culture	Fungi	36	287c9e1d-4cfa-49ac-935a-37b1db863d2c
CNR-IBE-CRYOBANK - Cryobank of Fruit Trees Tissues	Cryopreserved collection	Plantae	44	78c9b49e-d4a3-4451-b3b3-ef34cf06caa4

Collection label and title	Collection Type	Kingdom	No. Records	PUID (Click to open)
CNR-IPSP-BAC - Collection of Bacterial Species	Cryopreserved collection	Bacteria	36	976e0760-4db6-4209-afa4-17c4098b78f3
CNR-IPSP-CMC - Cypress Multipurpose Collection	Field collection	Plantae	4,058	da91d8a9-ced3-4b51-9970-faa9d18e2e41
CNR-IPSP-FPOFUAOPWS - Fungal Pathogens of Forest, Urban and Ornamental Plant Woody Species	Cryopreserved collection	Fungi	134	3266e65b-77cb-4d96-9ba8-5f392445f867
CNR-IPSP-FTFC - Forest Tree Fungal Collection	Cryopreserved collection	Fungi	491	6635259a-e0c9-436e-af0f-7a4a62217bf9
CNR-IPSP-PLAVIT Plant Virus Italy	Cryopreserved collection, Dried samples	Viruses	438	7b62ecf0-2d6e-4fc2-a22c-be7b674e690b
CNR-ISA-CC - ISA Culture Collection	Cryopreserved collection	Bacteria	318	04f88630-cb9a-42f6-b46b-c1f2f9a9d430
CNR-ISB-MGH - Medicinal Herb Garden	Field collection, Seed collection	Plantae	476	c084b58e-5cb5-44e0-be99-91b92f70c8ae
CNR-ISPA-ITEM Agri-Food Microbial Culture Collection	Cryopreserved collection	Becteria, Fungi, Yeasts	8,171	21c9b326-b756-4391-b302-f3e095b947cb
CNR-ISPAAM-GPASS - Germplasm Collection of Pasture Species	Cryopreserved collection, Seed collection	Plantae	680	a435ac07-ad8c-42bd-a103-1cdf98accd8d

Currently, 52 digitized datasets from the CNR-DiSBA collections are accessible as JSON objects through local APIs at the CNR-IBBR-BA repository (see **Chapter 7.7**). Of these, 35 occurrence datasets have already been published through the IPT platform, indexed by GBIF, and assigned a DOI (**Tab. 3**). These occurrence datasets are complemented by 9 DNA sequence datasets (598 sequence records from 416 accessions, totaling 438,175 base pairs), 3 SSR markers datasets (3,054 SSR markers profiles at 22 loci), two trait score datasets (63 accessions characterized for 47 different morpho-phenological traits, overall 1,521 records), two datasets with climatic data at the sites of specimen's collection, downscaled from past climatic series and corrected for elevation (1,038 sampling sites and 46 climatic parameters, totaling 47,748 records), and one dataset including all the literature references cited (319 records, each endowed with DOI/URL). All 17 complementary datasets listed above have been incorporated into the 35 occurrence datasets accessible in GBIF through the use of appropriate Darwin Core (DwC) extensions.

Finally, a dataset including the meta-information related to 24 simplified protocols for NSC specimen preparation/conservation is available on the digital platform (see **Chapter 4.1**). For all the above datasets, funding by ITINERIS Activity 6.5 is acknowledged through suitable tagged info and links.

Tab. 3 *- List of the 35 CNR-DiSBA occurrence datasets already registered and indexed on the GBIF platform. (*) endpoints for downloading specimen data in JSON format.*

Institute Dataset Label	Dataset Title	No. records	DOI	End points*
CNR-IBBA-MI IBBA-DUCKWEED-01	IBBA Duckweed dataset	521	10.15468/hstsnr	137
CNR-IBBA-PI IBBA-MLIP-FUNGI	Microbiology Lab IBBA Pisa Collection of Fungi	58	10.15468/gxet5m	126
CNR-IBBA-PI IBBA-MLIP-BACTERIA	Microbiology Lab IBBA Pisa Bacteria Collection	204	10.15468/wn7dgu	125
CNR-IBBR-BA ITA436-MGG-Hordeum	Mediterranean Germplasm Genebank - *Hordeum* dataset	2,003	10.15468/bxc8pe	102
CNR-IBBR-BA ITA436-MGG-Miscellaneous	Mediterranean Germplasm Genebank - Miscellaneous species	10,337	10.15468/rpej8x	103
CNR-IBBR-BA ITA436-MGG-Triticum	Mediterranean Germplasm Genebank - *Triticum* collection	14,346	10.15468/frm59p	101
CNR-IBBR-FI IBBR-DOUGLAS-FIR	Douglas fir (*Pseudotsuga menziesii* [Mirb.] Franco) trees from old-growth Italian plantations	718	10.15468/u63jc7	159
CNR-IBBR-FI CNR-IBBR-C-SEMPERVIRENS	Simple sequence repeats (SSRs) polymorphisms in *Cupressus sempervirens*	712	10.15468/224jzj	158
CNR-IBBR-FI IBBR-QUERCUS-CERRIS	*Quercus cerris* dataset from Eastern European populations	1,175	10.15468/qwefvq	148
CNR-IBBR-FI CNR-IBBR-PINUS-LEUCODERMIS	*Pinus heldrichii* var. *leucodermis* tissue and DNA extract collection	513	10.15468/9868ye	121
CNR-IBBR-FI IBBR-ABIES-ALBA	*Abies alba* natural populations of Italy and Balkans	1,167	10.15468/3ccqf4	112
CNR-IBBR-NA CELITABASE-01	*C. elegans* Italian collection (CelITAbase)	417	10.15468/k2pzpk	141
CNR-IBBR-NA CNR-IBBR-SUB	Rhizobacteria promoting plant growth and bacteria degrading toxic pollutants (CNR-IBBR-SUB)	84	10.15468/bprp46	108
CNR-IBBR-PA IBBR-MEPP-01	Mediterranean Plant Endophyte and Pathogens Culture Collection (IBBR-MEPP-01)	63	10.15468/bcvuf5	119
CNR-IBBR-PA PPGR-CITRUS-01	Clonal Citrus tree collection	216	10.15468/unnc2d	128
CNR-IBBR-PA PPGR-OLIVE-01	Clonal Olive tree collection	39	10.15468/4ghd28	130
CNR-IBBR-PA PPGR-VITIS-01	Sicilian grapevine clones dataset	169	10.15468/4tsjth	136

Institute Dataset Label	Dataset Title	No. records	DOI	End points*
CNR-IBBR-PG IBBR-CNR-FABI-01	Dataset of endophytic and forest	684	10.15468/arjsuy	120
CNR-IBBR-PG CNR-IBBR-OLIVE-MINOR-CV	Minor Olive tree cultivars from Umbria, Italy	122	10.15468/tmkk2g	136
CNR-IBBR-PG CNR-IBBR-PG-OLIVE-CROSSES	Progenies from various crosses of Olive tree cultivars	1,016	10.15468/4ngbyr	152
CNR-IBBR-PG CNR-IBBR-PG-LEDA-CROSS	Hybrid olive seedling collection from Leccino × Dolce Agogia clone crosses	319	10.15468/q6aj5s	151
CNR-IBBR-PG ITA463-Vicia	*Vicia ervilia* collection dataset	122	10.15468/7hkc3t	139
CNR-IBE-BO IBE-BO-ACOGER	Autochthonous Certified Olive Germplasm of Emilia Romagna	29	10.15468/mrkcvc	135
CNR-IBE-FI CNR-IBE-CRYOBANK-01	Cryobank of Tissues from Ancient Varieties of Fruit Trees	44	10.15468/7rhn8v	134
CNR-IBE-FI IBE-WOOD-BIODECAY	Fungi for wood biodegradation and preservation	36	10.15468/bpzgz2	109
CNR-IPSP-FI IPSP-FTFC	Forest Tree Fungal Collection	491	10.15468/rapbdp	138
CNR-IPSP-FI IPSP-CYPRESS	IPSP Cypress Multipurpose Clonal Dataset	4,058	10.15468/bxwze2	133
CNR-IPSP-FI CNR-IPSP-TFPD	CNR-IPSP Tree Fungal Pathogens Dataset	134	10.15468/2wsrke	107
CNR-IPSP-FI IPSP-BAC	IPSP Dataset of Bacterial Species	36	10.15468/vs7fvr	155
CNR-IPSP-TO PLAVIT-01	Plant Virus Italy - CNR-IPSP collection in Turin	438	10.15468/trnr4h	142
CNR-ISA-AV ISA01-Dataset	ISA Agri-food Microorganism Dataset	318	10.15468/2ze5cv	113
CNR-ISB-RM CNR-ISB-MGH	Medicinal Herb Garden	476	10.15468/d5zvvy	117
CNR-ISPA-BA ISPA-ITEM-02	The ITEM fungal repository dataset	7,672	10.15468/chznqw	154
CNR-ISPA-BA ISPA-ITEM-01	The ITEM fungal repository dataset	499	10.15468/57z3hx	106
CNR-ISPAAM-SS ISPAAM-CNR-GPASS	GPasS - Germplasm Collection of Pasture Species	680	10.15468/ymy5hr	124

5.2 Main characteristics of the digitized collections

Most specimens/accessions in the CNR-DiSBA collections (about 90%) are living organisms maintained under a wide range of preservation conditions, depending on the specific needs of each case. These range from cryoconservation at different temperatures (up to -196 °C) to maintenance *in vitro* on axenic agar media, up to clonal accessions grown in open-field fruit orchards. In other cases, specimens consist of DNA or tissue from georeferenced trees in the wild, preserved at -20 °C/-80 °C.

The majority of organisms preserved in the CNR-DiSBA collections are plants (approximately 78%), followed by fungi (19%), bacteria (1.65%), viruses (0.85%), and animals (0.84%). Most specimens have been collected across European countries (63.2%), but all other continents except Antarctica are also represented. Specimen data are mostly (94%) available under the Creative Commons Attribution-NonCommercial 4.0 International license (CC BY-NC 4.0), and more than 70% are classified as MIDS-1 (level of completeness in digitization – see **Chapter 4.3** for more details). The main characteristics of the digitized CNR-DiSBA collections are summarized in **Fig. 2**.

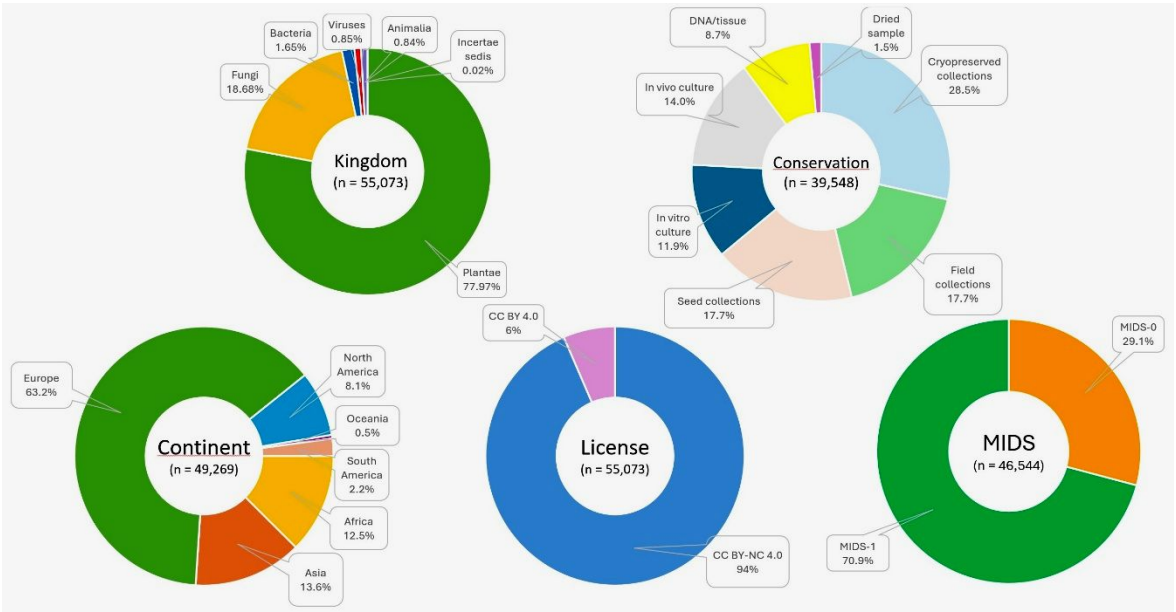


Fig. 2 *– Main characteristics of the specimens/accessions included in the CNR-DiSBA collections. The number of specimens shown in each pie chart varies depending on the available data for each characteristic throughout the entire dataset.*

Few "historical" collections at CNR-DiSBA, such as CNR-IPSP-PLAVIT (viruses), CNR-IBBR-MGG (plant seed bank), and CNR-ISPA-ITEM (fungi, bacteria) began acquiring specimens as early as the 1950s and 1970s. Since 2000, new collections (mainly fungi) have been established and have expanded rapidly (**Fig. 3**). The average timespan across collections is 17.58 ± 3.75 (SE) years, while the average maintenance time (i.e., the period between the first and last acquisition date) is 25.27 ± 3.60 years (n = 40,461).

The number of preserved accessions in each CNR-DiSBA collection varies widely, averaging 1,487.46 ± 747.50. Likewise, the mean number of taxa represented in these collections differs greatly, with an average of 50.32 ± 17.24 and a maximum of 486 (CNR-ISPA-ITEM), while the mean number of families is up to 57, averaging 8.95 ± 2.43.

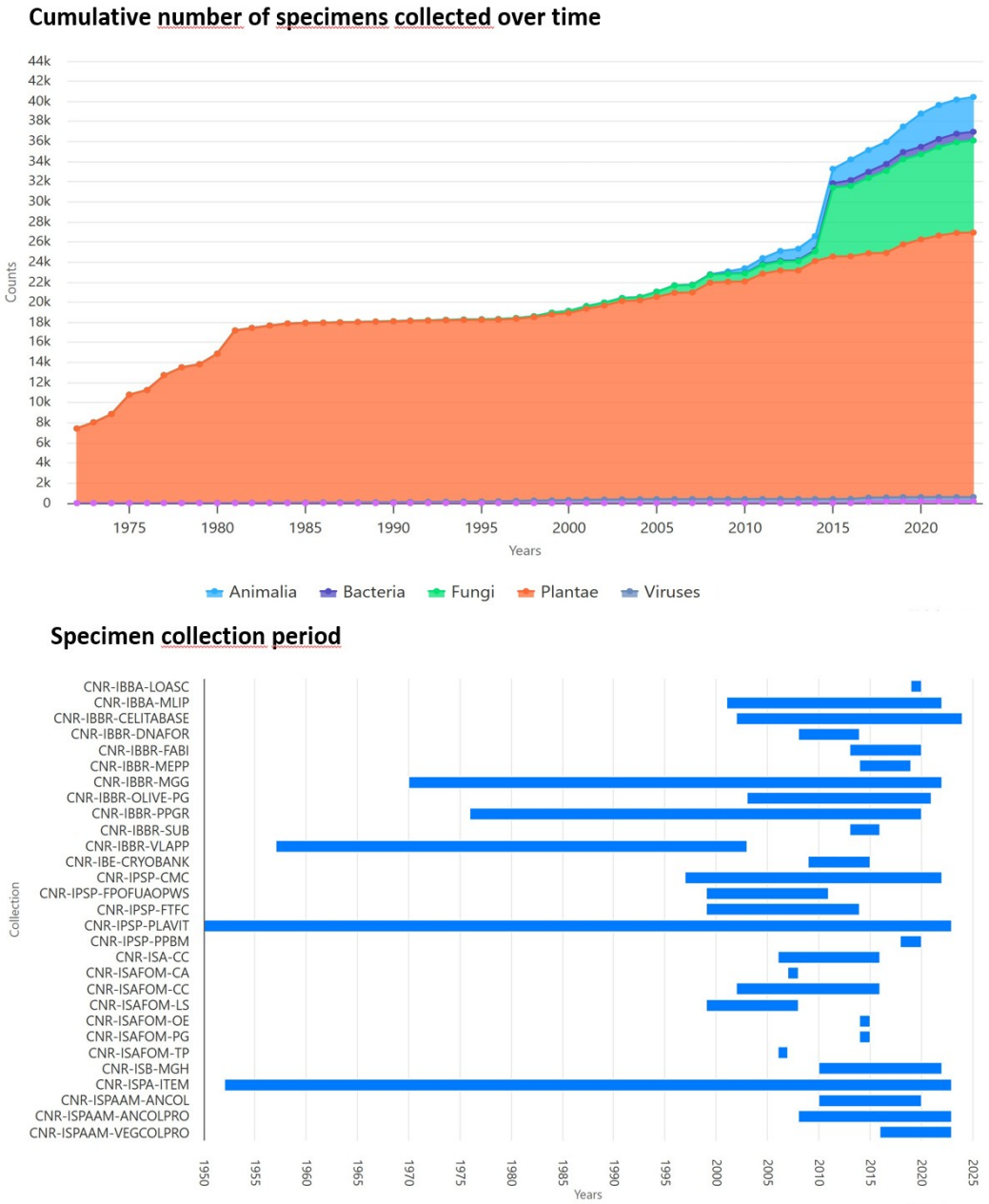


Fig. 3 *– (top) Cumulative count of accessions collected over time at the CNR-DiSBA collection, grouped by kingdom (different colors); (bottom) specimen collecting periods for several CNR-DiSBA collections.*

Notably, among the organisms maintained in the CNR-DiSBA collections, 307 accessions belonging to 16 taxa (six herbaceous plants, eight forest trees, one fruit tree, one fungus) are listed as threatened according to the IUCN Red List: *Brassica macrocarpa* (endemic to Egadi islands, Sicily) and *Lathyrus odoratus* (overall, 72 accessions) are classified as Critically Endangered (CE); *Aegilops mutica, Cupressus dupreziana, C. guadalupensis, C. goveniana* (plants), and *Fomitopsis officinalis* (fungus) are classified as Endangered (EN - overall, 152 accessions); *Aegilops sharonensis, Cupressus macrocarpa, C. chengiana, C. bakeri*, and *C. sargentii* (overall, 45 accessions) are classified as Vulnerable (VU).

5.3 Associated Info / Data Richness

Fig. 4 reports the current status of the information associated with the 55,073 specimens/accessions from the CNR-DiSBA collections. As expected, old records from historical collections have less information available than recently registered records, especially geographic coordinates at the sampling site, which are known for about 24% of accessions. It is noteworthy that a considerable number of specimens are accompanied by information regarding their preservation, acquisition date, country of origin, life stage, and data provenance details (such as agents like collectors, identifiers, etc.). Conversely, only about 6,000 records include links to related references, and roughly 3,700 records are tagged with a link to DNA sequence/SSR Marker information (e.g., DNA sequences deposited in NCBI GenBank). However, the missing data is still being retrieved (wherever available) by contacting NSC curators at CNR-DiSBA. Additionally, over 4,000 images, mainly of plant specimens, will soon be uploaded to the CNR-IBBR-BA mini-cloud (see **Chapter 7.5**) and will be made available and linked to the parent accessions.

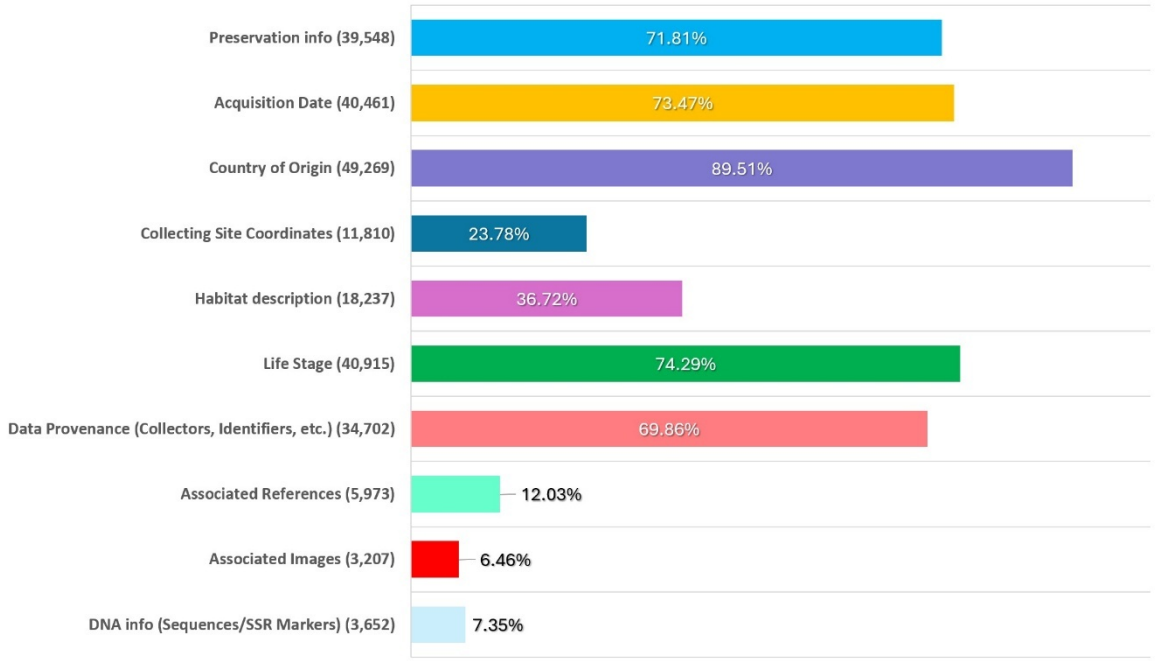


***Fig. 4** - Current status of the information associated with the 55,073 digital specimens in the CNR-DiSBA collections.*

5.4 MIDS score

The MIDS (Minimum Information of Digital Specimens) score indicates the level of completeness in the digitization of each dataset (Haston et al. 2023). Its value ranges from 0 to 3 (see **Chapter 4.3** for more details). We calculated the MIDS score for each dataset published on GBIF using the GBIF API MIDS Calculator available on GitHub (https://naturalhistorymuseum.github.io/gamc/). The results showed that 70.9% of the tested 46,544 records scored MIDS-1, while 29.1% scored MIDS-0. However, MIDS computation is based on the information available for a limited number of characteristics, which could not fully reflect the real data richness of the sample. **Fig. 5** reports an example of a CNR-DiSBA dataset that does not include any picture of the specimens (since they are represented by DNA stored in test tubes) and is therefore classified as MIDS-1, despite its data richness providing meaningful insights.

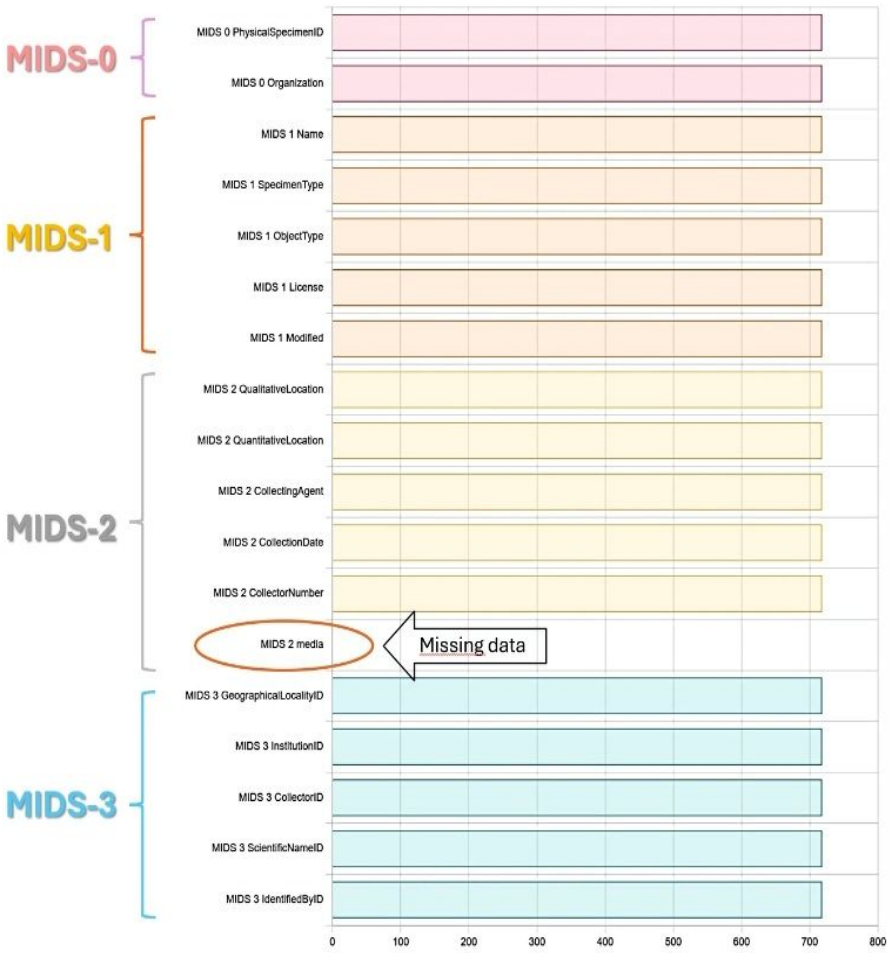


Fig. 5 *- MIDS calculation for a CNR-DiSBA dataset published on GBIF (IBBR-QUERCUS-CERRIS, n = 712). Despite the data richness of all specimens, they have been classified as MIDS-1 as no images were available.*

5.5 Data FAIRness

The compliance of the CNR-DiSBA datasets with the FAIR principles (Findable, Accessible, Interoperable, and Reusable) has been assessed using suitable tools freely available online. We evaluated our datasets with two such FAIR portals, namely the FAIR-Checker developed at INRAE for Elixir-France (Gaignard et al. 2023) and the F-UJI Automated FAIR Data Assessment Tool, created within the EU project FAIRSFAIR (Devaraju & Huber 2020).

Fig. 6 compares the results of two assessments based on an example dataset (IBBR-QUERCUS-CERRIS). The comparison was performed using the IPT web interface, producing consistent results across all CNR-DiSBA datasets. While F-UJI returned a FAIRness score of 83.33%, FAIR-Checker estimated only 65%. Notably, FAIR-Checker pointed out insufficient reusability of our datasets (see the red sectors in **Fig. 6** on the left), whereas F-UJI estimated the reusability at 100% (shown on the right in **Fig. 6**). Additional tests are required to achieve a reliable assessment of the FAIRness of the CNR-DiSBA datasets.

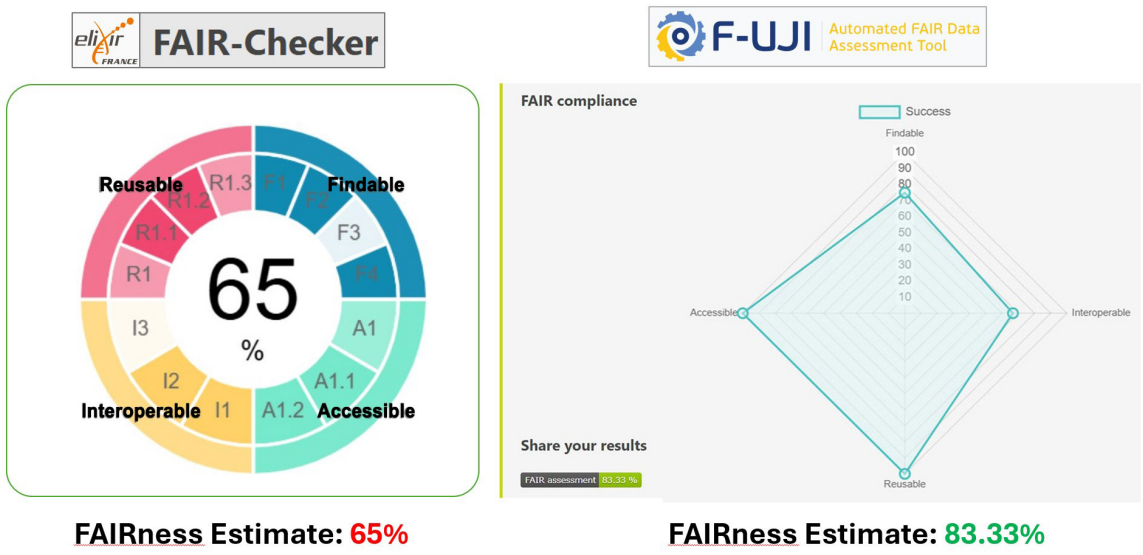


***Fig. 6** - FAIRness assessment of the CNR-DiSBA datasets using two different tools available online. For more details, see the text.*

5.6 Geographic distribution of collecting sites

Fig. 7 (left) illustrates the geographic distribution of the 11,893 georeferenced CNR-DiSBA specimens, i.e., those tagged with the geographic coordinates of their sampling sites. Most sites are located around the Mediterranean basin; however, the georeferenced localities are also distributed across Northern and Eastern Europe, Africa, Northern and Central America, and Asia. In Italy, specimens have been collected from 6,701 distinct localities, as shown in **Fig. 7** (right).

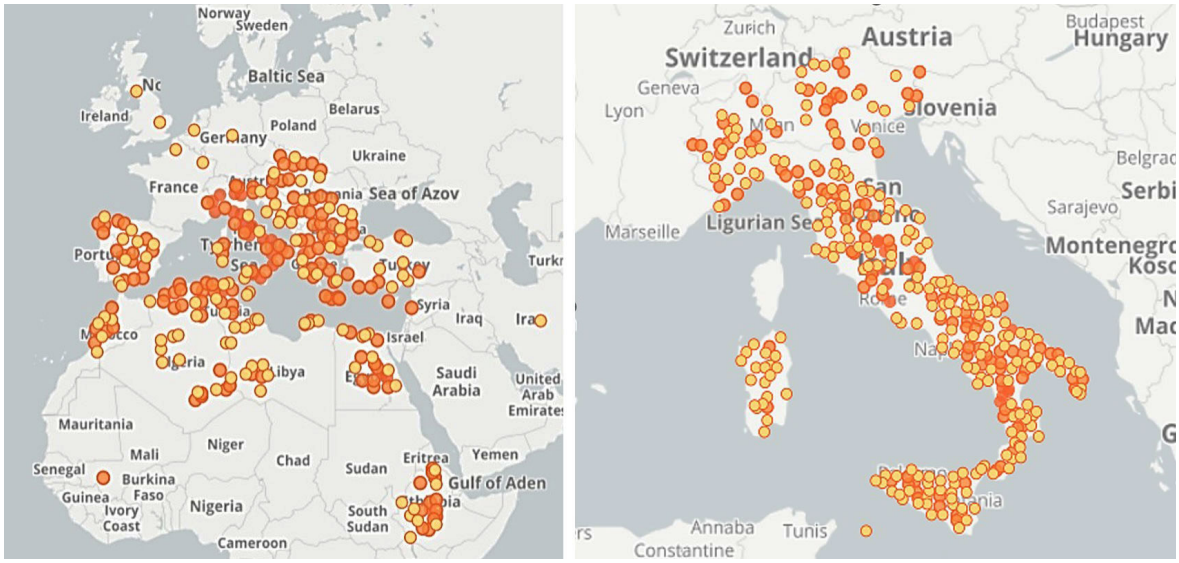


***Fig. 7** - Geographic distribution of sampling sites for the CNR-DiSBA specimens. The total number of distinct sites is 11,893 (left),of which 6,701 are located on the Italian territory (right). The color intensity of the mapped points corresponds to the number of sampling sites in each area. Maps are sourced from the GRSciColl website (https://scientific-collections.gbif.org/).*

Such biodiversity data covering a large part of Italy will help assess Essential Biodiversity Vari-

ables (EBV) across different areas and over time for many species or groups of species. Additionally, it will help in modeling the distribution of various species under future climate scenarios (Marchi et al. 2024, 2025) and mapping Italian regions at risk of biodiversity loss due to global change.

5.7 DOI assignment to digital resources

To date, all 35 datasets available on the GBIF platform have been assigned a DOI, as well as the 24 protocols for the maintenance of CNR-DiSBA collections (see **Chapter 4.1**). As for the specimens, 26,686 accessions (48,5%) have been assigned a DOI so far, specifically those of the CNR-IBBR-MGG collection. However, we developed *ad hoc* suitable routines to automate the minting process and register the DOI with their associated metadata on DataCite Fabrica by the available REST APIs[37], which takes about 1.0 seconds per accession. This allows for DOI assignment to the other CNR-DiSBA accessions to be completed in about 8 hours. The same will be done for the 77,000+ images currently stored at the MinIO object store (see **Chapter 7.5**), which takes about 21 hours (less than one day). All the above operations will be completed before the end of the ITINERIS project.

5.8 Projected size of the final data upon completion

The final number of datasets and specimens available at the CNR-IBBR-BA data repository is still unknown, as part of the CNR-DiSBA research collections has not yet been either inventoried or digitized (see **Chapter 5.1**). The repository is expected to contain approximately 80,000 individual records, with about 80 datasets anticipated to be accessible upon completion.

However, the number of data and datasets is expected to substantially increase in the coming years, as more natural history collections from museums, botanical gardens, and other collections will be digitized and indexed through the facilities deployed d uring Activity 6.5. The CNR-IBBR-BA data center has been implemented to host data/datasets from any prospective data providers and/or from past and future national and international research projects on biodiversity, as well as to host datasets from other PNRR projects, such as Agritech, National Center of Biodiversity, etc. The large capacity of data storage available at the CNR-IBBR-BA data center allows hosting a number of data/metadata to be stored, along with images, PDFs, and other documents.

6. DISSCO-ITINERIS METADATA CATALOG

6.1 Rationale

A data portal that aggregates all outcomes from Activities 6.4, 6.5, and 6.6 of the ITINERIS project was initially planned to be hosted on the GBIF platform[38] through sharing the (meta)data prepared by the OUs CNR-IBBR-BA, UNIFI-SMA, CNR-ISMAR-VE, CNR-IRSA-VB, in collaboration with GBIF's IT team. The ITINERIS data could be displayed by filtering the whole GBIF catalog (currently more than 3 billion records and 100,000+ datasets) using specific identifiers assigned to the ITINERIS datasets. A similar option has already been deployed for DiSSCo-UK[39], which implemented a hosted data portal on GBIF, collecting data on more than 17 million specimens from 290 different British institutions.

So far, the above scenario has proven unfeasible (and unlikely in the near future), as the Italian par-

[37] Introduction to the DataCite REST API (https://support.datacite.org/docs/api).
[38] GBIF hosted portals (https://www.gbif.org/hosted-portals).
[39] The DiSSCo-UK Data Portal (https://dissco-uk.org/specimen/search).

ticipation in the GBIF consortium (ruled out by the European Environmental Agency, Copenhagen, Denmark) is still under review at the Italian Ministry of Research (MUR). Therefore, we decided to implement a "Plan B" by developing an independent metadata catalog (and the parent web portal) that gathers all outcomes of Activities 6.4, 6.5, and 6.6 already published on the GBIF platform. Should the Ministry grant participation in the consortium in the future, the initial plan will be re-considered, and a DiSSCo-ITINERIS data portal on GBIF will be implemented shortly.

6.2 Description of the web portal

A shared web portal ("DiSSCo-ITINERIS Metadata Catalog" – https://dissco-itineris.it) has been implemented at the CNR-IBBR-BA data center, with the ultimate goal of deploying a "one-stop shop" that exposes all digital resources generated through the ITINERIS project regarding the DiSSCo-related activities (**Fig. 8**). The portal is hosted on a dedicated VM deployed on hardware purchased with the ITINERIS funds at the CED of the *Area della Ricerca* CNR in Bari. It has been developed in PHP/MySQL/NGINX on a dedicated virtual machine (VM), and dockerized to enhance its safety and resilience against potential security threats and facilitate the portability towards new infrastructures.

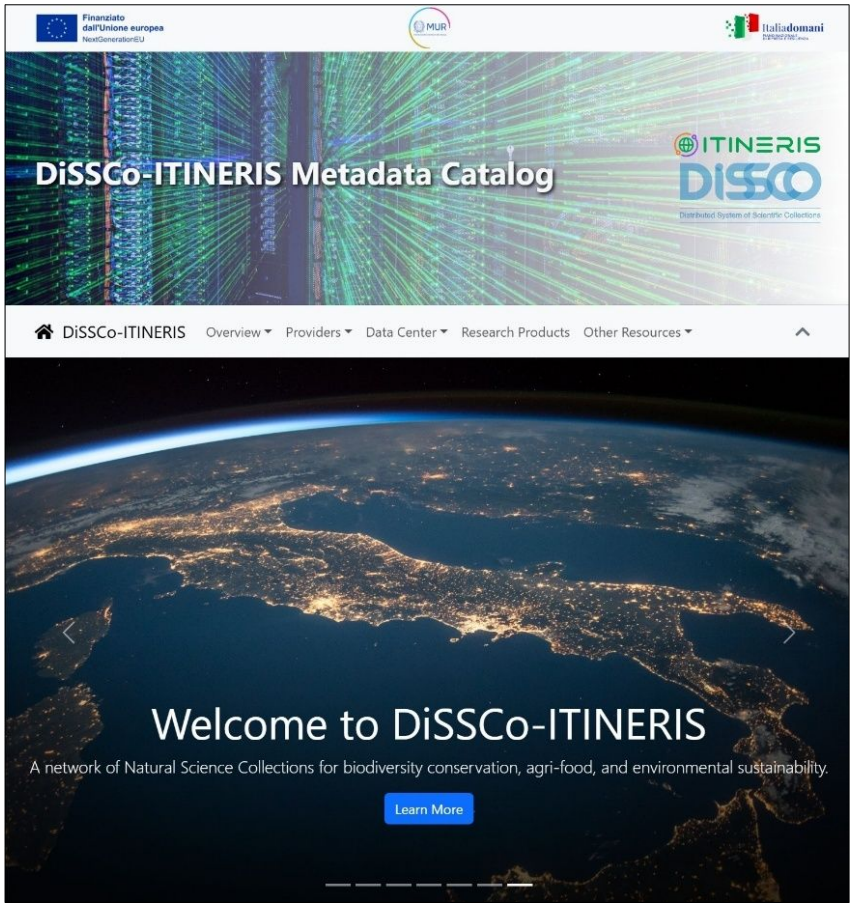


***Fig. 8** – Screenshot of the web portal hosting the DISSCo-ITINERIS Metadata Catalog.*

The metadata catalog makes extensive use of an API through HTTP-based asynchronous calls to the data/metadata endpoints available on the GBIF platform[40]. This API works against the central registry, which makes all datasets, installations, organizations, nodes, and networks discoverable. A detailed technical documentation describes how users can access data from GBIF. It provides information on the available data, how to retrieve it, where it has been cleaned or checked for data qual-

[40] GBIF API Reference (https://techdocs.gbif.org/en/openapi/)

ity, and how it should be cited. Additionally, a Swagger UI (API doc) is available where developers can independently test the results of their calls and easily consult the online documentation[41].

The MySQL relational database hosting the metadata catalog on the DiSSCo-ITINERIS web portal contains the PUID of datasets (DOI, GBIF UUID), collections (GRSciColl UUID), institutions (GRSciColl UUID, ROR), and people (ORCID), which are used to retrieve in real-time the updated meta-information from the GBIF registry. Using this approach, any update to the digital resources maintained on the remote platform is automatically displayed on the DiSSCo-ITINERIS metadata catalog.

6.3 Contents and data providers

According to ITINERIS deliverable D2.5, the available resources and the relative meta-information on the DiSSCo-ITINERIS web portal have been categorized into Providers, Services, Datasets, Research products, and others (News, Services, Training Resources, etc).

To date, the DiSSCo-ITINERIS Metadata Catalog includes the meta-information related to 154 collections, 101 datasets, 164,763 specimens, and 211 contributors from 11 institutions that shared their digital objects published on the GBIF platform. These institutions include: one University Museums (UNIFI-SMA, with its associates UNINA-SMA and DDL, the latter being a Civic Museum), eight CNR-DiSBA Institutes (CNR-IBBR, CNR-IBBA, CNR-IBE, CNR-IPSP, CNR-ISA, CNR-ISB, CNR-ISPA, CNR-ISPAAM), and two CNR-DSSTT Institutes (CNR-ISMAR, CNR-IRSA). A preliminary description of the DiSSCo-ITINERIS community can be found at https://doi.org/10.5281/zenodo. 15187687.

It is important to note that the number of datasets with available meta-information on DiSSCo-ITINERIS may rapidly change, increasing as new datasets are uploaded via the IPT installation at the CNR-IBBR-BA data center.

Further data providers are expected to publish their datasets through the CNR-IBBR-BA facilities (e.g., IPT platform) in the upcoming 10 years. These datasets will be included in the DiSSCo-ITINERIS Metadata Catalog after their registration on the GBIF platform.

6.4 Filters

A graphical user interface (GUI) allows datasets to be filtered using both pre-defined values available through dropdown controls (such as Parent collection and Institution, Dataset Type, Keywords, Publisher, and License) or searched by free text. Likewise, collections can be filtered by Collection Type, Preservation Type, Taxonomic Coverage, Geographic Coverage, and Institution, and/or searched using customized keywords.

6.5 Download center

In addition to graphical interfaces, DiSSCo-ITINERIS features a “Download center” that provides links to dataset metadata in JSON/XML formats. Here, end-users can easily download metadata for each dataset as DataCite Kernel 4.0 JSON, GBIF JSON, or EML/XML. Data itself can be retrieved as JSON objects from GBIF or as DwC-A archives via the IPT platform. A list of links to the parent collection and institution on GRSciColl is also available, along with additional links to the GUI pages on both GBIF and IPT platforms.

The complete list of datasets generated by Activities 6.4, 6.5, and 6.6 (totaling 101 datasets) is also available in RSS/XML format, which includes links to remote repositories containing data and

[41] The GBIF Registry API (https://techdocs.gbif.org/en/openapi/v1/registry)

metadata. This can be especially useful for machine-to-machine interactions, such as interactions with the ITINERIS central Hub, or harvesting by crawlers and webbots, which will improve the discoverability of the datasets across the internet.

6.6 Dashboard

A dashboard that summarizes all the contents of DiSSCO-ITINERIS has been prepared on the GRSciColl graphical interface (as no specimen data is maintained in the local metadata catalog), and a suitable link has been included in the web portal (https://dissco-itineris.it/dashboard/). The dashboard displays several statistics and counts related to the DiSSCo-ITINERIS NSC (including the CNR-DiSBA research collections), summarizing the geographic distribution of specimens, the abundance of taxa, the distribution of sampling dates, and the inclusion in IUCN Red List special categories, among others.

6.7 Research products

The list of research products (scientific papers, communications to meetings, posters, etc.) is also reported on a dedicated web page for all the OU participating in the construction of one of the Italian nodes of DiSSCO-RI (UNIFI-SMA, CNR-IBBR, CNR-IRSA, CNR-ISMAR). Those produced by the OU CNR-IBBR-BA are available at https://www.dissco-itineris.it/products/?ou=CNR-IBBR. Overall, 38 research products by CNR-IBBR are listed therein, of which nine papers are published in scientific journals, three are oral contributions to meetings, two are posters, and 24 protocols (see **Chapter 4.1**).

6.8 Other resources

All the other resources available on the DiSSCo-ITINERIS web portal are grouped under the section "Other resources." This section aims to promote external resources related to NSCs and DiSSCo activities through a brief summary, an image, and a link to each resource. These include news on initiatives related to biodiversity and natural science collection, along with valuable services for end users, such as ClimateDT (see **Chapter 7.9**), training activities, and videos like webinars organized by DiSSCo-RI, among others.

7. TECHNICAL DESCRIPTION OF FACILITIES AND SERVICES

7.1 Facilities

The data center of the OU CNR-IBBR-BA is located at the CED of the *Area della Ricerca* CNR in Bari (Italy), which is fully equipped with *ad-hoc* facilities (e.g., air conditioning, UPS, disk space, computational power, etc.) to guarantee primary data storage and backup copies of the project data, thereby providing the security needed to long-term storage of data and metadata. Backup copies of the project data are made at regular intervals on multiple supports to prevent accidental loss of the data (see below).

All the information and communication technology (ICT) services for end users are available at the CNR-CED in Bari, thus ensuring the efficiency, stability, and operational continuity of services needed for real-time data exchange with the ITINERIS Central Hub. Moreover, the CED is endowed with a monitoring system that can send multiple alerts via email upon the occurrence of connectivity problems, enabling rapid restoration of service after accidental setbacks or breakage.

The data repository has been implemented on a DELL PowerEdge R750 Server equipped with 2 × CPU Intel Xeon Gold 6342 (2.8 GHz, 24C/48T, 11.2 GT/s), 16 × 32GB RDIMM (3200 MT/s), 2 ×

480GB SSD SATA HD, and 7 × 16TB HD SATA (6Gbps 7.2K 512e 3.5in Hot-Plug). This configuration guarantees optimal performance in terms of high connectivity, fast computation, and ample storage capacity. The server above was purchased specifically using ITINERIS funds.

7.2 Data backup

Data/datasets are stored at the data center of the CNR *Area della Ricerca* in Bari, which is endowed with all the facilities needed for daily, weekly, and monthly backups of all the datasets, images, documents, etc. A safety copy is also daily transferred via SSH to a remote server with limited access, set up *ad hoc* at the *Area della Ricerca* CNR in Sesto Fiorentino (FI).

Additionally, the metadata of each specimen and dataset are available on the GBIF platform, which represents an additional safety copy as it can be freely retrieved and restored whenever needed.

7.3 Networking architecture of the CNR-IBBR-BA data center

Fig. 9 illustrates the IT facilities at the CNR-IBBR-BA data center and their machine-to-machine connections with remote repositories, research infrastructures, and the ITINERIS Hub. Such architecture has been designed to support current and future data providers who want to publish or update their biodiversity data or datasets to be harvested and indexed by DiSSCo-RI servers. It is worth noting that the data repository ("GeneRAP") is being finalized and is currently offline, but will be completed and equipped with full functionalities before the end of the ITINERIS project. Meanwhile, some functionalities, like several endpoints for recovering data and metadata of CNR-DiSBA collections, have been set up on the BioMemory platform (see **Chapter 7.7**) and will be replaced by those on GeneRAP once completed and online.

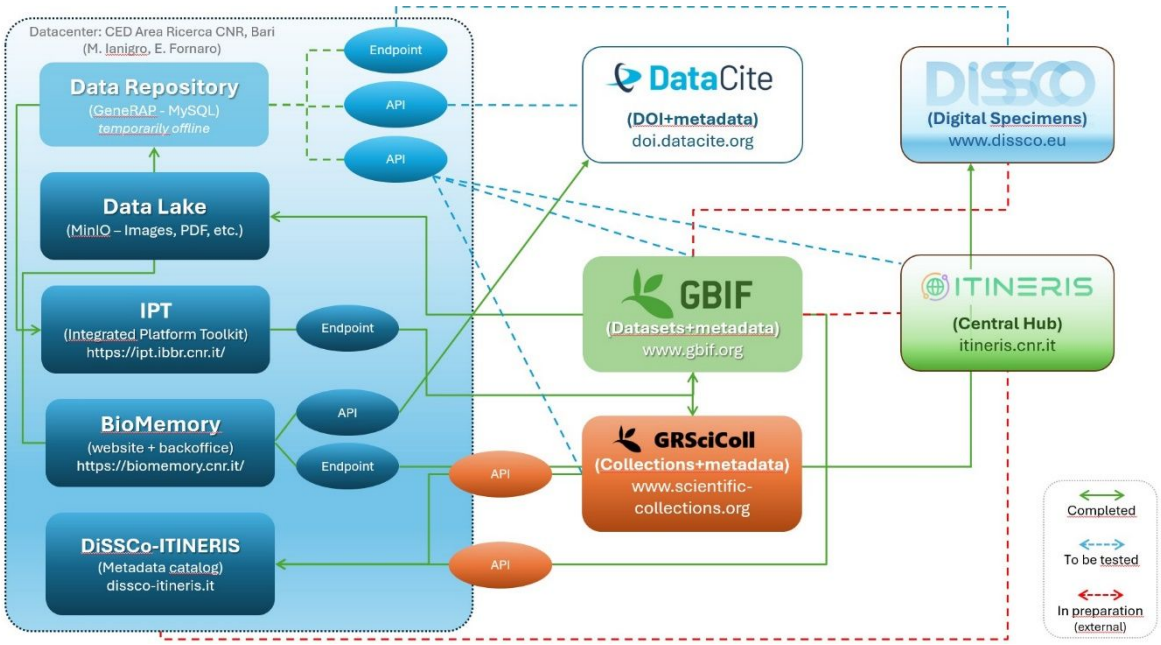


***Fig. 9** - Networking ecosystem of the CNR-IBBR-BA data center.*

7.4 The GenRAP Data Repository

7.4.1 General features

GeneRAP ("Genetic Resources Application") is a digital platform developed specifically to provide a comprehensive management system for collections of genetic resources and, more broadly, for NSCs. It was created within the ITINERIS project to be closely connected to one of the Italian nodes of DiSSCo-RI. GeneRAP offers all the features needed for digitizing, managing, and publishing biodiversity data and datasets, along with all the tools to connect with the world's leading biodiversity data aggregators such as GBIF, DiSSCo, EURISCO, GENESYS, and others. It also connects with remote repositories of genetic and genomic data, like NCBI GenBank, EMBL-ENA, and more.

The GeneRAP platform is being finalized, and it is currently offline, but will be operational before the end of the ITINERIS project. Therefore, the main characteristics of the digital platform will shortly be described here, supported by several screenshots taken from the offline portal. Once completed and made available online, it is expected to replace and enhance the functionalities currently available on the BioMemory platform in the networking ecosystem of the CNR-IBBR-BA data center (**Fig. 9**).

The GeneRAP data repository consists of (i) a containerized environment orchestrated by Docker Compose; (ii) REST APIs with Swagger UI for programmatic access; (iii) interconnection with the MinIO Object Storage for media file handling; (iv) structured SQL schemas with rich foreign key relationships; (v) dynamic JS interfaces for submissions, dashboard, filtering, and file uploads. A simplified schema of the GeneRAP technical architecture is displayed in **Fig. 10**.

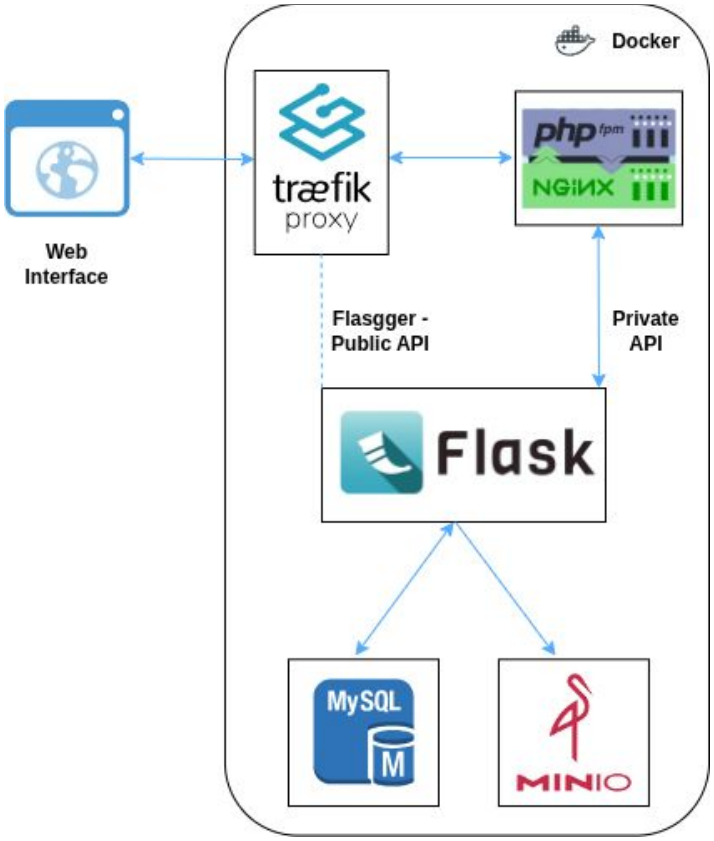


***Fig. 10** - A simplified schema of the GeneRAP technical architecture.*

GeneRAP has been built using a

modular system capable of scaling and integrating with other biodiversity infrastructures, with particular attention to the use of controlled vocabularies (DarwinCore), authentication strategies (institutional email-based registration), and modular data presentation (dashboard, requests, archive, files). The system has been devised to strictly follow the FAIR principles, making data Findable, Accessible, Interoperable, and Reusable. Several examples of usage and their compliance with the FAIR principles are reported in **Appendix 2**.

Data within the system are uniquely identified, richly described, and seamlessly indexed in a searchable interface, making them both human- and machine-discoverable. Through standardized RESTful APIs, secure role-based access, and careful tracking of provenance data, it ensures that all records remain accessible and transparent throughout their lifecycle. The platform's commitment to structured metadata, controlled vocabularies, and domain-relevant identifiers supports interoperability across systems and facilitates integration with global plant genetic resources and biodiversity networks.

7.4.2 Back office / Restricted-access Section

The GeneRAP platform is endowed with a back office to handle all CRUD (Create, Read, Update, Delete) operations necessary for managing the contents of the underlying MySQL database. It serves as the private interface for authenticated users, providing authorized personnel with advanced tools for managing biodiversity data collections in a structured, secure, and user-friendly environment.

Access to the back office is fully secured through a login system relying on institutional credentials (via LDAP) and email verification. Role-based permissions are granted to each authorized user, allowing different actions depending on the assigned role (e.g., admin, supervisor, contributor).

Contributors can access a user-friendly interface where biodiversity information in the data repository can be dynamically filtered, edited, exported, and summarized through suitable reports. Any operation on the database is sent to admins or supervisors for approval or rejection (via a suitable buffer table) before it takes effect. Additionally, all operations are tracked and stored in the database for safety and long-term reconstruction purposes.

The GeneRAP back office is also prearranged to assign each record a DOI (or update its related metadata) using a suitable API through the CNR account on DataCite.

Currently, the back office has been implemented only for the Mediterranean Germplasm Genebank (MGG – the CNR-IBBR genebank in Bari) through the interface MGD (Mediterranean Germplasm Database). Extensive documentation about the MGD Managing System, including several screenshots showing the main characteristics, is reported in **Appendix 3**.

7.4.3 API interface

The public API interface has been developed to improve the transparency, interoperability, and technical extensibility of the platform, making biodiversity data actionable, shareable, and reusable on a large scale. It is currently in its initial implementation, with future improvements anticipated before the end of the project.

The API interface can be accessed via the "API" link in the top-right corner of the platform. Presented via a clean, interactive Swagger UI (**Fig. 11**), it provides structured access to the underlying database, enabling end users to explore, retrieve, and integrate biodiversity data programmatically.

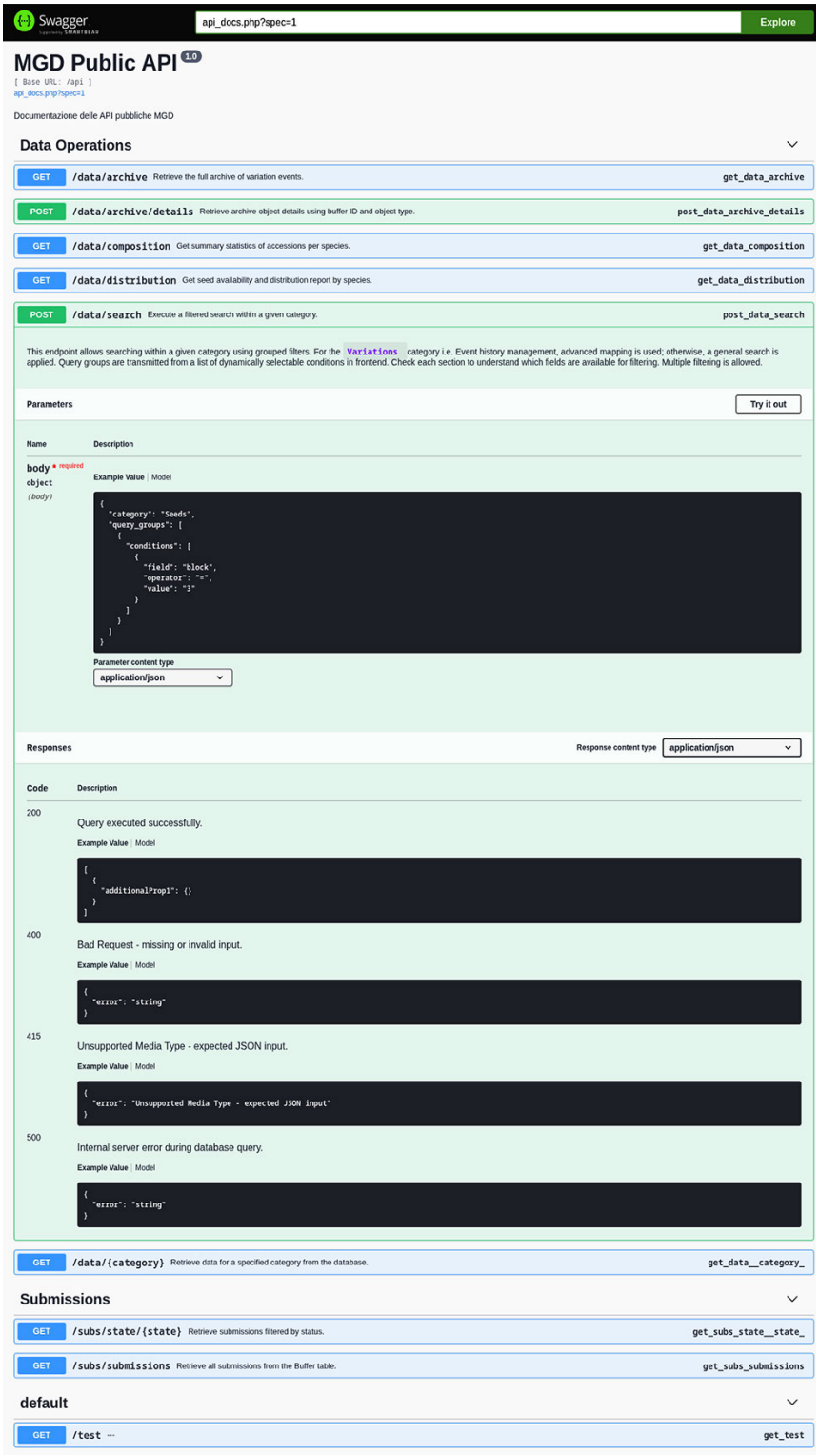


***Fig. 11** - Screenshot of the Swagger UI showing the APIs available on the GeneRAP platform.*

The interfaces available are grouped into “Data Operations” and “Submissions” sections, with suitable endpoints available for querying variation events, distribution data, archive details, and taxonomic composition. For example, the /data/search POST endpoint accepts grouped filters and allows complex queries to filter NSC data using a flexible JSON schema. This endpoint supports machine-actionable interactions aligned with FAIR principles.

The API interface features a comprehensive documentation on parameters, content types, and example responses, along with a "Try it out" feature for direct testing. This enhances accessibility for data scientists and developers working on third-party applications or integrations.

7.4.4 Web portal

The web portal is currently being finalized (**Fig. 12**) and provisionally presents four main modules, simulating different sections targeted to different collection types within the DiSSCo projects, of which only the first one (the Mediterranean Germplasm Genebank) has been developed so far.

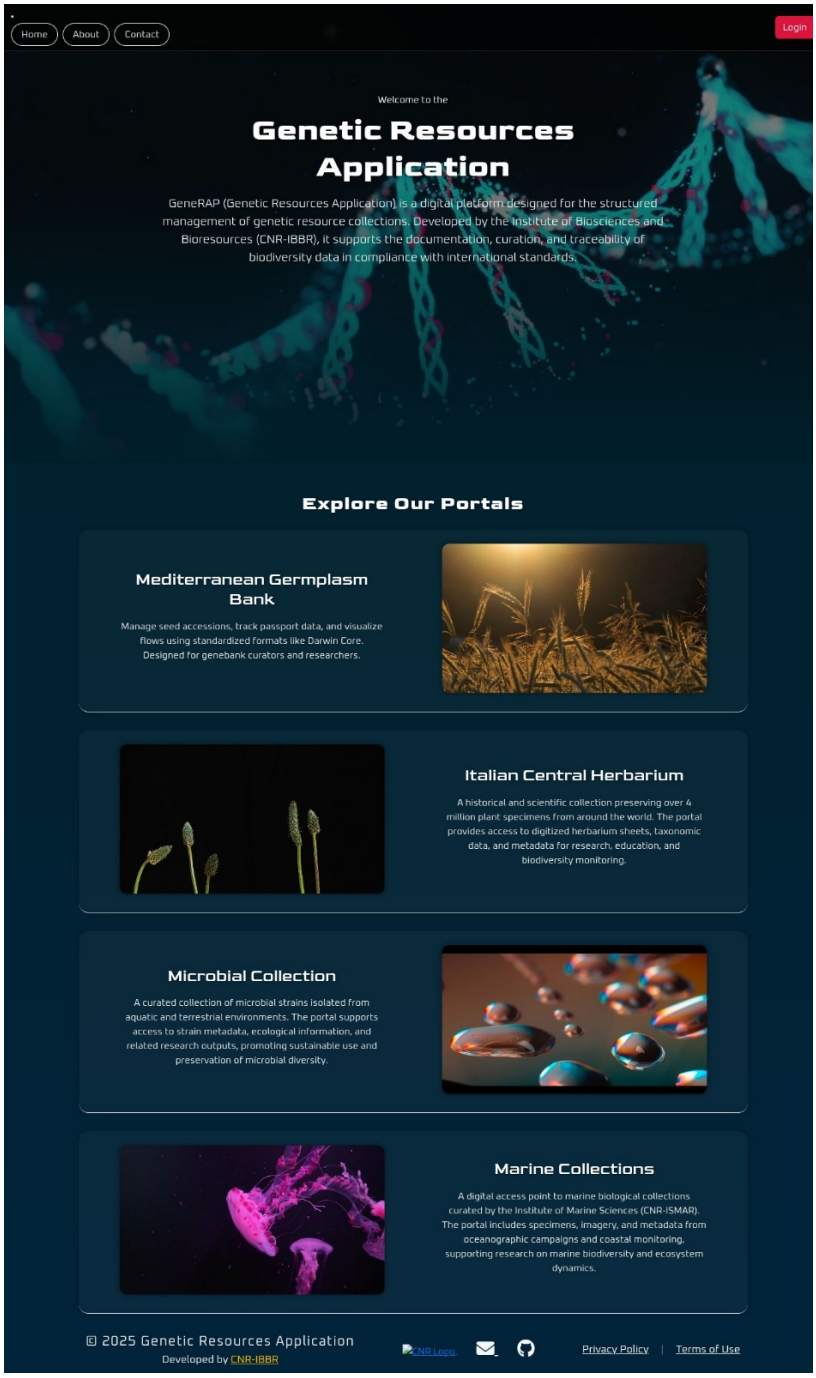


***Fig. 12** - The web portal of the GeneRAP digital platform (under construction).*

The Mediterranean Germplasm Database (MGD) is dedicated to the structured documentation and dissemination of agri-food plant germplasm collections curated at CNR-IBBR in Bari, Italy (**Fig. 13**). It provides a gateway for accessing and exploring plant genetic resource data. Compared to the current MGD web portal (https://ibbr.cnr.it/mgd), it also provides enhanced information accessibility through organized documentation, public project summaries, and a structured index. Finally,

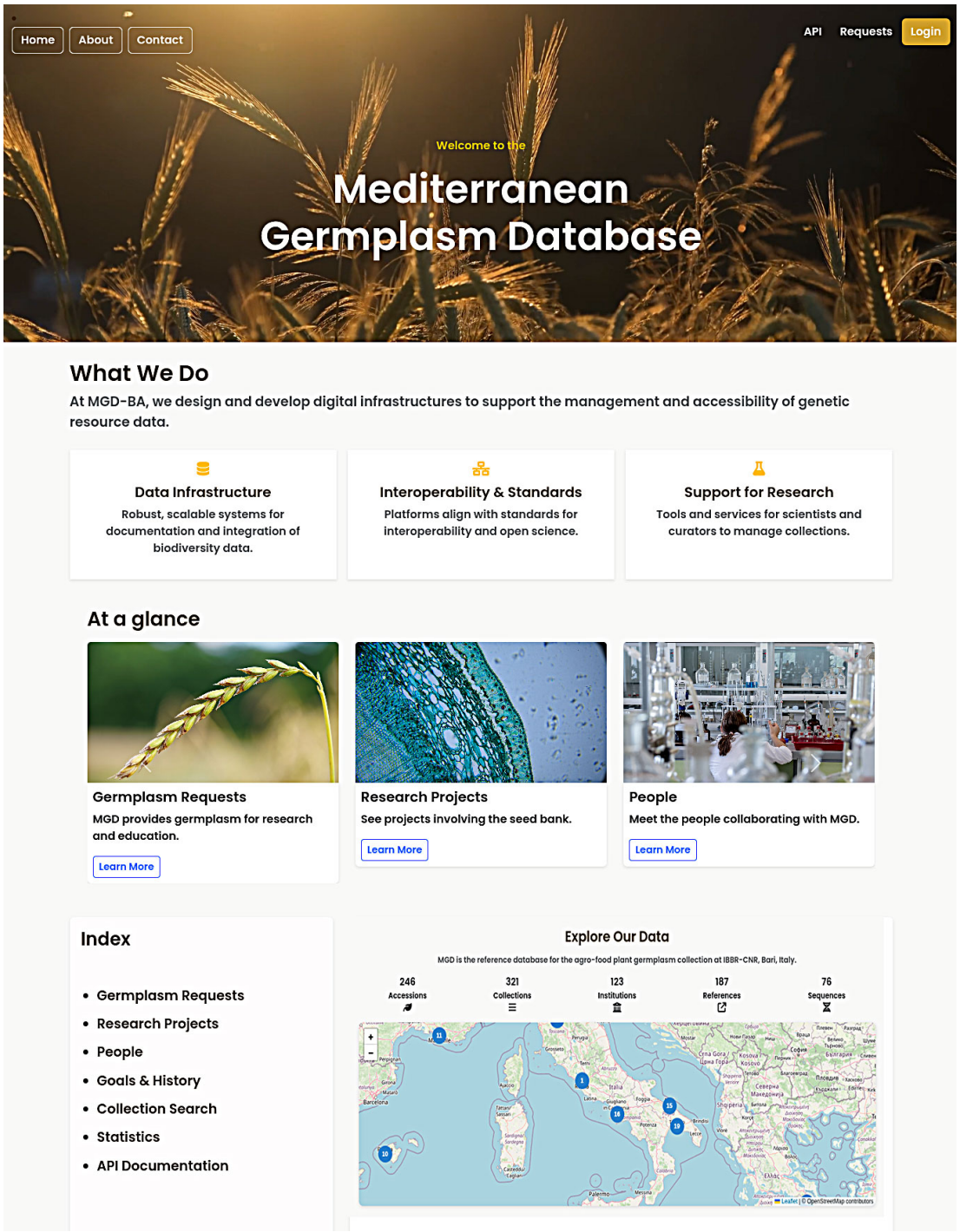


***Fig. 13** - The MGD web portal on the GeneRAP digital platform.*

transparency and data governance have been ameliorated through an appropriate privacy policy and terms of use for the content.

7.4.5 Current status and future development

As mentioned above, the GeneRAP platform is going to be finalized and, therefore, is still not available online. A suitable domain will shortly be minted to host the platform, and all the functionalities will be deployed. The ultimate goal is to create an efficient, scalable, flexible, and easy-to-use data repository for the Italian NSC collections, which can help future data providers to publish their biodiversity data in DiSSCo-RI easily.

7.5 The Mini-cloud/Data Lake

A mini-cloud based on freeware object store (MinIO[42]) has been implemented at the CNR-IBBR-BA data center to host images, documents (PDF), papers, protocols, and other digital objects directly related to NSC specimens. MinIO is an open-source, high-performance, distributed object storage system designed for cloud-native applications. It is S3-compatible, making it an excellent option for hybrid cloud environments, and it is widely used for large-scale data workloads, AI/ML pipelines, and Kubernetes-based storage solutions.

The object repository is currently endowed with 40 TB (terabytes) of local disk capacity. However, MinIO facilitates further expansion of storage capacity through the integration of remote storage units, which can be managed as a single virtual, local unit. This hybrid cloud environment option allows for large amounts of digital objects to be stored, ensuring their long-term preservation for the next 10 years. Furthermore, fine-grained control of access to independent "buckets" (cloud folders) guarantees that each authorized user can upload, edit, or delete objects solely within the authorized buckets, and different users can be permitted to manage objects within the same bucket depending on their privileges (see also **Chapter 10.3.4**).

Currently, over 77,000 high-resolution images of specimens are stored in the CNR-IBBR-BA data lake, totaling 340 GB. A large number of high-resolution images of NSC specimens from prospective data providers can thus be readily hosted in the upcoming decade.

7.6 The IPT Platform

7.6.1 IPT description

The Integrated Publishing Toolkit (IPT[43]) is an open-source freeware developed by the ICT team of GBIF and used by organizations around the world to share biodiversity data. The IPT has public pages showing rich metadata and data versioning for each dataset and a restricted-access section where data and metadata can be easily managed. Moreover, datasets can be tagged with DOIs upon registration in the GBIF platform and downloaded as zip archives along with their rich metadata.

IPT includes several features that make it easy to share structured information, such as:

- An RSS channel, listing all datasets available and their source URL.
- Endpoints for each dataset exposing enriched metadata in EML/XML format.
- Endpoints for each dataset where data and metadata can be freely downloaded as a DwC-A file (zip archive).

[42] MinIO – Exascale Object Store for AI Data, Agentic Computing, and Analytics (https://www.min.io/)
[43] GBIF Integrated Publishing Toolkit (IPT) User Manual (https://ipt.gbif.org/manual/en/ipt/latest/)

The IPT platform set up on a dedicated VM at the CNR-IBBR-BA data center can serve as a shared tool among multiple data providers. The possibility of adding multiple publishers to a single IPT installation enables many different institutions to share the same platform, as long as they are registered as GBIF publishers. Currently, the IPT platform hosts 59 datasets uploaded independently by the UNIFI-SMA staff (from Activity 6.4), which have been indexed to GBIF after agreement with the IPT installation managers. This paves the way for more data providers to participate in the coming years, thus enlarging the number of collections hosted on the platform.

7.6.2 RSS channel

The IPT exposes a metadata catalog (RSS/XML channel[44]) listing the main information of the datasets, including a short description of each dataset, its version and the date of last release, the URL to the parent enriched metadata in EML/XML format (see below) and the URL to the zip archive containing the specimen data (DwC-A archives). Such a catalog could be efficiently utilized by web crawlers and data harvesters to identify newly published datasets on the platform and to extract associated metadata from the parent EML/XML endpoint.

7.6.3 Enriched dataset metadata

Each dataset uploaded on the IPT is associated with enriched metadata encoded in EML/XML format which includes additional information about: (i) identifiers (PUID, DOI); (ii) access policy, data owner, and data licensing; (iii) creators/contributors, metadata providers and contacts (name, address, e-mail, ORCID, etc.); (iv) description, keywords and purposes of the dataset; (v) associated parties, parent project(s) and funding (including information of the project team); (vi) geographic, taxonomic and temporal coverage of the dataset; (vii) preservation and sampling methods used, and additional metadata (bibliographic citation, parent collection name and its permanent identifiers, links to additional sources, etc.).

A suitable endpoint is available for each dataset included in the IPT (and listed in the RSS channel – see above). The URL of the endpoint has the following form: https://ipt.ibbr.cnr.it/ipt/eml.do?r={archive-name}, where {archive-name} is the dataset label published by the user. Such information is easily machine-readable by the metadata harvester of the ITINERIS Hub, as well as by those of DiSSCo-RI, thus increasing the interoperability across platforms.

7.6.4 DwC-A archives

Each dataset is freely downloadable at the IPT platform as a standard zip archive in DwC-A format, which contains the following files:

- "occurrences.txt": a text file in tabular format with columns separated by tabulators (\t), one row per record (row separator: \n); the first row contains the headers with the name of the DwC descriptors; the first column contains the PUID of each specimen; null values are represented by blanks.
- A text file (txt) for each DwC extension used in the dataset (e.g., "preparation.txt") having the same characteristics reported above; the first column contains the PUID, which must match the PUID of the corresponding record in the "occurrence.txt" data file.
- "meta.xml": a file in XML format containing the list of DwC descriptors in each column of the occurrence file and the reference (URI) to the corresponding ontology source term.
- "eml.xml": a file containing the enriched metadata in EML/XML format described above.

[44] Resource metadata of Institute of Biosciences and Bioresources CNR-IBBR, Bari, Italy (https://ipt.ibbr.cnr.it/ipt/rss.do)

A suitable endpoint where the whole archive can be downloaded is available on the IPT platform for each dataset. The URL of the endpoint has the following form: https://ipt.ibbr.cnr.it/ipt/archive.do?r={archive-name}, where {archive-name} is the label of the dataset uploaded by the user.

Additional endpoints for dataset downloading are available on the CNR-DiSBA data portal on the GBIF platform[45], where a customized dashboard showing several statistics on the datasets and their usage is also available.

7.6.5 Integration with GeneRAP and the BioMemory platform

Datasets on the IPT platform can be uploaded as DwC-A archives (including EML/XML metadata), as comma-separated value (.csv), or directly read from a MySQL database using suitable queries in SQL language. The latter option has several advantages compared to the former, as any update in the database tables is straightforwardly included in a new version of the dataset to be uploaded to GBIF. For this reason, we integrated the IPT with the BioMemory platform, where CNR-DiSBA datasets were stored as JSON objects in a MySQL 8.0.1 database. When the GeneRAP platform is completed and online, the IPT will be integrated with the MySQL database of the latter platform.

7.7 The Local APIs (Application Programming Interface)

In addition to the IPT endpoints described above, datasets, data, and metadata of CNR-DISBA collections are available to end users and crawlers/harvesters (including those of the ITINERIS hub) as JSON objects at the CNR-IBBR-BA data center. To this end, a specific set of APIs (Application Programming Interfaces) has been developed (Python, PHP) to ensure full semantic interoperability with the main aggregators of biodiversity data and their compliance with the FAIR principles.

APIs are essential for data accessibility and (re-)usability and have been implemented to expose the repository data, both internally (e.g., for data visualization on the website) and externally (e.g., for remote machines/data harvesters). This enables external users to access the data programmatically and build additional tools for data visualization, selection, or analysis on top of the CNR-IBBR-BA platforms. The developed APIs provide several endpoints, each based on the best practices coming from the JSON:API specification. A set of endpoints is dedicated to the retrieval of the data, including bulk operations, while others have been developed for data management through the back office.

It is worth noting that, since the GeneRAP platform is still offline, the APIs are temporarily hosted on the BioMemory platform. When GeneRAP is operative, all the endpoints currently available on the BioMemory platform will be redirected to GeneRAP. Moreover, to facilitate the machine-to-machine data exchange and (meta)data bulk retrieval, a Swagger UI (API doc) will be available on the GeneRAP platform to illustrate all the functionality available for users and developers (see **Chapter 7.4.3**).

The complete list of available datasets (occurrences, SSR markers, DNA sequences, and literature references – see also **Chapter 5.1**) can be retrieved at the following URLs:

- https://biomemory.cnr.it/api/dataset/json/ (JSON object, schema version: DataCite kernel 4.0)
- https://biomemory.cnr.it/api/dataset/xml/ (XML format, schema version: RSS 2.0 channel)

The developed APIs expose the data of the available specimens as JSON objects. Each JSON ob-

[45] Consiglio Nazionale delle Ricerche, Istituti del Dipartimento di Scienze Bio-AgroAlimentari (CNR-DiSBA) (https://www.gbif.org/publisher/6563e0ba-fab7-431c-b897-b6bf364f4f1e)

ject complies with the DwC data model (see **Chapter 3.2.1**), including classes and properties taken from several DwC extensions (see **Chapter 3.2.2**). Furthermore, as soon as the final openDS data model is released, specific APIs will be developed to facilitate the exchange of information with the DiSSCo-RI harvesting machines.

Current APIs also allow users to filter available information using various taxonomy-related parameters (e.g., kingdom, family, genus, species), country (either by name or ISO 3166-1 alpha-2 code), or keywords (free-text strings). Upon an HTTP_GET call, a UTF8-encoded JSON object containing information of no more than 3,000 specimens (FDO) per call is returned. Larger datasets can be retrieved by making additional calls and adjusting each time the query parameters: offset (the number of initial records to be excluded) and limit (the maximum number of records returned). An example of the JSON objects returned via API, which includes occurrence data, the associated genetic information (DNA sequences/SSR markers), and details on the related references, can be seen at the following URL:

- https://biomemory.cnr.it/api/occurrences/json/?dataset=112
  (JSON object, schema version: simple DwC + DwC extensions)

The complete list of available endpoints at the CNR-IBBR-BA data center is reported in **Tab. 3**. The URL of each endpoint has the form https://biomemory.cnr.it/api/occurrences/json/dataset={ID}, where {ID} is the number reported in the last column of **Tab. 3**.

7.8 The BioMemory platform

The BioMemory platform (https://biomemory.cnr.it) was created in 2021 to host the meta-information of the CNR-DiSBA research collections. During the three-year ITINERIS project, the platform has been transferred to the CNR-IBBR-BA data center in Bari and enhanced through the implementation of a set of new functionalities and services. Within ITINERIS Activity 6.5, the BioMemory platform has been endowed with an advanced back office[46] aimed at facilitating the management of (meta)data related to CNR-DiSBA research collections, and several APIs (see above) to foster the machine-to-machine interactions with the leading biodiversity data aggregators worldwide (GBIF, DiSSCo, EURISCO, etc.) and with the other facilities at the CNR-IBBR-BA data center.

The main limitation of the current BioMemory platform arises from its origin as a CNR-DiSBA project, which hampers the participation of data providers from other CNR departments or outside the CNR. Additionally, several issues related to the current data storage can slow down the retrieval of large amounts of data, such as those expected from one of the Italian nodes of the European research infrastructure DiSSCo.

As already mentioned, once the GeneRAP platform will be completed and made available online, all the functionalities currently implemented on the BioMemory platform will be replaced on the GeneRAP data repository, which will allow security improvements and enhanced networking activities, thus ensuring optimal machine-to-machine interactions with DiSSCo harvesters.

7.9 ClimateDT portal: Climate Downscaling Tool

The Climate Downscaling Tool (ClimateDT – https://climatedt.org) is a geo-web service designed to downscale 46 climatic variables and indices derived from both past climatic series and future simulations by General Circulation Models (GCM). The core of ClimateDT is the 1 km 1981–2010 climatology from CHELSA Climate (version 2.1), where the CRU-TS layers for the period 1901-current are overlayed to generate a historic time series. ClimateDT also provides future scenarios

[46] BioMemory Back Office (https://biomemory.cnr.it/bkoff/)

from CMIP5 using UKCP18 projections (rcp2.6 and rcp8.5) and CMIP6 using 5 GCMs, also available on the CHELSA website. The system can downscale the grids and adjust for local elevation using a dynamic approach (scale-free) by computing a local environmental lapse rate for each location as an adjustment for spatial interpolation. More technical details on the underlying computations are reported in Marchi et al. (2024).

ClimateDT has been initially developed as a standalone R-based software (R Core Team 2020) within the EU-funded H2020 project B4EST[47] (2018-2022). During the ITINERIS project, a web portal has been implemented that provides a user-friendly web interface and additional functionalities for end users. More recently, a dedicated domain has been minted ("climatedt.org"), and the web portal has been moved to a dedicated VM on a server purchased using ITINERIS funds.

Through the web portal, users can upload up to 512 locations worldwide by submitting a standard .csv file that contains their geographic coordinates and elevation data. Both geographic coordinates and elevation of the uploaded locations can be corrected before submission by dragging points on a map, taking advantage of the Google Maps™ DEM and its associated geoservices. To calculate the requested parameters, the timespan for downscaling both past climate and future scenarios can be selected by dragging a range slider. Additionally, a set of radio buttons allows users to choose the preferred assessment report (AR) for future climate scenarios (AR5: rcp2.6, rcp8.5; AR6: ssp135, ssp370, ssp585), and select the GCM (only for requests including AR6: GFDL-ESM4, IPSL-CM6A-LR, MPI-ESM1-2-HR, MRI-ESM2-0, UKESM1-0-LL). After submission, the requested parameters are calculated for each location and corrected for the elevation using a worldwide DEM. Finally, an email is sent to the user's mailbox containing the URL where the downscaled file can be downloaded.

Climate-DT has proven to be a successful tool for achieving reliable and unbiased estimations of climatic variables for environmental assessments. Since its release, a total of 2,084 submissions were processed and downscaled (updated: Aug 24, 2025), with an overall number of locations successfully processed equal to 302,989 (145.4 ± 202.4 locations per request, on average). This demonstrates that ClimateDT represents a valuable free-access service for end users.

The availability of fine-scaled climatic data from climatic layers using ClimateDT will enable the modeling of several species' distributions under future climate scenarios and help identify Italian regions at risk of biodiversity loss due to global change. The large number of specimens maintained in the NSC collections with known geographic coordinates of the sampling site could be used to model the distribution of taxa in the context of future climatic scenarios (e.g., SDM - Spatial Distribution Models, ENM - Environmental Niche Models, etc.) using the ClimateDT service for environmental (climate) assessment.

Future development of the ClimateDT service will include the implementation of specific APIs to obtain downscaled datasets programmatically.

8. CONCLUSIONS

The final goal of Activity 6.5 of the ITINERIS project is to deploy the facilities, infrastructures, and services needed for the Italian natural science collections to be interconnected with the ITINERIS central hub and included in the European research infrastructure DiSSCo-RI. To this end, the construction of the data repository could be considered a "pilot" tool which can foster the digitization, reorganization, standardization, harmonization, and indexing of all current and prospective natural science collections in Italy (university/civic museums, botanical gardens, addi-

[47] B4EST Adaptive BREEDING for productive, sustainable and resilient FORESTs under climate change (https://b4est.eu/)

tional research collections, etc.) for their participation in the European RI.

According to the above considerations, efforts have been undertaken in the frame of the ITINERIS project to provide shared protocols, easy and ready-to-use IT tools, and workflows to be exploited in the future by biodiversity experts and data providers currently not included in the project, such as collection curators, data managers, researchers, etc. with the ultimate goal of mobilizing the large number of Italian NSC datasets currently unavailable, offline or fragmented across different online repositories.

9. CITED REFERENCES

- Alercia A et al. (2015). FAO/Bioversity Multi-Crop Passport Descriptors V.2.1 [MCPD V.2.1]. Bioversity International, Food and Agriculture Organization of the United Nations, Rome, Italy, pp. 11. - https://hdl.handle.net/10568/69166
- Blätke M-A et al. (2021). Advances in applied bioinformatics in crops. Frontiers in Plant Science 12: 1-3. - https://doi.org/10.3389/fpls.2021.640394
- Cortea IM (2025). Improving the findability, interoperability, and trustworthiness of the INFRA-ART Spectral Library – a dedicated data service for the heritage science domain. Zenodo. https://doi.org/10.5281/zenodo.15756790
- Devaraju A, Huber R (2020). (2025). F-UJI - An Automated FAIR Data Assessment Tool (v3.5.0). Zenodo. https://doi.org/10.5281/zenodo.6361400
- Gaignard A et al. (2023). FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. Journal of Biomedical Semantics 14; Article number: 7 - https://doi.org/10.1186/s13326-023-00289-5
- Haston EM et al. (2023). Mapping across Standards to Calculate the MIDS Level of Digitisation of Natural Science Collections. Biodiversity Information Science and Standards 7: e112672 - https://doi.org/10.3897/biss.7.112672
- Hardisty A et al. (2019). Provisional data management plan for DiSSCo infrastructure. ICEDIG Project - Deliverable D6.6. - https://doi.org/10.5281/zenodo.3532937
- Hardisty A et al. (2022). Digital extended specimens: enabling an extensible network of biodiversity data records as integrated digital objects on the internet. BioSciences 72 (10): 978-987. - https://doi.org/10.1093/biosci/biac060
- Lannom L et al. (2020). FAIR data and services in biodiversity science and geoscience. Data Intelligence 2: 122-130. - https://doi.org/10.1162/dint_a_00034
- Lymer G, Paleco C, Scory S, Graber-Soundry O, Dusolier F, Worley K, Alonso E (2024). DiSSCo Transition Project – Data policy, version 1.1. Milestone 6, Work Package 1, EU Project HORIZON-INFRA-DEV-01-02, Grant no. 101130121, pp.20. [unpublished]
- Marchi M, Bucci G, Iovieno P, Ray D (2024). ClimateDT: A Global Scale-Free Dynamic Downscaling Portal for Historic and Future Climate Data. Environments 11 (4): 82. -https://doi.org/10.3390/environments11040082

- Marchi M, Cocozza M, Bucci G, Iovieno P (2025). Spatial Modeling of Douglas-Fir Plantations in Italy After 120 Years of Experimentation. Ecology and Evolution 15 (8): e71943. - https://doi.org/10.1002/ece3.71943
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org/
- Weise S et al. (2020). Document or lose it - On the importance of information management for genetic resources conservation in genebanks. Plants 9 (8): 1050. - https://doi.org/10.3390/plants9081050
- Wieczorek J et al. (2012). Darwin Core: an evolving community-developed biodiversity data standard. PLoS One 7: e29715. - https://doi.org/10.1371/journal.pone.0029715
- Wilkinson MD et al. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3: 160018. - https://doi.org/10.1038/sdata.2016.18
- Wu M et al. (2024). Ten Principles to Improve Dataset Discoverability (1.0). Research Data Alliance. https://doi.org/10.15497/rda/00120

10. SUPPLEMENTARY MATERIAL

10.1 Appendix 1 – Recommended DwC standard terms for NSCs

This appendix is intended to serve as a guide for providers, curators, and managers of Natural Science Collection (NSC) data to be shared online through indexing on the main biodiversity data aggregators. It illustrates a set of shared best practices for acquiring, organizing, standardizing, and maintaining (meta)data related to NSC. In Activity 6.5 of the ITINERIS project, we focused on the key information that should be included for NSC specimens in the datasets, along with guidance for mapping this information to a selection of Darwin Core (DwC) standard terms, which is often the most challenging step for beginners.

Assigning the correct DwC term to each specimen characteristic can streamline the process of publishing the dataset through the IPT platform, enabling automatic mapping of the file columns with Darwin Core. The use of DwC vocabularies and thesauri (whenever available) is strongly recommended for all characteristics included in the dataset. Several examples of the use of DwC standard entries are provided below.

Following is a selection of 72 Darwin Core (DwC) standard terms (taken from both the Simple DwC and its extensions among more than 900 terms available) suggested to data curators for accurately describing each specimen of their Natural Science Collections (NSC). This list aims to support data managers in the challenging task of selecting the relevant information of the specimens to be included in the datasets. It is assumed that datasets are compiled using a standard spreadsheet file, where each row (record) represents an NSC specimen, and each column represents a specific characteristic (DwC term).

A permanent identifier (PID) must uniquely identify each specimen or accession in NSC (see **10.1.1**). It can be a local unique identifier (e.g., CNR-DiSBA-8176, CNR-IBBR-MGG-4988, ...) or a 36-character alphanumeric string generated by MD5 (e.g., ea283838-9b1c-aa4a-68dc-61e858666d7c). The latter option is strongly recommended; however, other options are reported below. Wherever the specimen has been endowed with a DOI, this identifier should be used. The PID should appear in the first column of the file and be labeled with the DwC term "occurrenceID".

The minimal mandatory terms (columns) that should be present for each record (row) in the datasets are:

- occurrenceID
- catalogNumber
- verbatimIdentification
- genus (if known)
- preservationType
- preparationMaterials

When no catalog number has been ever assigned your specimens/accessions, assign a progressive number (ID) to each specimen (row) in your dataset (e.g., 1, 2, 3, ..., 99166, ...), and label this column with the DwC term "catalogNumber". Otherwise, the original label attached to the specimen (e.g., the label on the tube or box where it is conserved) can be used. Such an ID must be unique for the accession throughout the collection and must indicate the same specimen across different datasets within the same collection.

Each specimen (row) in the file should include a column with the taxonomic identification of the accession as it appeared in the original record (e.g., “Peromyscus sp.”, “Trifolium subterraneum”, “Anser anser × Branta canadensis”, “Aspergillum flavus”, “Pachyporidae”, “Vicia faba minor”, etc.), to be labeled using the DwC term “verbatimIdentification” (see description below). Further taxonomic information of the specimen is desirable and could be described using the DwC terms belonging to the “Taxon” class/extension (see below, **section 10.1.7**).

A field in the dataset matching the DwC term “preservationType” is also required (see **10.1.9**). Please report the kind of preservation of each specimen synthetically, e.g., dried, silica, alcohol, FTA card, tube, QIAsafe, etc. Multiple preservation types must be concatenated within the same cell using a vertical bar (“|”), e.g., “paper bags | sealed cans”. Preservation temperature should be included in a field apart, which should be labeled with the DwC term “preservationTemperature”.

A field containing information on the preparation of each specimen is also required and should be labeled with the DwC term “preparationMaterials”. Materials and chemicals used to prepare the specimen, tissue, DNA, or RNA sample should be included here, e.g., for DNA: DNeasy blood and tissue kit, CTAB, etc. If the specimen did not undergo any preparation process, please use the string “none”.

All the other terms are optional, though it is strongly encouraged to include as much information in the file (dataset) as possible, to be labeled with any of the 72 suitable terms listed below, trying to strictly conform to the format of the values to the description provided for each term. Particularly relevant is the information regarding:

- the collection site (country, locality, minimumElevationInMeters, decimalLatitude, decimalLongitude, etc.), which should be labeled with the suitable DwC terms (see **10.1.5**);
- the reference(s) (if any) associated with each accession, which should be provided either as paper DOI(s) (wherever available) and labeled by the DwC term “associatedReferences” or as full text reference(s); multiple references for the same specimen should be concatenated in each cell and separated by “|”;
- the DNA sequences available for each accession (if any), which should be provided as (concatenated “|”) URL(s) of the remote repositories where the sequence(s) is/are deposited (e.g., ”http://www.ncbi.nlm.nih.gov/nuccore/U34853.1 | http://www.ncbi.nlm.nih.gov/nuccore/GU328060“) and labeled by the DwC term “associatedSequences”;
- the URL(s)/DOI(s) of images for each specimen (if any) retrievable from the internet (concatenated by “|”), which should be labeled by the DwC term “associatedMedia”;
- for pathogens, pests, or symbiotic organisms, a list (concatenated and separated by “|”) of identifiers or names of taxa and the associations of this to each of them, in the form 'association type':'associated organism' (e.g., 'host':'Quercus alba','host': 'gbif.org/species/2879737','parasitoid of':'Cyclocephala signaticollis' | 'predator of':'Apis mellifera', etc.), to be labeled by the DwC term “associatedTaxa”. Please refer to the following URL for the definition of association types: http://purl.obolibrary.org/obo/IAO_0000589. As an alternative, the DwC term “specific_host” of the “DNA data” extension could be used (see the full list of DwC terms).
- the name, surname, and institution label of the person who identified (“identifiedBy”), recorded (“recordedBy”), or prepared (“preparedBy) the specimen/accession. Multiple persons must be separated by a vertical bar (“|”). If the ORCID of each person is available, it should preferably be included in the suitable field (“identifiedByID”, “recordedByID”, etc.).

The list of 72 recommended DwC terms that can effectively convey most NSC specimen information is reported below. This list is deemed large and detailed enough to cover all the different as-

pects of the accessions required for inclusion in the data repository. However, more DwC terms are available at the following URLs: https://dwc.tdwg.org/terms/ and https://rs.gbif.org/extensions.html.

10.1.1 Occurrence	
1. catalogNumber (MANDATORY FIELD)	
DwC class	Occurrence
Description	A numeric identifier (unique for each collection) for the record within the collection. Ex: "145732", "145732", "2008", "4313"
Expected values	Integer (min: 1), String (min chars: 2; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#catalogNumber
2. otherCatalogNumbers	
DwC class	Occurrence
Description	A list (concatenated and separated by "|") of previous or alternate fully qualified catalog numbers or other human-used identifiers for the same Occurrence, whether in the current or any other data set or collection. Ex: "FMNH_Mammal_1234" or "NPS YELLO6778 | MBG 33424"
Expected values	Text (min chars: 0; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#otherCatalogNumbers
3. occurrenceID (MANDATORY FIELD)	
DwC class	Occurrence
Description	An identifier for the Occurrence (as opposed to a particular digital record of the occurrence). In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the occurrenceID globally unique. Recommended best practice is to use a persistent, globally unique identifier. Ex: "http://arctos.database.museum/guid/MSB:Mamm:233627", "000866d2-c177-4648-a200-ead4007051b9", "urn:catalog:UWBM:Bird:89776"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#occurrenceID
4. recordNumber	
DwC class	Occurrence
Description	An identifier given to the Occurrence at the time it was recorded. Often serves as a link between field notes and an Occurrence record, such as a specimen collector's number. Ex: "OPP 7101"
Expected values	Text (min chars: 0; max chars: 80)

Flags	simple
More Info	https://dwc.tdwg.org/terms/#recordNumber
5. recordedBy	
DwC class	Occurrence
Description	A list (concatenated and separated by "|") of names of people, groups, or organizations responsible for recording the original Occurrence. The primary collector or observer, especially one who applies a personal identifier (recordNumber), should be listed first. Ex: "José E. Crespo | Oliver P. Pearson | Anita K. Pearson" (where the value in recordNumber "OPP 7101" corresponds to the collector number for the specimen in the field catalog of Oliver P. Pearson).
Expected values	Text (min chars: 0; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#recordedBy
6. individualCount	
DwC class	Occurrence
Description	The number of individuals present at the time of the Occurrence. Ex: "0", "1", "25"
Expected values	Integer (min: 1)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#individualCount
7. organismQuantity	
DwC class	Occurrence
Description	A number or enumeration value for the quantity of organisms.An organismQuantity must have a corresponding organismQuantityType. Ex: "27" (organismQuantity) with "individuals" (organismQuantityType). "12.5" (organismQuantity) with "% biomass" (organismQuantityType). "r" (organismQuantity) with "Braun Blanquet Scale" (organismQuantityType). "many" (organismQuantity) with "individuals" (organismQuantityType).
Expected values	Decimal (min chars: 1)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#organismQuantity
8. organismQuantityType	
DwC class	Occurrence
Description	The type of quantification system used for the quantity of organisms. A "organismQuantityType" must have a corresponding "organismQuantity". Ex: "27" (organismQuantity) with "individuals" (organismQuantityType). "12.5" (organismQuantity).
Expected values	Any of the following: percentageOfSpecies, percentageOfBiovolume, percentageOfBiomass, percentageCoverage, individuals, domainScale, braunlalnquetScale, biomassAFDG, biomassG, biomassKg, biovolumeCubicMicrons,

	biovolumeMl
Flags	simple
More Info	https://dwc.tdwg.org/terms/#organismQuantityType
9. sex	
DwC class	Occurrence
Description	The sex of the biological individual(s) represented in the Occurrence. Ex: Recommended best practice is to use a controlled vocabulary. Ex: "female", "male", "hermaphrodite"
Expected values	Any of the following: female, male, hermaphrodite, na
Flags	simple
More Info	https://dwc.tdwg.org/terms/#sex
10. lifeStage	
DwC class	Occurrence
Description	The age class or life stage of the Organism(s) at the time the Occurrence was recorded. Recommended best practice is to use a controlled vocabulary. Ex: "zygote", "larva", "juvenile", "adult", "seedling", "flowering", "fruiting"
Expected values	Any of the following: zygote, embryo/seed, larva, juvenile/seedling, adult, sporophyte, spore, gametophyte/pollen, gamete/sperm, pupa
Flags	simple
More Info	https://dwc.tdwg.org/terms/#lifeStage
11. reproductiveCondition	
DwC class	Occurrence
Description	The reproductive condition of the biological individual(s) represented in the Occurrence. Recommended best practice is to use a controlled vocabulary. Ex: "non-reproductive", "pregnant", "in bloom", "fruit-bearing"
Expected values	Text (min chars: 0; max chars: 255)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#reproductiveCondition
12. establishmentMeans	
DwC class	Occurrence
Description	Statement about whether an organism or organisms have been introduced to a given place and time through the direct or indirect activity of modern humans. Ex: Recommended best practice is to use controlled value strings from the controlled vocabulary designated for use with this term, listed at http://rs.tdwg.org/dwc/doc/em/. For details, refer to https://doi.org/10.3897/biss.3.38084. Ex: "native", "nativeReintroduced", "introduced", "introducedAssistedColonisation", "vagrant", "uncertain"
Expected values	Any of the following: native, nativReintroduced, introduced, introducedAssistedColonisation, vagrant, uncertain

Flags	simple
More Info	https://dwc.tdwg.org/terms/#establishmentMeans
13. degreeOfEstablishment	
DwC class	Occurrence
Description	The degree to which an Organism survives, reproduces, and expands its range at the given place and time. Ex: Recommended best practice is to use controlled value strings from the controlled vocabulary designated for use with this term, listed at http://rs.tdwg.org/dwc/doc/doe/. For details, refer to https://doi.org/10.3897/biss.3.38084 Ex: "native", "captive", "cultivated", "released", "failing", "casual", "reproducing", "established", "colonising", "invasive", "widespreadInvasive"
Expected values	Any of the following: native, captive, cultivated, released, failing, casual, reproducing, established, colonising, invasive, widespreadInvasive
Flags	simple
More Info	https://dwc.tdwg.org/terms/#degreeOfEstablishment
14. preparations	
DwC class	Occurrence
Description	A list (concatenated and separated by "|") of preparations and preservation methods for a specimen. Ex: "fossil", "cast", "photograph", "DNA extract", "skin | skull | skeleton", "whole animal (ETOH) | tissue (EDTA)"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#preparations
15. disposition	
DwC class	Occurrence
Description	The current state of a specimen with respect to the collection identified in collectionCode or collectionID. Recommended best practice is to use a controlled vocabulary. Ex: "in collection", "missing", "voucher elsewhere", "duplicates elsewhere"
Expected values	Any of the following: in collection, missing, voucher elsewhere, duplicates elsewhere
Flags	simple
More Info	https://dwc.tdwg.org/terms/#disposition
16. associatedMedia	
DwC class	Occurrence
Description	A list (concatenated and separated by "|") of identifiers (publication, global unique identifier, URI) of media associated with the Occurrence. Ex: "https://arctos.database.museum/media/10520962 | https://arctos.database.museum/media/10520964"
Expected values	Text (no length limits)

Flags	simple
More Info	https://dwc.tdwg.org/terms/#associatedMedia
17. associatedOccurrences	
DwC class	Occurrence
Description	A list (concatenated and separated by "|") of identifiers of other Occurrence records and their associations to this Occurrence. Ex: This term can be used to provide a list of associations to other Occurrences. Note that the ResourceRelationship class is an alternative means of representing associations, and with more detail. Ex: 'parasite collected from':'https://arctos.database.museum/guid/MSB:Mamm:215895?seid=950760'
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#associatedOccurrences
18. associatedReferences	
DwC class	Occurrence
Description	A list (concatenated and separated by "|") of identifiers (publication, bibliographic reference, global unique identifier, URI) of literature associated with the Occurrence. Note that the ResourceRelationship class is an alternative means of representing associations, and with more detail. Note also that the intended usage of the term "references" in Darwin Core when applied to an Occurrence is to point to the definitive source representation of that Occurrence if one is available. Note also that the intended usage of bibliographicCitation in Darwin Core when applied to an Occurrence is to provide the preferred way to cite the Occurrence itself. Ex: "http://www.sciencemag.org/cgi/content/abstract/322/5899/261", "Christopher J. Conroy, Jennifer L. Neuwald. 2008. Phylogeographic study of the California vole, Microtus californicus Journal of Mammalogy, 89(3) 755-767."
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#associatedReferences
19. associatedSequences	
DwC class	Occurrence
Description	A list (concatenated and separated) of identifiers (publication, global unique identifier, URI) of genetic sequence information associated with the Occurrence. Ex: "http://www.ncbi.nlm.nih.gov/nuccore/U34853.1", "http://www.ncbi.nlm.nih.gov/nuccore/GU328060 | http://www.ncbi.nlm.nih.gov/nuccore/AF326093"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#associatedSequences
20. associatedTaxa	

DwC class	Occurrence
Description	A list (concatenated and separated by "|") of identifiers or names of taxa and the associations of this Occurrence to each of them. This term can be used to provide a list of associations to Taxa other than the one defined in the Occurrence. Note that the ResourceRelationship class is an alternative means of representing associations, and with more detail. This term is not apt for establishing relationships between Taxa, only between specific Occurrences of an Organism with other Taxa. Ex: 'host':'Quercus alba', 'host':'gbif.org/species/2879737','parasitoid of':'Cyclocephala signaticollis' | 'predator of':'Apis mellifera'
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#associatedTaxa
21. occurrenceRemarks	
DwC class	Occurrence
Description	Comments or notes about the Occurrence. Ex: "found dead on road"
Expected values	Text (min chars: 0; max chars: 255)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#occurrenceRemarks
10.1.2 Organism	
22. organismID	
DwC class	Organism
Description	An identifier for the Organism instance (as opposed to a particular digital record of the Organism). May be a globally unique identifier or an identifier specific to the data set. Ex: "http://arctos.database.museum/guid/WNMU:Mamm:1249"
Expected values	Text (min chars: 0; max chars: 255)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#organismID
23. organismScope	
DwC class	Organism
Description	A description of the kind of Organism instance. Can be used to indicate whether the Organism instance represents a discrete organism or if it represents a particular type of aggregation. Recommended best practice is to use a controlled vocabulary. This term is not intended to be used to specify a type of taxon. To describe the kind of "Organism" using a URI object in RDF, use rdf-type (http://www.w3.org/1999/02/22-rdf-syntax-ns#type) instead. Ex: "multicellular organism", "virus", "clone", "pack", "colony"
Expected values	Any of the following: multicellular organism, virus, clone, pack, colony

Flags	simple
More Info	https://dwc.tdwg.org/terms/#organismScope
24. associatedOrganisms	
DwC class	Organism
Description	A list (concatenated and separated by "|") of identifiers of other Organisms and the associations of this Organism to each of them. This term can be used to provide a list of associations to other Organisms. Note that the ResourceRelationship class is an alternative means of representing associations, and with more detail. Ex: 'parent of':'http://arctos.database.museum/guid/MSB_Mamm_196523' | 'sibling of':'http://arctos.database.museum/guid/MSB_Mamm_142638'
Expected values	Text (min chars: 0; max chars: 255)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#associatedOrganisms
25. previousIdentifications	
DwC class	Organism
Description	A list (concatenated and separated) of previous assignments of names to the Organism. Ex: Recommended best practice is to separate the values in a list with a vertical bar ("|"). Ex: "Chalepidae", "Pinus abies", "Anthus sp., field ID by G. Iglesias | Anthus correndera, expert ID by C. Cicero 2009-02-12 based on morphology"
Expected values	Text (min chars: 0; max chars: 255)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#previousIdentifications
26. organismRemarks	
DwC class	Organism
Description	Comments or notes about the Organism instance. Ex: "One of a litter of six"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#organismRemarks
10.1.3 MaterialSample	
27. materialSampleID	
DwC class	MaterialSample
Description	An identifier for the MaterialSample (as opposed to a particular digital record of the material sample). In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the materialSampleID globally unique. Recommended best practice is to use a persistent, globally unique identifier. Ex: "06809dc5-f143-459a-

	be1a-6f03e63fc083"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#materialSampleID
10.1.4 Event	
28. fieldNumber	
DwC class	Event
Description	An identifier given to the event in the field. Often serves as a link between field notes and the Event. Ex: "RV Sol 87-03-08"
Expected values	Text (min chars: 0; max chars: 80)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#fieldNumber
29. eventDate	
DwC class	Event
Description	The date-time or interval during which an Event occurred. For occurrences, this is the date-time when the event was recorded. Not suitable for a time in a geological context. Recommended best practice is to use a date that conforms to ISO 8601-1:2019. Ex: "1963-03-08T14:07-0600" (8 Mar 1963 at 2:07pm in the time zone six hours earlier than UTC)..
Expected values	Date
Flags	simple
More Info	https://dwc.tdwg.org/terms/#eventDate
30. habitat	
DwC class	Event
Description	A category or description of the habitat in which the Event occurred. Ex: "oak savanna", "pre-cordilleran steppe"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#habitat
31. samplingProtocol	
DwC class	Event
Description	The names of, references to, or descriptions of the methods or protocols used during an Event. Recommended best practice is describe an Event with no more than one sampling protocol. In the case of a summary Event with multiple protocols, in which a specific protocol can not be attributed to specific Occurrences, the recommended best practice is to separate the values in a list with space vertical bar space (" |"). Ex: "UV light trap", "mist net", "bottom

	trawl", "ad hoc observation | point count", "Penguins from space: faecal stains reveal the location of emperor penguin colonies, https://doi.org/10.1111/j.1466-8238.2009.00467.x", "Takats et al. 2001. Guidelines for Nocturnal Owl Monitoring in North America. Beaverhill Bird Observatory and Bird Studies Canada, Edmonton, Alberta. 32 pp., http://www.bsc-eoc.org/download/Owl.pdf"
Expected values	Text (min chars: 0; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#samplingProtocol
32. fieldNotes	
DwC class	Event
Description	One of a) an indicator of the existence of, b) a reference to (publication, URI), or c) the text of notes taken in the field about the Event. Ex: "Notes available in the Grinnell-Miller Library."
Expected values	Text (min chars: 0; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#fieldNotes
10.1.5 Location	
33. country	
DwC class	Location
Description	The name of the country or major administrative unit in which the Location occurs. Ex: Recommended best practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names. Recommended best practice is to leave this field blank if the Location spans multiple entities at this administrative level or if the Location might be in one or another of multiple possible entities at this level. Multiplicity and uncertainty of the geographic entity can be captured either in the term higherGeography or in the term locality, or both. Ex: "Denmark", "Colombia", "Espana"
Expected values	Text (min chars: 1; max chars: 255)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#country
34. countryCode	
DwC class	Location
Description	The standard code for the country in which the Location occurs. Ex: Recommended best practice is to use an ISO 3166-1-alpha-2 country code. Ex: "AR", "SV"
Expected values	Any of the following:
Flags	simple

More Info	https://dwc.tdwg.org/terms/#countryCode
35. locality	
DwC class	Location
Description	The specific description of the place. Less specific geographic information can be provided in other geographic terms (higherGeography, continent, country, stateProvince, county, municipality, waterBody, island, islandGroup). This term may contain information modified from the original to correct perceived errors or standardize the description. Ex: "Bariloche, 25 km NNE via Ruta Nacional 40 (Ruta 237)", "Queets Rainforest, Olympic National Park"
Expected values	Text (min chars: 0; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#locality
36. verbatimLocality	
DwC class	Location
Description	The original textual description of the place. Ex: "25 km NNE Bariloche por R. Nac. 237"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#verbatimLocality
37. minimumElevationInMeters	
DwC class	Location
Description	The lower limit of the range of elevation (altitude, usually above sea level), in meters. Ex: "-100", "802"
Expected values	Integer (min: -11000; max: 8848)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#minimumElevationInMeters
38. locationRemarks	
DwC class	Location
Description	Comments or notes about the Location. Ex: "under water since 2005"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#locationRemarks
39. decimalLatitude	
DwC class	Location
Description	The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive. Ex: "-41.0983423"

Expected values	Decimal (min chars: -90; max chars: 90)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#decimalLatitude
40. decimalLongitude	
DwC class	Location
Description	The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive. Ex: "-121.1761111"
Expected values	Decimal (min chars: -180; max chars: 180)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#decimalLongitude
41. verbatimCoordinates	
DwC class	Location
Description	The verbatim original spatial coordinates of the Location. The coordinate ellipsoid, geodeticDatum, or full Spatial Reference System (SRS) for these coordinates should be stored in verbatimSRS and the coordinate system should be stored in verbatimCoordinateSystem. Ex: "41 05 54S 121 05 34W", "17T 630000 4833400"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#verbatimCoordinates
42. georeferencedDate	
DwC class	Location
Description	The date on which the Location was georeferenced. Recommended best practice is to use a date that conforms to ISO 8601-1:2019. Ex: "1963-03-08".
Expected values	Date
Flags	simple
More Info	https://dwc.tdwg.org/terms/#georeferencedDate
10.1.6 Identification	
43. identificationID	
DwC class	Identification
Description	An identifier for the Identification (the body of information associated with the assignment of a scientific name). May be a global unique identifier or an identifier specific to the data set. Ex: "9992"
Expected values	Integer

Flags	simple
More Info	https://dwc.tdwg.org/terms/#identificationID
44. verbatimIdentification (MANDATORY FIELD)	
DwC class	Identification
Description	A string representing the taxonomic identification as it appeared in the original record. This term is meant to allow the capture of an unaltered original identification/determination, including identification qualifiers, hybrid formulas, uncertainties, etc. This term is meant to be used in addition to "scientificName" (and "identificationQualifier" etc.), not instead of it. Ex: "Peromyscus sp.", "Ministrymon sp.", "Anser anser x Branta canadensis", "Pachyporidae"
Expected values	Text (min chars: 3; max chars: 255)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#verbatimIdentification
45. identifiedBy	
DwC class	Identification
Description	A list (concatenated and separated "|") of names of people, groups, or organizations who assigned the Taxon to the subject. Ex: "James L. Patton", "Theodore Pappenfuss | Robert Macey"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#identifiedBy
10.1.7 Taxon	
46. scientificName	
DwC class	Taxon
Description	The full scientific name, with authorship and date information if known. When forming part of an Identification, this should be the name in lowest level taxonomic rank that can be determined. This term should not contain identification qualifications, which should instead be supplied in the IdentificationQualifier term. When applied to an Organism or Occurrence, this term should be used to represent the scientific name that was applied to the associated Organism in accordance with the Taxon to which it was or is currently identified. Ex: "Coleoptera" (order). "Vespertilionidae" (family). "Manis" (genus). "Ctenomys sociabilis" (genus + specificEpithet). "Ambystoma tigrinum diaboli" (genus + specificEpithet + infraspecificEpithet). "Roptrocerus typographi (Gyarfi, 1952)" (genus + specificEpithet + scientificNameAuthorship), "Quercus agrifolia var. oxyadenia (Torr.) J.T. Howell" (genus + specificEpithet + taxonRank + infraspecificEpithet + scientificNameAuthorship).
Expected values	Text (min chars: 0; max chars: 254)
Flags	simple

More Info	https://dwc.tdwg.org/terms/#scientificName
47. genus (MANDATORY FIELD)	
DwC class	Taxon
Description	The full scientific name of the genus in which the taxon is classified. Ex: "Puma", "Monoclea"
Expected values	Text (min chars: 3; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#genus
48. specificEpithet	
DwC class	Taxon
Description	The name of the first or species epithet of the scientificName. Ex: "concolor", "gottschei"
Expected values	Text (min chars: 0; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#specificEpithet
49. infraspecificEpithet	
DwC class	Taxon
Description	The name of the lowest or terminal infraspecific epithet of the scientificName, excluding any rank designation. In botany, where there can be more than one infraspecific rank, name strings may be provided, in literature and in identifications, that have more than two epithets. Only the last of these epithets is the infraspecificEpithet and only the first and the last epithets belong to the scientificName. For example: the infraspecificEpithet in the string "Indigofera charlieriana subsp. sessilis var. scaberrima" is "scaberrima" and the scientificName is "Indigophera charlieriana var. scaberrima". Ex: "concolor" (for scientificName "Puma concolor concolor"), "oxyadenia" (for scientificName "Quercus agrifolia var. oxyadenia"), "laxa" (for scientificName "Cheilanthes hirta f. laxa"), "scaberrima" (for scientificName "Indigofera charlieriana var. scaberrima").
Expected values	Text (min chars: 0; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#infraspecificEpithet
50. taxonRank	
DwC class	Taxon
Description	The taxonomic rank of the most specific name in the scientificName. Ex: Recommended best practice is to use a controlled vocabulary. Ex: "subspecies", "varietas", "forma", "species", "genus"
Expected values	Any of the following: , subspecies, infrasubspecificname, variety, subvariety, form, subform, pathovar, biovar, chemovar, morphovar, phagovar, serovar, chemoform, formaspecialis, cultivarGroup, cultivar, strain, informal, unranked

Flags	simple
More Info	https://dwc.tdwg.org/terms/#taxonRank
51. verbatimTaxonRank	
DwC class	Taxon
Description	The taxonomic rank of the most specific name in the scientificName as it appears in the original record. Ex: "Agamospecies", "sub-lesus", "prole", "apomict", "nothogrex", "sp.", "subsp.", "var."
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#verbatimTaxonRank
52. scientificNameAuthorship	
DwC class	Taxon
Description	The authorship information for the scientificName formatted according to the conventions of the applicable nomenclaturalCode. Ex: "(Torr.) J.T. Howell", "(Martinovsky) Tzvelev", "(Gyarfi, 1952)"
Expected values	Text (min chars: 0; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#scientificNameAuthorship
53. vernacularName	
DwC class	Taxon
Description	A common or vernacular name. Ex: "Andean Condor", "Condor Andino", "American Eagle", "Gà¤nsegeier"
Expected values	Text (min chars: 0; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#vernacularName
54. taxonomicStatus	
DwC class	Taxon
Description	The status of the use of the scientificName as a label for a taxon. Requires taxonomic opinion to define the scope of a taxon. Rules of priority then are used to define the taxonomic status of the nomenclature contained in that scope, combined with the experts opinion. It must be linked to a specific taxonomic reference that defines the concept. Recommended best practice is to use a controlled vocabulary. Ex: "invalid", "misapplied", "homotypic synonym", "accepted"
Expected values	Any of the following: accepted, doubtful, synonym, heterotypicSynonym, homotypicSynonym, proParteSynonym, misapplied
Flags	simple
More Info	https://dwc.tdwg.org/terms/#taxonomicStatus

55. taxonRemarks	
DwC class	Taxon
Description	Comments or notes about the taxon or name. Ex: "this name is a misspelling in common use"
Expected values	Text (min chars: 0; max chars: 254)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#taxonRemarks

10.1.8 DNA data

56. ploidy	
DwC class	DNA data
Description	The ploidy level of the genome (e.g., allopolyploid, haploid, diploid, triploid, tetraploid, etc.). It has implications for the downstream study of duplicated genes and regions of the genomes (and perhaps for difficulties in assembly). For terms, please select terms listed under class ploidy (PATO:001374) of Phenotypic Quality Ontology (PATO), and for a browser of PATO (v 2018-03-27) please refer to http://purl.bioontology.org/ontology/PATO. Ex: Aneuploid, Haploid, Diploid, Triploid, Tetraploid, Polyploid, Allopolyploid
Expected values	Any of the following: Aneuploid, Haploid, Diploid, Triploid, Tetraploid, Polyploid, Allopolyploid
Flags	extension
More Info	https://w3id.org/gensc/terms/MIXS:0000021
57. pathogenicity	
DwC class	DNA data
Description	To what is the entity pathogenic The names of organisms that the entity is pathogenic to. Ex: human, animal, plant, fungi, bacteria, etc.
Expected values	Text (min chars: 0; max chars: 255)
Flags	extension
More Info	https://w3id.org/gensc/terms/MIXS:0000027
58. biotic_relationship	
DwC class	DNA data
Description	Description of relationship(s) between the subject organism and other organism(s) with which it is associated. E.g., parasite on species X; mutualist with species Y. The target organism is the subject of the relationship, and the other organism(s) is the object. Ex: https://rs.gbif.org/vocabulary/mixs/biotic_relationship.xml
Expected values	Any of the following: free living, parasitism, commensalism, symbiotic, mutualism
Flags	extension

More Info	https://w3id.org/gensc/terms/MIXS:0000028
59. host_disease_stat	
DwC class	DNA data
Description	List of diseases with which the host has been diagnosed; it can include multiple diagnoses. The value of the field depends on the host; for humans, the terms should be chosen from the DO (Human Disease Ontology) at https://www.disease-ontology.org, non-human host diseases are free text. Ex: https://w3id.org/gensc/terms/MIXS:0000031
Expected values	Text (no length limits)
Flags	extension
More Info	https://w3id.org/gensc/terms/MIXS:0000031
60. trophic_level	
DwC class	DNA data
Description	Trophic levels are the feeding positions in a food chain. Microbes can be a range of producers (e.g. chemolithotroph). Ex: https://w3id.org/gensc/terms/MIXS:0000032, https://rs.gbif.org/vocabulary/mixs/trophic_level.xml
Expected values	Any of the following: autotroph, carboxydotroph, chemoautotroph, chemoheterotroph, chemolithoautotroph, chemolithotroph, chemoorganoheterotroph, chemoorganotroph, chemosynthetic, chemotroph, copiotroph, diazotroph, facultative, autotroph, heterotroph, lithoautotroph, lithoheterotroph, lithotroph, methanotroph, methylotroph, mixotroph , obligate, chemoautolithotroph, oligotroph, organoheterotroph, organotroph, photoautotroph, photoheterotroph, photolithoautotroph, photolithotroph, photosynthetic, phototroph
Flags	extension
More Info	https://w3id.org/gensc/terms/MIXS:0000032
61. propagation	
DwC class	DNA data
Description	This field is specific to different taxa. For phages: lytic/lysogenic, for plasmids: incompatibility group, for eukaryotes: sexual/asexual (Note: there is the strong opinion to name phage propagation obligately lytic or temperate, therefore we also give this choice Ex: https://w3id.org/gensc/terms/MIXS:0000033
Expected values	Text (no length limits)
Flags	extension
More Info	https://w3id.org/gensc/terms/MIXS:0000033
62. rel_to_oxygen	
DwC class	DNA data
Description	Is this organism an aerobe or an anaerobe? Please note that aerobic and anaerobic are valid descriptors for microbial environments. Ex: https://w3id.org/gensc/terms/MIXS:0000015. https://rs.gbif.org/vocabulary/mixs/rel_to_oxygen.xml

Expected values	Any of the following: aerobe, anaerobe, facultative, microaerophilic, microanaerobe, obligate aerobe, obligate anaerobe
Flags	extension
More Info	https://w3id.org/gensc/terms/MIXS:0000015

10.1.9 Preparation

63. preservationType (MANDATORY FIELD)	
DwC class	Preparation
Description	Type of Specimen, Tissue, DNA or DNA Preservation or Storage. Ex: dried, silica, alcohol, FTA card, tube, QIAsafe, etc.
Expected values	Text (min chars: 4)
Flags	extension
More Info	https://dwc.tdwg.org/terms/#preservationType
64. preservationTemperature	
DwC class	Preparation
Description	Temperature of Specimen, Tissue, DNA or RNA Preservation or Storage. Ex: -20°C, RT, -80°C, -196°C, LN
Expected values	Text (min chars: 3)
Flags	extension
More Info	https://dwc.tdwg.org/terms/#preservationTemperature
65. preservationDateBegin	
DwC class	Preparation
Description	Start of current specimen, tissue, DNA or RNA Preservation. Ex: 2014-02-19
Expected values	Date
Flags	extension
More Info	https://dwc.tdwg.org/terms/#preservationDateBegin
66. preparationType	
DwC class	Preparation
Description	Description of preparation type (specimens, tissues, DNA, HTS Libraries). Ex: for DNA: gDNA, eDNA, aDNA; for tissues/specimens: leaf, muscle, leg, blood; for HTs libraries: Whole genome shotgun sequencing, Amplicon sequencing, RAD sequencing
Expected values	Text (no length limits)
Flags	extension
More Info	https://dwc.tdwg.org/terms/#preparationType
67. preparationProcess	

DwC class	Preparation
Description	Process used in preparing the specimen or sample, can also be used to describe Phage/Plasmid propagation, Process used in extracting the DNA/RNA; adaptions made; SPREC code
Expected values	Text (no length limits)
Flags	extension
More Info	https://dwc.tdwg.org/terms/#preparationProcess
68. preparationMaterials (MANDATORY FIELD)	
DwC class	Preparation
Description	Materials and chemicals used in the preparation of the specimen, tissue, DNA or RNA sample. Ex: for DNA: DNeasy blood and tissue kit, CTAB. If any, please insert "none".
Expected values	Text (min chars: 4)
Flags	extension
More Info	https://dwc.tdwg.org/terms/#preparationMaterials
69. preparedBy	
DwC class	Preparation
Description	Person and/or institution responsible for or effecting the preparation/extraction. Ex: Jane Doe, NMNH
Expected values	Text (no length limits)
Flags	extension
More Info	https://dwc.tdwg.org/terms/#preparedBy
70. preparationDate	
DwC class	Preparation
Description	The date of preparation/extraction. Ex: 2011-01-18
Expected values	Date
Flags	extension
More Info	https://dwc.tdwg.org/terms/#preparationDate
10.1.10 Record-level	
71. dynamicProperties	
DwC class	Record-level
Description	A list of additional measurements, facts, characteristics, or assertions about the record. Meant to provide a mechanism for structured content. Ex: Recommended best practice is to use a key:value encoding schema for a data interchange format such as JSON. Ex: "{"heightInMeters":1.5}", "{"tragusLengthInMeters":0.014, "weightInGrams":120}", "{"natureOfID":"expert identifica-

	tion", "identificationEvidence":"cytochrome B sequence"}", "{"relativeHumidity":28, "airTemperatureInCelsius":22, "sampleSizeInKilograms":10}", "{"aspectHeading":277, "slopeInDegrees":6}", "{"iucnStatus":"vulnerable", "taxonDistribution":"Neuquàn, Argentina"}"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#dynamicProperties
72. references	
DwC class	Record-level
Description	A related resource that is referenced, cited, or otherwise pointed to by the described resource. This property is intended to be used with non-literal values. This property is an inverse property of Is Referenced By." The intended usage of this term in Darwin Core is to point to the definitive source representation of the resource (e.g.,Taxon, Occurrence, Event in Darwin Core), if one is available. Note that the intended usage of dcterms:bibliographicCitation in Darwin Core, by contrast, is to provide the preferred way to cite the resource itself. Ex: MaterialSample example: "http://arctos.database.museum/guid/MVZ:Mamm:165861", Taxon example: "https://www.catalogueoflife.org/data/taxon/32664"
Expected values	Text (no length limits)
Flags	simple
More Info	https://dwc.tdwg.org/terms/#references

10.2 Appendix 2: Compliance of GeneRAP usage with the FAIR principles.

The central GeneRAP data repository has been developed in full accordance with FAIR principles, making data Findable, Accessible, Interoperable, and Reusable. Here are some examples of how these principles are applied in GeneRAP:

10.2.1 Findable

Principle	Justification
F1. (Meta)data is assigned a globally unique and persistent identifier	GeneRAP assigns unique, persistent identifiers to accessions and related entities (e.g., idAccession, catalogNumber, occurenceID, etc.).
F2. Data are described with rich metadata	Each object (e.g., accessions, datasets, seeds, etc.) is described with detailed attributes like collection source, biological status, conservation status, regulation reference, and more.
F3. Metadata includes the identifier of the data they describe	Metadata explicitly references the object's identifier, such as foreign keys linking to idAccessionSpecies, idAccessionDataset, etc.
F4. (Meta)data are registered or indexed in a searchable resource	GeneRAP includes a web-based interface for search abilities and a /data/search API to search and filter data across multiple categories. This fulfills both human and machine discoverability.

10.2.2 Accessible

Principle	Justification
A1. (Meta)data is retrievable by its identifier using a standardized communication protocol	The system exposes REST APIs over HTTP(S), which is universally supported.
A1.1 Protocol is open, free, and universally implementable	HTTP is a free, widely adopted protocol.
A1.2 Protocol supports authentication and authorization	GeneRap includes user and institution authentication and role-based access (e.g., supervisor, operator), controlling access to data and operations.
A2. Metadata remains accessible even if the data is no longer available	All operations (creations, updates, deletions) are tracked as Submission and Variations tables, providing historical traceability even if the main data is removed.

10.2.3 Interoperable

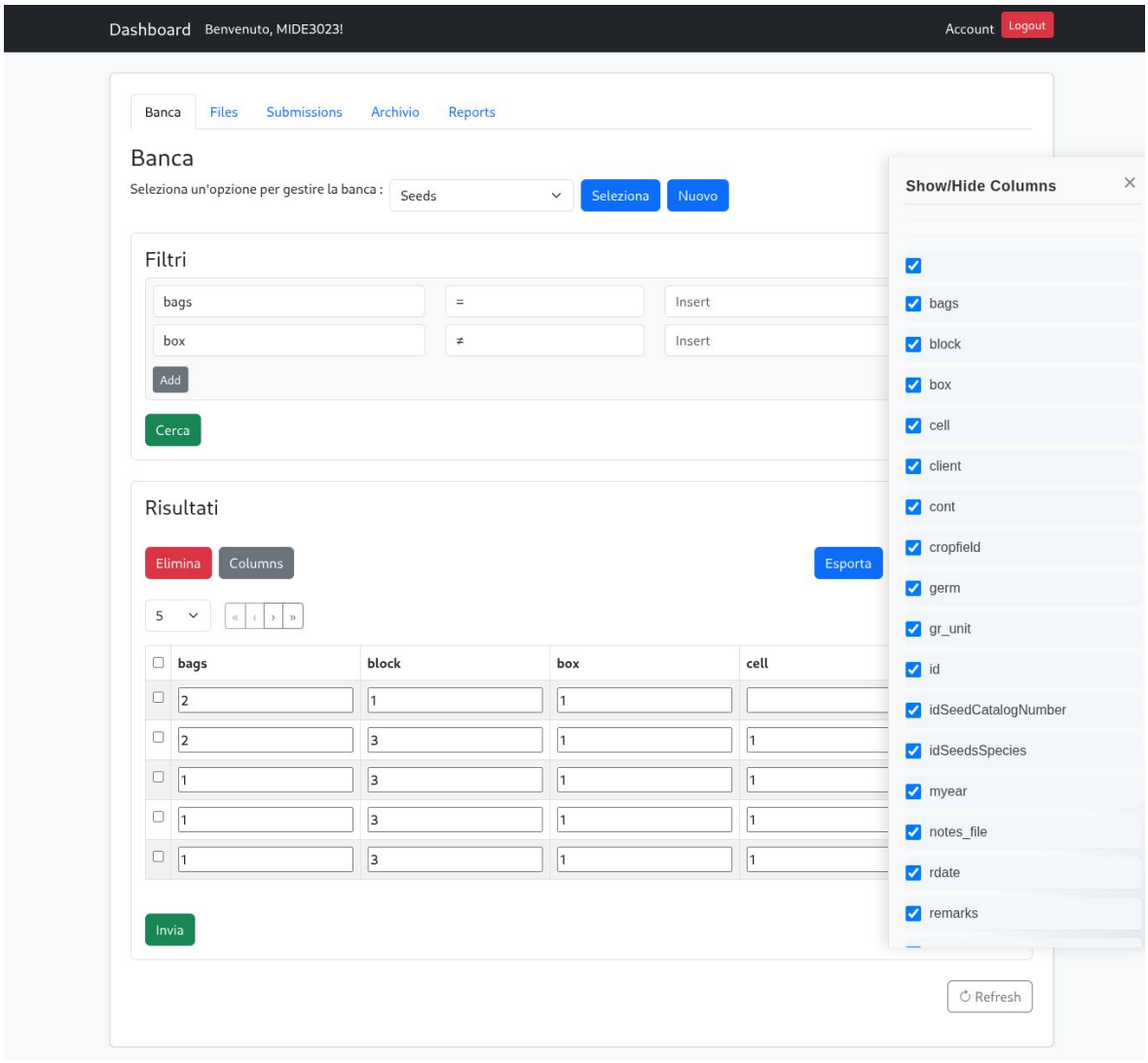
Principle	Justification
I1. (Meta)data uses a formal, accessible, shared, and broadly applicable language	GeneRAP uses standard formats such as JSON for APIs and SQL for data management, with fields following standard patterns (e.g., ISO dates, country codes).
I2. (Meta)data use vocabularies that follow FAIR principles	Controlled vocabularies and Standards such as DarwinCore are used to enhance semantic interoperability.
I3. (Meta)data include qualified references to other (meta)data	A strong use of foreign keys and linked relationships between datasets (e.g., accessions linked to species or datasets) ensures relational interoperability. Many links to other resources are stored and searchable.

10.2.4 Reusable

Principle	Justification
R1. (Meta)data is richly described with accurate and relevant attributes	Data entries are richly annotated with contextual and domain-specific metadata (e.g., collecting institutes, conditions, legal frameworks).
R1.1 (Meta)data are released with a clear and accessible usage license	As explicitly declared, the website will show an open data license (e.g., CC-BY 4.0), which will fully satisfy this criterion.
R1.2 (Meta)data is associated with detailed provenance	Metadata is enriched with records about actions, timestamps, user roles, and detailed descriptions about logical and physical provenance.
R1.3 (Meta)data meets domain-relevant community standards	GeneRAP supports domain-specific identifiers, aligning with common biodiversity and plant genetic resource standards.

10.3 Appendix 3: The MGD Management System in the GeneRAP back office

The MGD-BA Dashboard represents the core of the private interface for authenticated users, providing authorized personnel with advanced functionalities for managing biodiversity data collections in a structured, secure, and user-friendly environment.



10.3.1 Navigation & Access

At the top of the page, a clean navigation bar grants access to all key management areas:

- Banca (Bank): Entry point for accessing and managing collection data.
- Files: Upload and associate relevant documents.
- Submissions: Review and curate pending or approved submissions.
- Archivio: Track historical variations and edits across records.
- Reports: Generate analytical outputs and downloadable summaries.

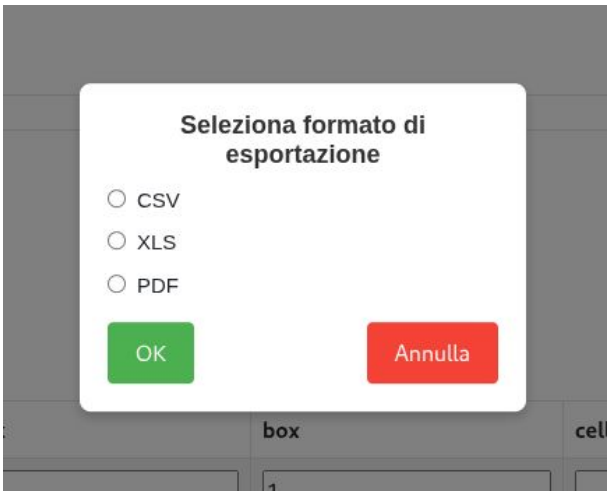
On the right, the interface greets the logged-in user (e.g., MIDE3023) and provides account/logout controls, ensuring secure session management.

10.3.2 Record Management Interface – Bank Section

This section is dedicated to the interactive management of biological records such as Seeds, Accessions, or Collections. The user can select the specific dataset from the dropdown menu and immediately perform operations on it.

Features include:

- Dynamic filtering: Users can add multiple conditional clauses using a simplified logic builder to query the database.
- Pagination and limit selection: A dropdown allows control over the number of rows displayed, enhancing usability for larger datasets.
- Editable fields: Records in the table are editable in-line, promoting quick data corrections or additions.
- Bulk actions:
    - Delete: Remove selected records.
    - Export: Download the current data view for offline processing.
    - Submit: Send modified or newly entered rows for approval.



Column Customization

A slide-in panel titled “Show/Hide Columns” empowers users to tailor the table to their specific task by toggling the visibility of individual columns. This modular view supports workflows ranging from detailed curation to high-level summaries, without clutter.

Security & Access Control

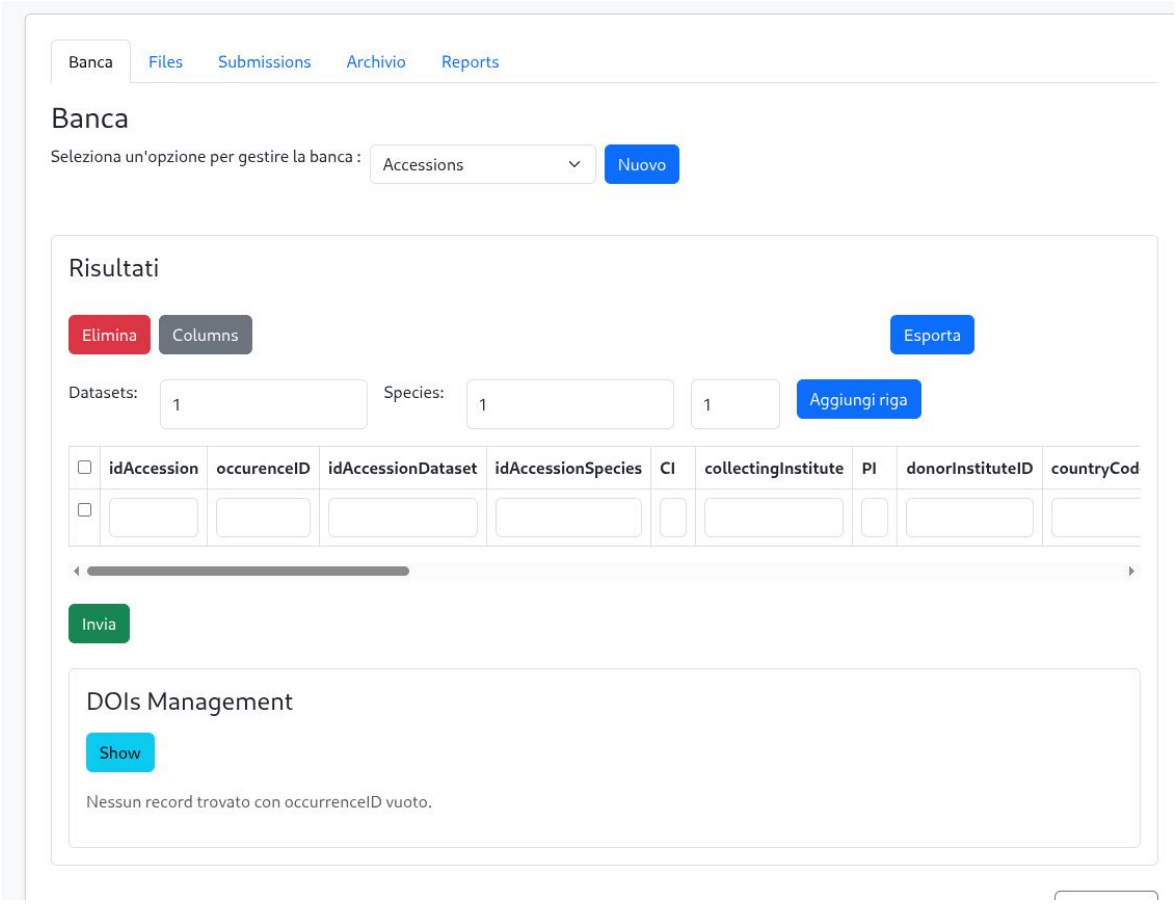
This section of the portal is fully protected behind the login system previously described, relying on institutional credentials and email verification. Once inside, the interface respects role-based permissions, limiting certain operations to supervisors or admins while allowing data entry roles to focus on their contributions.

Underlying Architecture

The Dashboard is powered by a combination of structured backend APIs and frontend dynamic JavaScript (as referenced in dashboard.js). This enables responsive behavior such as:

- Updating filters without reloading the page

- Real-time column visibility adjustments
- Role-aware rendering of actions and controls



Upon accessing the "Banca" section and selecting a category (e.g., Accessions), users can initiate the creation of a new record by clicking the "Nuovo" (New) button.

Interface & Workflow

When a new record is initiated:

- A blank row appears in the editable data table, which dynamically adapts based on the selected category.
- The interface enables direct, inline data entry for each field. This includes typical attributes such as idAccession, occurrenceID, collectingInstitute, donorInstituteID, and more.
- An "Aggiungi riga" (Add row) button allows the user to expand the data entry session by adding multiple records before submission.

Each input cell is fully editable and validated client-side via JavaScript functions in submissions.js, ensuring basic consistency before committing to the database.

Column Management & Dynamic Filters

The column visibility can be tailored through the "Columns" button, which opens a side panel allowing users to show or hide specific fields, depending on the current curation task. This feature is particularly useful when handling datasets with numerous attributes, ensuring focus and efficiency.

Additionally, filters (e.g., for Datasets or Species IDs) can be applied even during new record creation, helping users to contextualize their entry in relation to existing records.

DOI Management Integration

At the bottom of the submission interface is a DOIs Management module. Although in the example no DOI record is found (*Nessun record trovato con occurrenceID vuoto*.), the system is prepared for automatic linking or creation of Digital Object Identifiers for uniquely identifying data entries. The "Show" button likely triggers an API call or interface to retrieve linked DOI metadata for review.

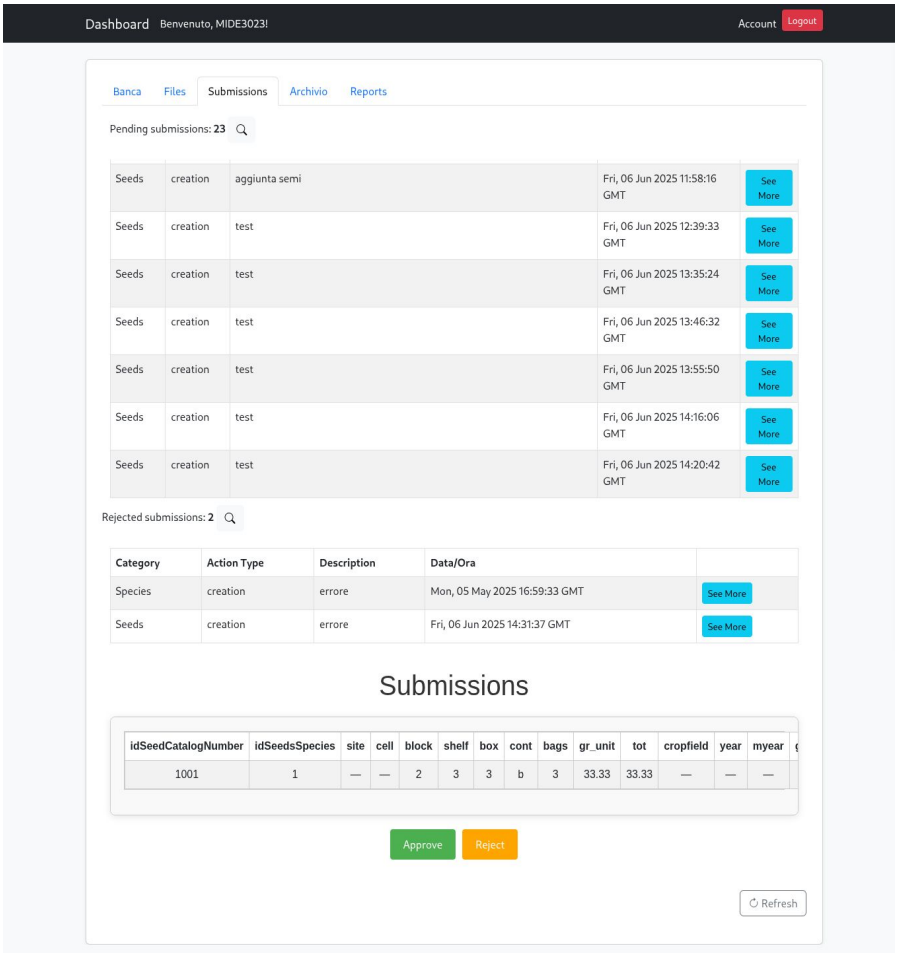
Backend Logic

From a functional perspective, submission files are managed:

- Validation and input sanitation before database submission.
- Asynchronous interaction with backend endpoints for storing or retrieving records.
- Pagination and export (CSV/Excel) functionalities via the "Esporta" button.
- Integration with controlled vocabularies or dynamic selection lists when applicable (e.g., predefined institutes, countries).

10.3.3 Submission

Once all fields are populated, clicking "Invia" (Submit) finalizes the insertion. Feedback from the system confirms success or alerts the user in case of validation failure.



The Submissions section of the MGD application plays a key role in managing the integration of new data. It is designed to provide curators and authorized users with a streamlined way to review, verify, and approve information before it becomes part of the public database.

What is the Submissions section for

When a user - such as a researcher or contributor - submits new data (e.g., a new accession or taxonomic record), that information is not immediately added to the central database. Instead, it is stored in a temporary area known as the buffer, where it waits for review and validation.

From the Submissions interface, users can:

- View all pending records, filtering them by status (e.g., "pending", "approved", "rejected").
- Edit individual data entries directly, fixing errors or completing missing fields.
- Approve or reject records, either individually or in bulk.

The work process

Upon accessing the section, a dynamic table displays all submissions. Each row represents a new dataset awaiting approval. Fields are editable directly in the table, allowing curators to make adjustments on the spot. Once a record is reviewed and confirmed, it can be validated with a simple click.

Reviewers can:

- Ensure that the data is complete and accurate (e.g., correct species names, valid institutional codes).
- Use clear actions like "Approve" or "Reject" depending on the quality of the submission.
- Once approved, records are automatically transferred to the official database.

Quality control and traceability

Each submission keeps a detailed history: who submitted it, who reviewed it, and when the decision was made. This ensures transparency and accountability in the curation process.

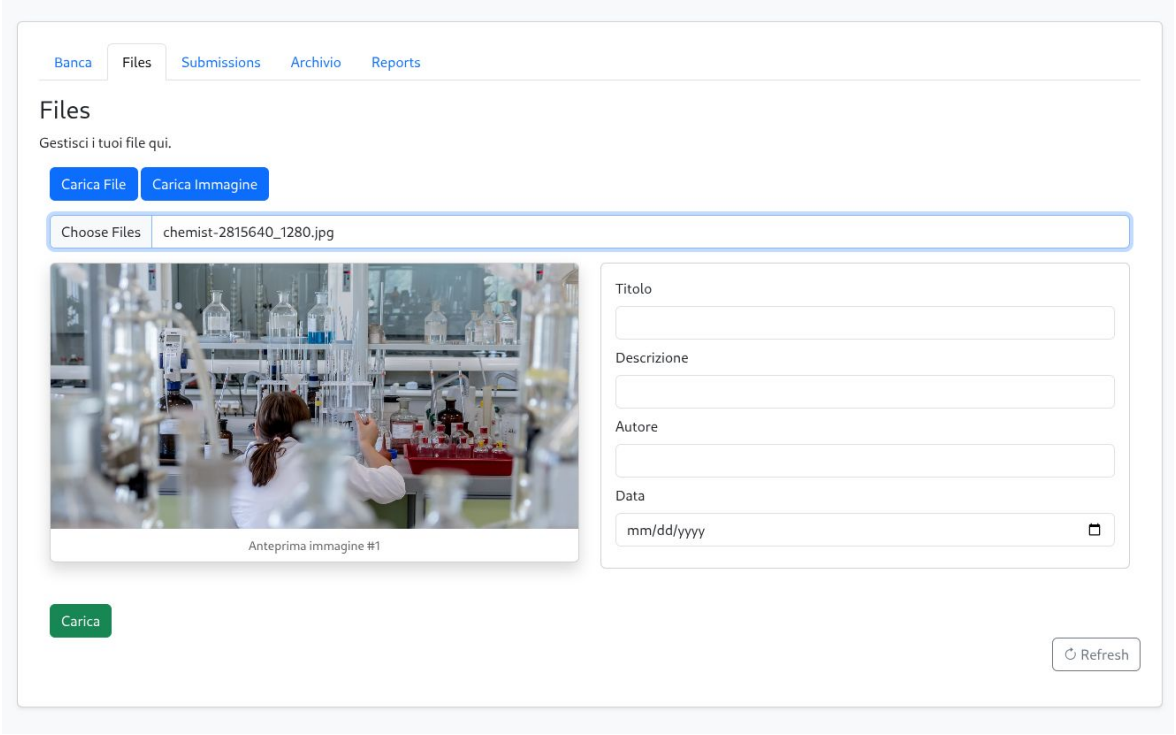
Moreover, the system helps prevent mistakes or inconsistencies by validating key fields (e.g., required inputs, format checks) before any approval is possible.

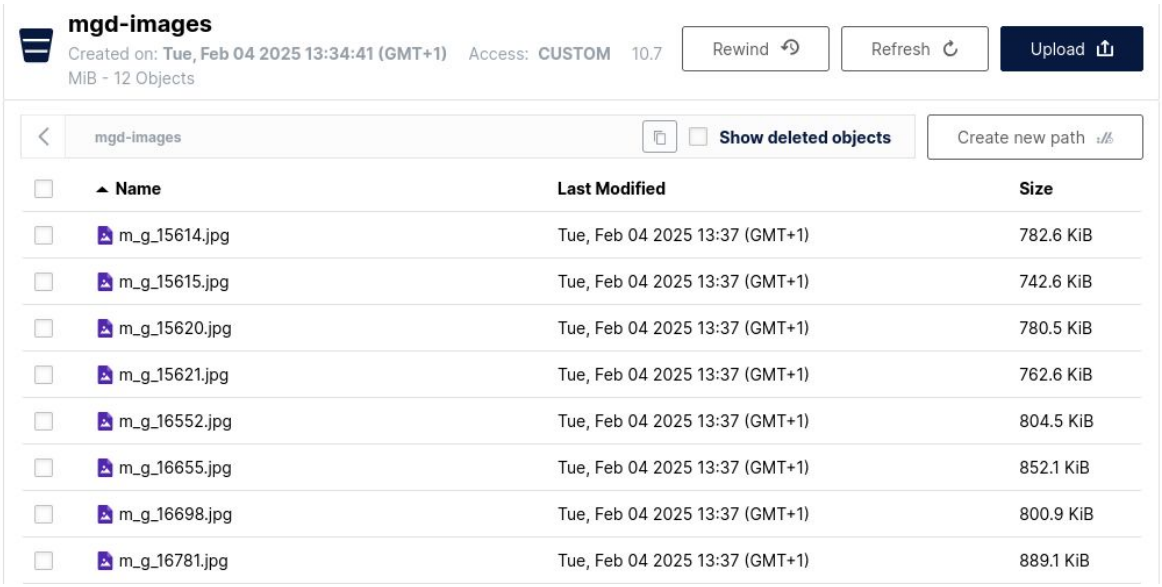
Why it matters

Thanks to this interface, MGD guarantees that only curated and validated data is made public, maintaining a high standard of data integrity.

At the same time, it promotes scientific collaboration, allowing multiple users and institutions to contribute valuable data in a structured and controlled environment, knowing each submission will be carefully reviewed.

10.3.4 Files

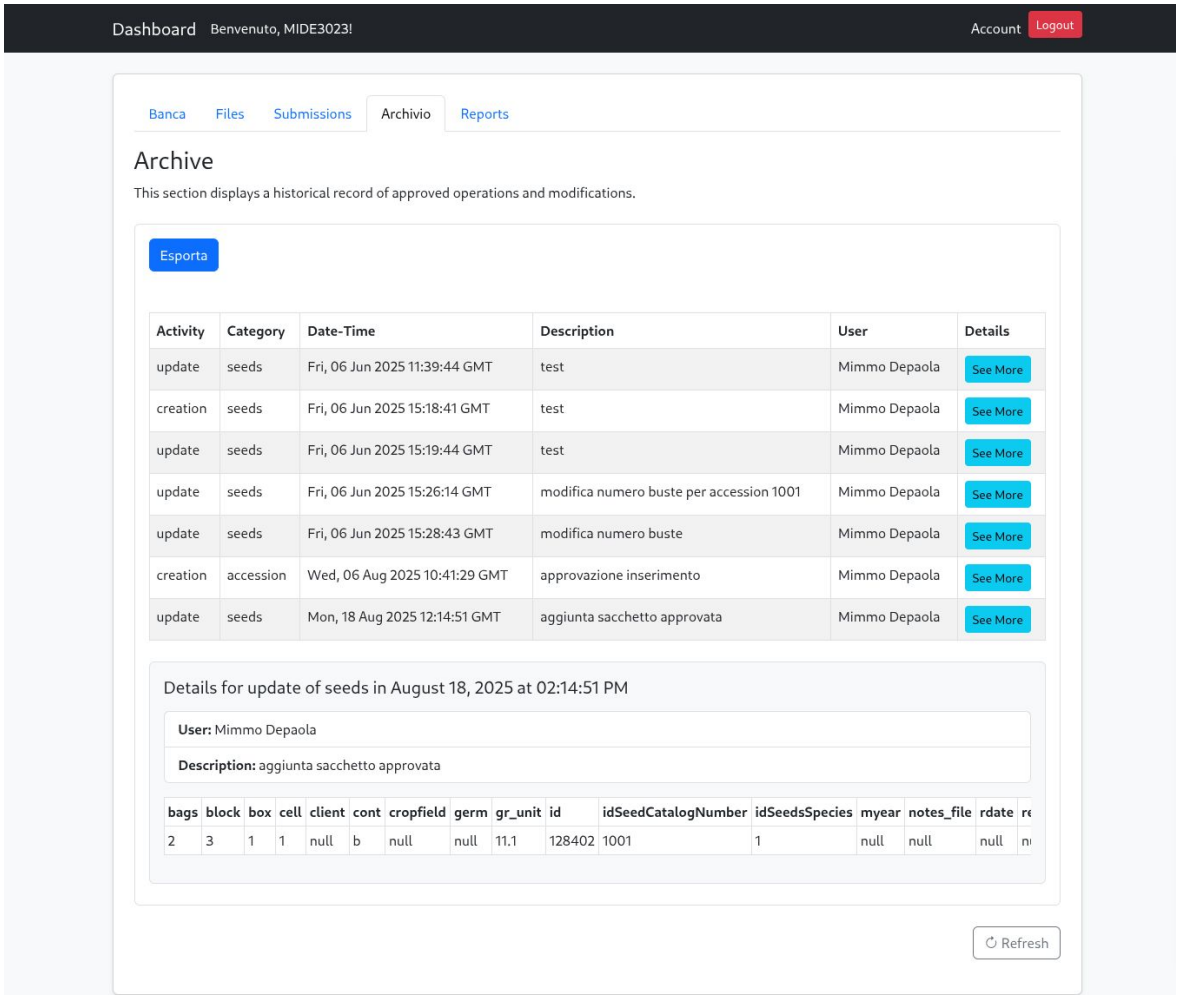




The Files section in the private dashboard of the MGD application provides a user-friendly interface to manage and upload images associated with germplasm records or research activities. Users can upload both general files and images using designated buttons. Once a file is selected, a live preview is displayed (for images), and users can optionally add metadata such as title, description, author, and date before finalizing the upload.

Under the hood, the uploaded files are managed through a connection with a MinIO object storage system. This ensures that files are stored securely, retrievably, and in a scalable way. The file_manager.py backend module confirms this by using the MinIO Python SDK to establish a client using environment variables for authentication and configuration.

10.3.5 Archive



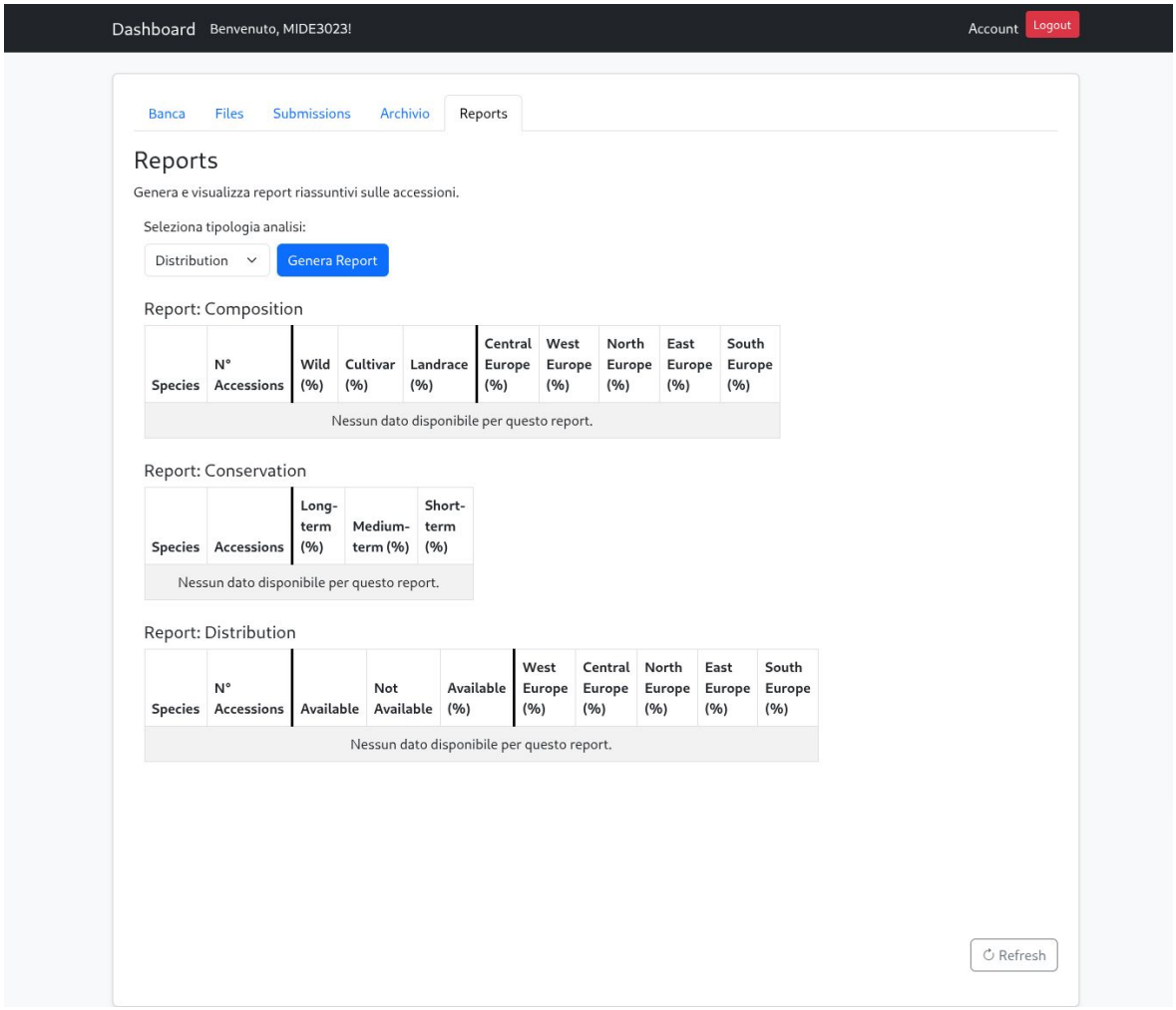
The Archive section of the MGD dashboard offers users a comprehensive view into the historical record of all operations that have been approved and committed to the system. It acts as an audit trail, providing both accountability and traceability over time.

At a glance, the archive table displays the type of activity (e.g., creation, update), the category involved (such as seeds or accessions), the exact date and time of the operation, a short description of what was done, and the user who performed the action. For more granular insight, each row includes a "See More" button that reveals the full details of the change, including all associated metadata and values before or after modification.

This section is especially useful for curators and administrators who need to verify data integrity, monitor recent submissions, or review previous decisions.

From a usability standpoint, the archive interface is streamlined and intuitive. A dedicated "Export" button allows data extraction for further reporting or analysis, and the structure of the table aligns with other parts of the dashboard, reinforcing consistency in user experience.

10.3.6 Reports



The Reports section of the MGD-BA portal, while still under active development, is conceptually inspired by internationally recognized Genebank Metrics frameworks—particularly those proposed by ECPGR, FAO, and initiatives like Pro-GRACE and CGIAR. These frameworks advocate for transparency, traceability, and measurable quality assurance in the management and documentation of genetic resources.

Within MGD-BA, the Reports area is envisioned to evolve into a comprehensive monitoring and evaluation module. Its purpose is to track the status and performance of the collections, aligning with core certification elements such as availability, conservation quality, and documentation standards. Although the current implementation may still lack full operational features, the existing structure already points toward the adoption of metric-based dashboards and key performance indicators (KPIs) similar to those outlined in the Genebank Metrics proposal.

In particular, the system is being shaped to handle both static indicators (e.g., total number of accessions, number of unique taxa, or wild/weedy populations) and dynamic performance indicators (e.g., number of germination tests conducted within a specific timeframe, distributions fulfilled, or safety duplications). These metrics are key components in certifying that a genebank operates following FAO Genebank Standards and ensures that plant genetic resources are conserved reliably and sustainably.

Moreover, the modular architecture of MGD-BA's reports aims to support future integration with

external monitoring tools and certification platforms. This ensures scalability and positions the platform to become a qualified node within larger biodiversity infrastructures. By embracing this vision early, the Reports section not only demonstrates MGD-BA's commitment to institutional accountability and scientific rigor but also establishes a foundation for contributing to Europe's long-term strategy for plant genetic resource conservation and accessibility.