MICROBE
the Microbiome Biobanking (RI) Enabler
www.microbeproject.eu

Deliverable D3.2

Search services for microbiome sample and dataset discovery across select core public repositories

Work package number and title	*WP3: Data infrastructure for microbiome biobanking*
Work package Leader	*EMBL*
Relevant Task	*Task 3.3*
Lead contributor to deliverable	*EMBL*
Dissemination Level	*public*
Due Date (month)	*M30*
Version	*1*

Contents

1. Background

Microbiomes comprise communities of microorganisms (i.e., microbiota that includes bacteria, archaea, protists, fungi and microalgae) and their "theatre of activity" (i.e., structural elements, metabolites, signal molecules, mobile genetic elements, as well as surrounding environmental conditions). Microbiomes exist in a diverse range of ecosystems on Earth, from deep oceans to fertile soil and seeds and play a key role in maintaining life on Earth by providing essential ecosystem services that are indispensable for the health of plants, animals and humans. There is a wide consensus that by harnessing microbiome functions, society would be better placed to tackle global challenges such as food security, health and wellbeing, food waste management, and climate change mitigation.

To facilitate the science necessary to achieve key advances in microbiome research, methodologies and technologies to capture or create, ensure stable long-term maintenance, and experimentally perturb microbiomes are required. Research infrastructures currently lack optimised methodologies and technologies to preserve and provide access to microbiome samples and extensive associated datasets. MICROBE is designed to address these issues by building upon and connecting: (1) technical solutions for microbiome preservation, propagation and functionality assessment, (2) novel ecological concepts (i.e. "core microbiome" and "microbial keystone taxa"), and (3) data infrastructures.

In light of the vast diversity and complexity of microbiomes, MICROBE will employ a focused approach that will ensure that the required developments and assessments can be comprehensively executed, resulting in a robust set of solutions. For this, a representative set of microbiome samples from soil, seed, marine and human systems have been collected. In the context of MICROBE, these can be linked to existing data sets, available knowledge and ongoing research projects, such as Biodiversity Exploratories[1] ensuring easy accessibility of samples and data, and access to external partners for the validation of methods developed during the project.

MICROBE Work Package 3 (WP3) is dedicated to the delivery of a comprehensive data infrastructure for microbiome biobanking. Such a data infrastructure requires a robust technical platform; data standardisation measures; support for cross-linkage between microbiome sample records and their derivatives, support for cross-linkage between sample records and any data that characterises those samples; and a strong capability for handling and managing data in accordance with the FAIR principles[2] (Findable, Accessible, Interoperable, Reusable).

Operational Framework for Microbiome Biobanking:

Technical Platform: Maintaining systems for efficiently managing and storing data related to microbiome samples, analytical data, and derived resources.

Data Standardisation Measures: Enforcing standardised procedures for the submission, storage and analysis of the data and metadata derived from microbiome samples. This guarantees consistency across various research endeavours and facilitates meaningful comparisons between studies.

Linkages Between Biological Samples, Analytical Data, and Derived Resources: Ensuring clear connections between the biological environmental samples and their sub-samples, the analytical data generated from these, and the resources derived from them (such as isolates). This linkage is important for traceability and reproducibility. Handling human-derived samples involves data privacy and GDPR considerations; a discussion of these issues is outside the scope of this deliverable. We expect to address any data privacy issues as we progress through the MICROBE project and will provide a detailed report on how MICROBE's data infrastructure handles GDPR in D3.4 (M36).

Handling and Managing FAIR data: To be a trusted research infrastructure requires that the data derived from research outputs is FAIR - Findable, Accessible, Interoperable and Reusable. Early steps on any journey towards ensuring data is FAIR tend to start with making those data readily discoverable and, where possible, retrievable using a standard identifier. Sharing records in established public repositories and ensuring searchability through those repositories is a good way to accomplish this aim.

This document illustrates the importance and implementation of cross-repository search services in supporting sample and dataset discovery. MICROBE partners will generate and archive a variety of different data types and datasets; the ability to efficiently locate and access these will be fundamental to fostering collaboration and maximising the value of existing data. **BioSamples**, along with the **European Nucleotide Archive (ENA), BioStudies** and **MGnify**, make up the key components of the MICROBE data infrastructure, and we have developed search services for the relevant data in these repositories.

2. Introduction

The data ecosystem is fragmented, and the majority of data is found and reused usually by virtue of being associated with publications. Depositing data and rich metadata describing the data is crucial to the findability and downstream reuse of the data. Data deposited with minimal metadata is of limited use as, without the context and provenance information, reliable analyses become more difficult and interoperability with other data sets virtually impossible.

Linking different data types stored in different repositories can also be a challenge. Here we describe the search services we have developed to ensure the findability of samples and associated data in the MICROBE project. Additionally, the established infrastructure for sample submission, linking, and discovery could be used as a blueprint for microbiome biobanking research infrastructure.

The ability to efficiently locate and access samples and datasets after deposition is vital for data reuse and collaboration. This document focuses on the four core public repositories (all ELIXIR Core Data Resources[3] and/or Deposition Databases[4]) where MICROBE data is deposited: BioSamples[5], ENA[6], BioStudies[7], MGnify[8] and the development and evaluation of search services for microbiome sample and dataset discovery across these repositories. Each repository has their own individual search service for public data, and we have developed a search service which allows search across harmonised MICROBE sample metadata, providing information on the availability of the genomic sequence data associated with these samples held within the consortium. This will ultimately be expanded to all data types once the data becomes publicly available whereby this unified search functionality will improve the findability and reusability of MICROBE samples, allowing further reuse and collaboration.

Our focus is on improving findability while ensuring traceability and connectivity between samples, to enable the associated data to be FAIR and sustainable beyond the life of the project. MICROBE aims to enable the establishment of a microbiome biobanking research infrastructure; capturing extensive standardised metadata linked to associated data will provide the provenance required for this. The data generated within the MICROBE project will be made publicly available towards the end of the project and will be permanently accessible as described below.

The data types generated by the MICROBE project are outlined in the Data Management Plan, D7.1, which is updated regularly. A list of the data types is shown below.

Type of data	Proposed hosting solution	Data sharing mechanism
Microbiome and microbial genome sequence data	European Nucleotide Archive (ENA) at EMBL-EBI	Open access (after publication)
Microbiome and microbial transcriptome sequence data	European Nucleotide Archive (ENA) at EMBL-EBI	Open access (after publication)
Microbiome and microbial sample descriptions	BioSamples database at EMBL-EBI	Open access
Functional microbiome profiles	MGnify, BioSamples and BioStudies databases at EMBL-EBI	Open access (after publication)
Taxonomic microbiome profiles, isolates profiles synthetic consortia profiles	MGnify, BioSamples and ENA databases at EMBL-EBI	Open access (after publication)
Metabarcode sequence data and metagenome- assembled genomes	ENA database at EMBL-EBI	Open access (after publication)

Table 1. Overview of data types that will be generated by the MICROBE project and their proposed hosting solutions

3. Selected Core Public Repositories

All the Core Repositories mentioned below provide unique persistent identifiers for meta(data) submissions or analysis, which are instrumental in the subsequent linking of different data types.

3.1 BioSamples at the European Bioinformatics Institute (BioSD)

The BioSamples repository (https://www.ebi.ac.uk/biosamples/) stores and supplies descriptions and metadata about biological samples used in research and development by academia and industry. Samples are either 'reference' samples (e.g. from 1000 Genomes, HipSci, FAANG) or have been used in an assay database such as the European Nucleotide Archive (ENA) or ArrayExpress. BioSamples provides links to assays and specific samples and accepts direct submissions of sample information.

Within MICROBE, BioSamples is used for structuring, cross-linking and tracking samples and sample collections used in microbiome biobanking research. Samples will first be registered in BioSamples and each sample will be issued with a unique persistent identifier and all samples will be made public at the point of submission. Samples will be linked back to the original environmental sample they were derived from, providing context and provenance to support downstream assessment and validation analysis.

A Guidance on microbiome sample accessioning, Deliverable 3.1, was submitted in month 12. This is a practical guide on submitting samples to the BioSamples database and explains how samples are structured and linked to the data, as well as the importance of rich metadata.

The BioSamples search services allow both programmatic searching via the JavaScript Object Notation Application Programming Interface (JSON API) or using the BioSamples User Interface (UI) to search with keywords. The UI also allows filtering of available metadata attributes and selecting combinations of both metadata attributes and values to display those samples that match the specified criteria. BioSamples also provides certain advanced search features, such as boolean, wildcard and range queries, or filtering for samples with associated data in an external archive.

All MICROBE samples can be found by filtering for samples with the project name 'MICROBE' (https://www.ebi.ac.uk/biosamples/samples?filter=attr%3Aproject+name%3AMICROBE) or by clicking on the MICROBE link on the BioSamples homepage.

3.2 European Nucleotide Archive (ENA)

The European Nucleotide Archive (https://www.ebi.ac.uk/ena; ENA) is a globally comprehensive data resource for nucleotide sequence, spanning raw data, alignments and assemblies, functional and taxonomic annotation and rich contextual data relating to sequenced samples and experimental design. Serving both as the database of record for the output of the world's sequencing activity and as a platform for the management, sharing and publication of sequence data, the ENA provides a portfolio of services for submission, data management, search and retrieval across web and programmatic interfaces. The ENA is part of the International Nucleotide Sequence Database Collaboration (INSDC), which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at the NCBI. These three organisations exchange data daily.

ENA datasets can be used by MICROBE partners, MGnify, and a wide range of data consumers in ENA-provided format without any conversion needed. As ENA is part of the INSDC, hosting MICROBE data there ensures global accessibility across all platforms.

Sample registration in BioSamples is required prior to sequence submission to ENA. Upon public release of the data, a bidirectional link exists between BioSamples and ENA ensuring samples and associated data relationships are clearly identifiable.

Within MICROBE, ENA will be used for storage of read data from microbiome samples and synthetic consortia, storage of metagenome assemblies, linking newly derived metagenome-assembled genomes (MAGs) back to read data, and taxonomic classification of MAGs.

ENA data can be searched and retrieved interactively and programmatically and visualised using the ENA browser. These searches include options for free text search based on keywords, searches based on nucleotide sequence similarities, searches using the cross-reference Xref search service which cross-references external data resources linked to ENA, and an ‘Advanced Search feature. The ‘Advanced Search’ features an intuitive UI that allows users to discover reads, assemblies and accompanying metadata. This search feature allows filtering on different metadata fields, selecting desired metadata fields to download, and generating scripts to download data.

Once public, all data in the MICROBE project can be discovered under the umbrella study PRJEB74917 (https://www.ebi.ac.uk/ena/browser/view/PRJEB74917) where all sequence data and associated samples will be linked.

ENA Data Hubs

The ENA Data Hubs system enables users to securely share pre-release data that has already been submitted to ENA with a selected group of collaborators.

Within the MICROBE project, partners can submit genomic sequence data to ENA for validation and subsequently add this to the Data Hub which allows the data to be shared within the consortium whilst kept private until a specified release date. This data includes read data and metagenome assemblies. Data held in the Data Hub are secured through EMBL-EBI's bespoke authentication service, WEBIN. The Data Hub includes services for data discovery, filtering based on metadata, and the ability to download metadata and other associated files.

The Data Hub uses a private version of the 'Advanced Search' feature provided in ENA that only allows permitted users to search, filter, and access private data.

3.3 BioStudies at the European Bioinformatics Institute

BioStudies (https://www.ebi.ac.uk/biostudies) can provide access to the data outputs of a life sciences study from a single place, organises links to data in other databases at EMBL-EBI or elsewhere, and hosts data and metadata that cannot be submitted to existing repositories. It also enables manuscript authors to submit supplementary information and link to it from the publication.

The database accepts submissions via an online tool, or in a simple tab-delimited format. BioStudies provides rich mechanisms for defining and using metadata guidelines specific for a particular data source such as a project or a community. It can organise datasets into 'collections', which allows for grouping of datasets with shared common features or originating from a single source.

Within MICROBE, BioStudies will be used for storage of functional assay data, such as the Biolog and 4MU data types in tabulated format. The data will be held in BioStudies, with bidirectional links to the relevant samples in BioSamples.

3.4 MGnify

MGnify (https://www.ebi.ac.uk/metagenomics) is a free-to-use resource aiming at supporting all metagenomics researchers. The service is an automated pipeline for the analysis and archiving of metagenomic data submitted to the ENA that aims to provide insights into the phylogenetic diversity as well as the functional and metabolic potential of a sample. All the public data in the repository can be freely browsed, while private data are accessible through the WEBIN authentication service.

MGnify provides functional and taxonomic analysis for an extensive range of data types, including metagenomic raw reads and assemblies, and amplicon (metabarcoding). This platform covers studies conducted in diverse environments, also referred to as "biomes", which can be searched and queried on the MGnify website. Among these environments, three of the four key areas of interest for MICROBE, namely human, soil, and marine ecosystems, have already been extensively covered, both under a metagenomic and a metabarcoding viewpoint. MGnify also provides biome-specific genome catalogues, including both isolate genomes and metagenome-assembled genomes (MAGs) from a collection of publicly available studies.

The MGnify resource provides a variety of options to search and access data products:

- “EBI Search” is a multi-resource search service enabling text-based and faceted searches across all EBI resources. This includes MGnify's samples, studies, and analyses. This allows users to locate, for example, all studies related to a particular biome.

- MGnify has a comprehensive set of API endpoints that allow for computational queries which can then be tailored through filtering on a variety of fields. For example, it is possible to fetch all MGnify studies pertaining to samples collected from the soil biome. API calls can also be combined to build complex queries, such as matching sample metadata to specific subsets of analyses.

- MGnify Genomes supports searching of a user-provided genome against the genome catalogues, using Sourmash. It uses a sketching approach, generating the sketch for the user-provided genome within the browser.

- MGnify Genomes also supports short gene/fragment sequence searching, based on Clustering of Basic Sequences (COBS). This search option uses a k-mer-based approach to compare query sequences against MGnify Genomes.

- MGnify Sequence Search uses HMMER technology to compare user-provided protein sequences against the MGnify Proteins database. This database is populated from the

MGnify assembly analyses, and thus provides links back to the MGnify analyses results for any matches.

MGnify also provides a selection of Jupyter notebooks (https://docs.mgnify.org/src/notebooks_list.html) which serve as an advanced search interface, offering ready-to-use and editable examples of downstream analyses, through interfacing with the MGnify API. Among these notebooks, users can find interactive tools to perform cross-study comparative metagenomics analyses.

Additionally, certain MGnify datasets are exclusively available through the EMBL-EBI FTP server (http://ftp.ebi.ac.uk/pub/databases/metagenomics). The MGnify Proteins database is freely available via FTP as a series of flat files to be downloaded and used locally. MGnify Proteins can also be queried through Google's BigQuery as the MGnify Proteins database is available as a BigQuery public dataset. However, users should be aware that BigQuery is a paid-for service.

MGnify also groups related studies into super studies, which serve as collections of individual studies that have contributed to the work of a specific consortium. For example, the MICROBE super study (https://www.ebi.ac.uk/metagenomics/super-studies/MICROBE) provides a unified entry point for accessing all data generated within the consortium, as well as publicly available datasets used as benchmarks.

The FTP server is also where users can access the full MGnify Genomes catalogues, including all genome cluster members not accessible through the MGnify web interface or API.

4. Data Flow

Metadata Checklists and Harmonisation across partners

Data harmonisation is essential when integrating data from multiple sources, especially in collaborative research environments involving different labs, organisations, and teams. Each contributor may use different metadata standards, field names, and file formats, which can lead to inconsistencies that hinder effective filtering, searching, and analysis.

To address this, the MICROBE consortium implemented a harmonised metadata framework by extending the Genome Standards Consortium Minimum Information about any (x) Sequence (GSC MIxS) checklist with additional terms based on partner feedback and alignment with community-driven standards like STORMS[9] and STREAMS[10]. This harmonisation process resulted in a clear checklist of mandatory, recommended, and optional fields, including defined data types, controlled vocabularies, and ontologised values.

For example, an ontologised field titled 'preservation method' was added to store vital metadata for MICROBE cryopreservation trials in BioSamples. Validating metadata against the checklist increases quality control, and easier data integration and searching through the data and metadata.

These repositories are connected by the unique BioSamples SAMEXXXXXXXXX accessions as a key. The harmonised metadata in each repository is indexed for search, and filtering, resulting in greater findability and reusability.

After sample collection and preservation assays, sample metadata is submitted to BioSamples and made publicly available immediately. Sequence data and metadata are submitted to ENA and subsequently linked to both the Data Hub and the MICROBE overarching umbrella study. Most data is still private at this point in the project; however, these can be accessed from the Data Hub by MGnify for metagenomic analysis. At any point during this journey, additional data can be added or updated as structured tables to BioSamples (see section 5.3 below), linking out to functional assays in BioStudies. Once data is made public on ENA, this is also immediately reflected in MGnify, making sequence reads and MAGs (in ENA), and metagenomic analyses (in MGnify) publicly available.

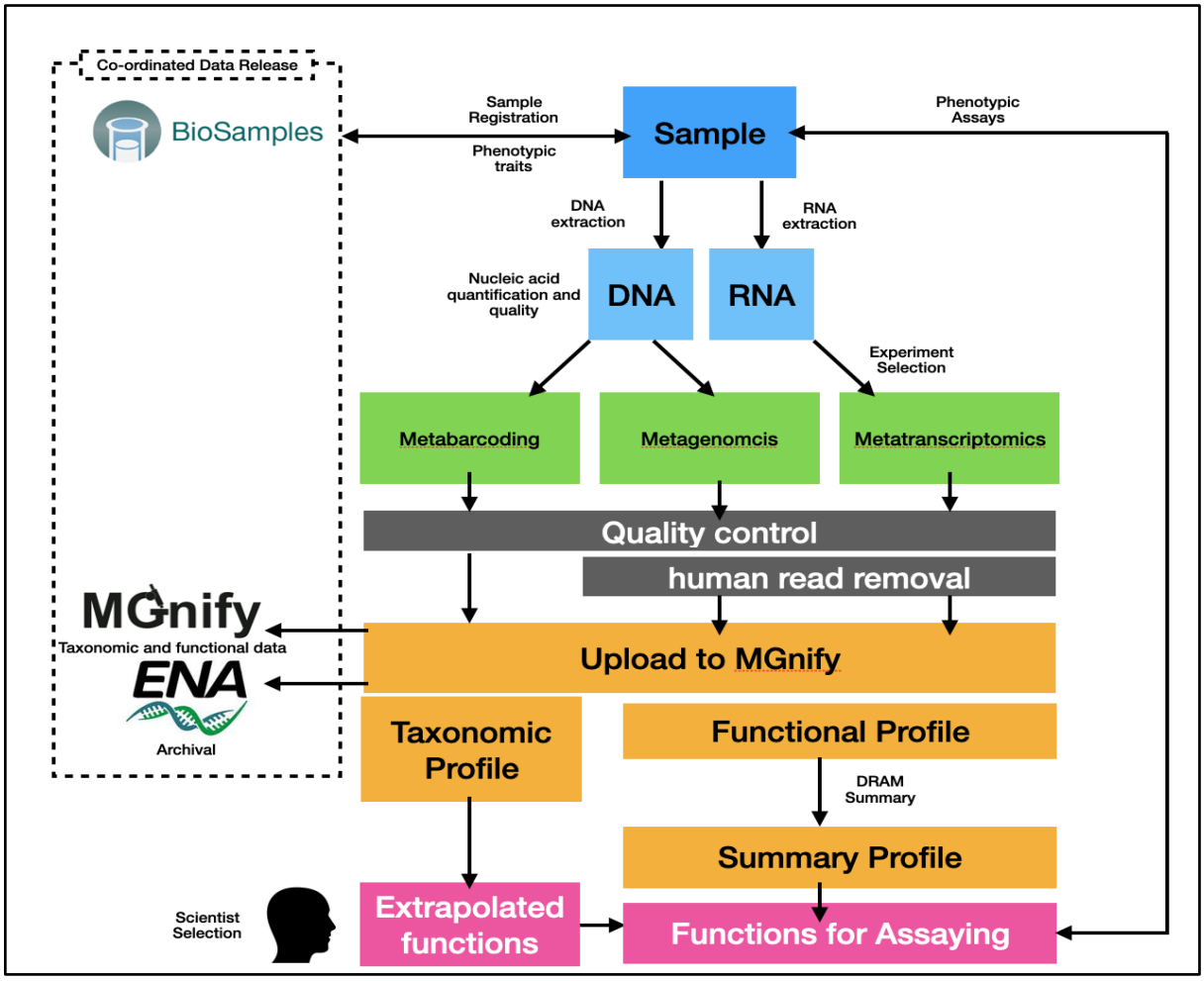


Figure 1: Illustration of data flow between selected core public repositories

5. Data Structure

Each of the selected core public repositories stores metadata and data of a different context, data types and data models. BioSamples stores sample metadata, including collection, transport, and sample descriptions. ENA stores metadata and data files related to genomic sequence reads, and structures data using the INSDC data model, with overarching studies, component studies, their samples, experiments, reads, and links out to downstream analysis files. MGnify primarily stores metagenomic analyses, functional and taxonomic analysis metadata and data files. The MICROBE data structure follows the ENA model, with an overarching MICROBE study: PRJEB74917, and component studies specific for soil, seed, and marine genomic/transcriptomic data and samples. All samples are associated with the MICROBE study and classified into one of the component studies.

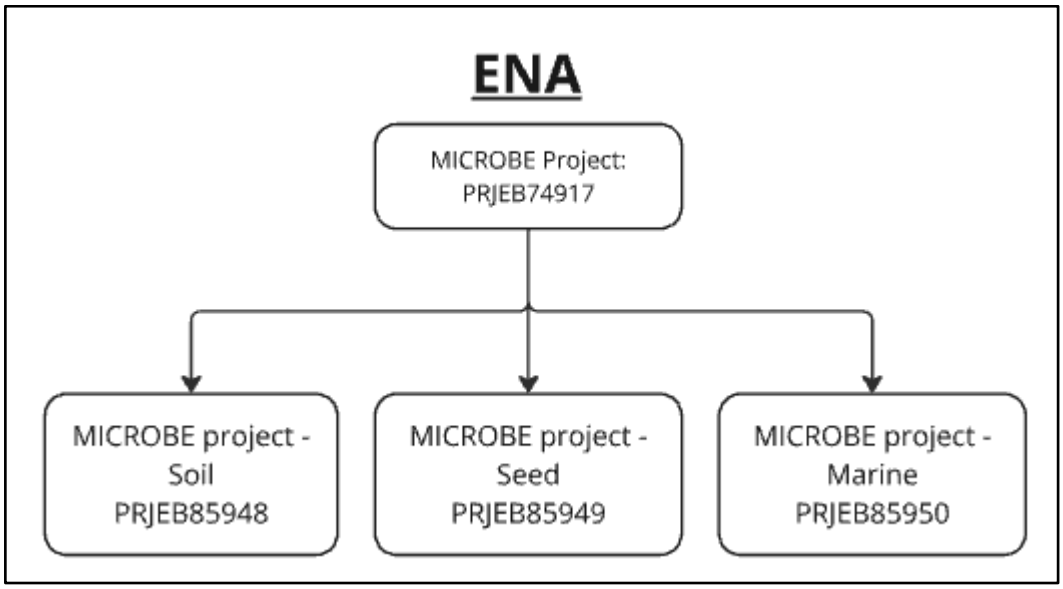


Figure 2: MICROBE data structure in ENA

5.1 Sample Structure

As search services are highly dependent on how the samples and data are structured, defining the structures and relationships between samples is essential to developing functional and scalable cross-repository search services.

Samples are uniquely identified using a BioSamples accession number (e.g. SAMEA115407051). As BioSamples accessions are supported by ENA, BioStudies and MGnify, the BioSamples repository functions as a 'central source of truth' from which all the other

repositories are linked. BioSamples itself stores a significant amount of sample metadata but cannot easily display the specific metadata and data stored in other repositories. The unique sample accession operates similarly to a key in a cross-repository relational database.

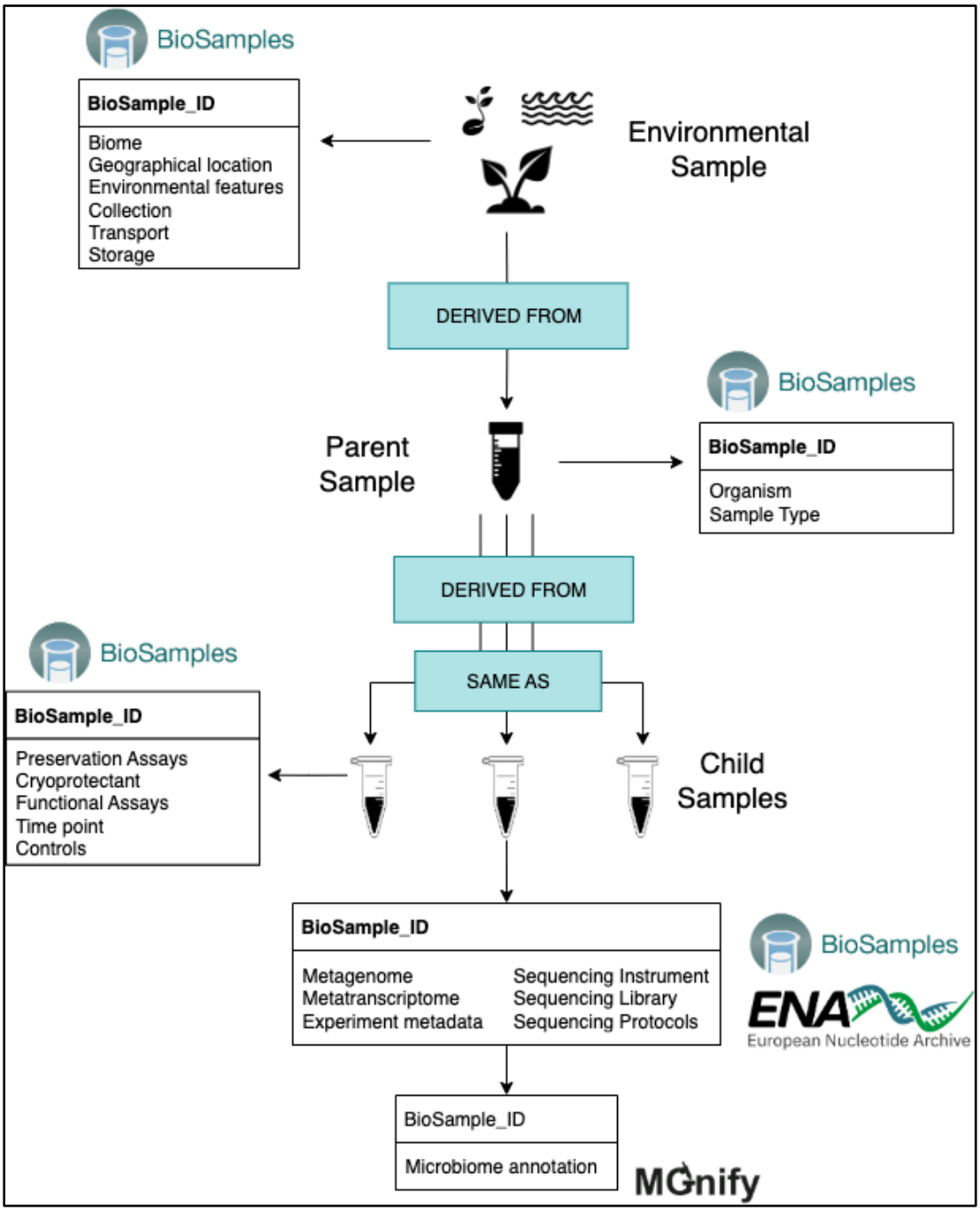


Figure 3: MICROBE sample structure

5.2 Sample Relationships

Sample relationships in BioSamples describe the connections between parent and child samples, biological replicates, technical replicates, and negative controls. Accurate and precise representation of these relationships is essential to efficient use of the data, and to building a search service that can filter and provide them. These relationships establish links between different samples and facilitate relationship-based graph searches.

When a submitter provides relationship information for one sample, corresponding reverse relationships in related samples are automatically generated.

Two types of sample relationships are used in MICROBE:

1. **Derived From:**
    - **Description:** Sample A is derived from Sample B.
    - **Example:** SAMEA115408618 is derived from SAMEA115407051
2. **Same As:**
    - **Description:** Sample A is the same as Sample B. Useful for linking duplicated samples.
    - **Example:** SAMEA115408618 is the same as SAMEA115408606

Relationships

Source	Type	Target
SAMEA115408618	derived_from	SAMEA115407051
SAMEA115408618	same_as	SAMEA115408606

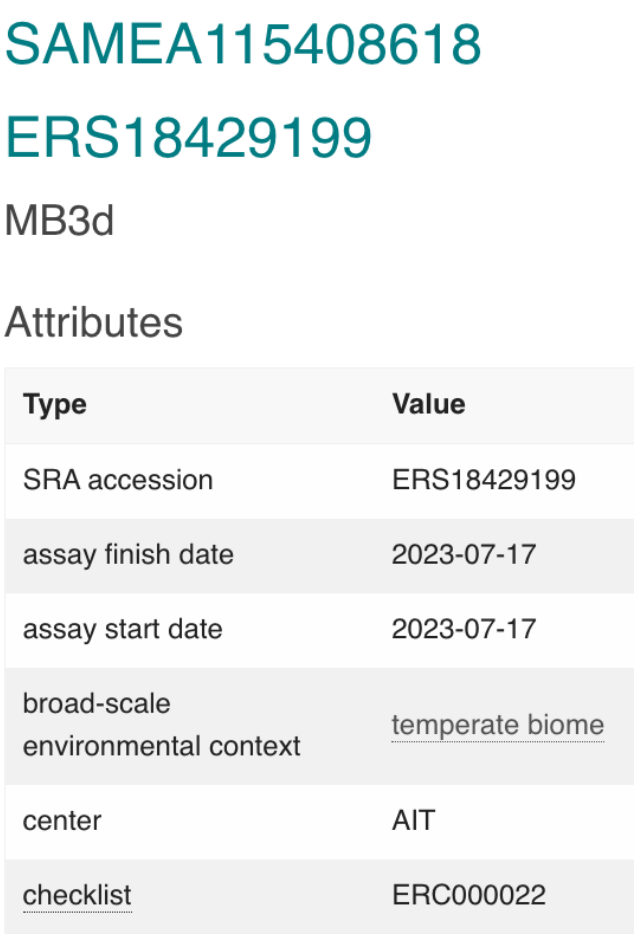
Figure 4: Representation of sample relationships in BioSamples

These relationships provide a structured framework for understanding the connections and associations between various samples within the BioSamples database. Furthermore, these relationships will feed into the data portal that will be developed as part of the final data infrastructure for microbiome biobanking.

5.3 Structured Tables

Samples within the BioSamples database store metadata as key/value pairs and as structured tables, which allow a more structured way of storing and displaying metadata. Structured tables are used for metadata and tabular data such as antimicrobial resistance or supporting assays. In MICROBE, structured tables are currently used for two supporting assays, Biolog and 4MU functional assays for determining metabolic activity. This is a flexible system that can include different assays developed as the project progresses.Samples can be filtered based on the type of structured data table e.g. Biolog data.

Structured tables can link out to data files stored in BioStudies, providing a link between the sample and the functional assay data file.


SAMEA115408618

ERS18429199

MB3d

Attributes

Type	Value
SRA accession	ERS18429199
assay finish date	2023-07-17
assay start date	2023-07-17
broad-scale environmental context	temperate biome
center	AIT
checklist	ERC000022

Figure 5: Key:value metadata fields in BioSamples

Structured Data (Biolog)

Incubation temperature	Plate shaken	Reader device	Type of plate	data_preprocessing_status	data_url	plate	timepoint
20 degree celsius	No	Omnilog	Ecoplate	Processed	https://ftp.ebi.ac.uk/pub/databases/biostudies/.beta/S-BSST/572/S-BSST2572/Files/hca_template%20(11).xlsx	A	24H
20 degree celsius	No	Omnilog	Ecoplate	Raw	https://ftp.ebi.ac.uk/pub/databases/biostudies/.beta/S-BSST/572/S-BSST2572/Files/hca_template%20(11).xlsx	A	72H
20 degree celsius	No	Omnilog	Ecoplate	Processed	https://ftp.ebi.ac.uk/pub/databases/biostudies/.beta/S-BSST/572/S-BSST2572/Files/hca_template%20(11).xlsx	A	48H

Figure 6: Structured data table in BioSamples

6. MICROBE Catalogue

To facilitate searching and data set discovery across different repositories, we have developed the MICROBE Catalogue as an internal consortium cross-repository search service for tracking samples.

The Catalogue (https://www.ebi.ac.uk/microbe/#/samples) enables discovery of MICROBE samples via filtering and searching on metadata fields such as timepoints, cryoprotectants, freezing methods and preservation temperature. This allows the tracking of which samples have been registered, and the status of the associated sequencing data; "Public', 'Private' or 'Not Sequenced'.

The catalogue pulls data from the select core repositories mentioned above. As shown in Table 1, most data files are still privately held in the ENA Data Hub and are not publicly accessible yet. Thus, the catalogue is mostly centred around publicly available metadata that is stored in BioSamples.

Figure 7: MICROBE Catalogue for sample tracking

7. Conclusion

In this deliverable we describe the search services for core public repositories utilised within the MICROBE project for tracking samples and associated data. We highlight the importance of structuring sample relationships and linking harmonised samples to both other samples and associated data, as this is essential in the findability of all the meta(data). As the data becomes publicly available, this structure will also support the accessibility, usability, and interconnectedness of microbiome data.

As the project progresses and additional data types, including functional assay data and synthetic community data, are generated, additional metadata fields, views, and data files will be added. This will include linking to Culture Collections. All of these enhancements will contribute to the development of a comprehensive MICROBE Data Portal as part of Deliverable 3.4, MICROBE's data infrastructure for microbiome biobanking. The Data Portal will be a public, external-facing user interface which will allow the searching, filtering, and downloading of all available samples, sequence and metagenomic data within the MICROBE project that resides in established public core repositories mentioned above, as well as links to external resources.

Depositing and linking metadata and data in these repositories and linking to external resources such as culture collections is a fundamental step in developing the Data Portal, and for ensuring the long-term accessibility and preservation of the data beyond the project's conclusion. The described infrastructure could be used as a blueprint for future microbiome biobanking efforts requiring long-term accessibility. The process of sample submission, cross-linking between harmonised samples and data, and standardised checklists can be taken as a base template which takes advantage of existing core public repositories, and is highly adaptable for different use cases. This would allow future projects to reuse, rather than reinvent infrastructure for FAIR data management.

8. References

[1] https://www.biodiversity-exploratories.de/en/

[2] Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[3] https://elixir-europe.org/platforms/data/core-data-resources

[4] https://elixir-europe.org/platforms/data/elixir-deposition-databases

[5] Courtot, M., Gupta, D., Liyanage, I., Xu, F. and Burdett, T. BioSamples database: FAIRer samples metadata to accelerate research data management. *Nucleic Acids Res.* 2022 Jan 7; 50(D1): D1500–D1507 doi: 10.1093/nar/gkab1046

[6] O'Cathail *et. al* The European Nucleotide Archive in 2024. The European Nucleotide Archive in 2024. *Nucleic Acids Res.* 2024 **53**(D1), D49–D59 https://doi.org/10.1093/nar/gkae975

[7] Sarkans *et. al.* The BioStudies database - one stop shop for all data supporting life science study *Nucleic Acids Res. 2018* **46** D1266-1270 https://doi.org/10.1093/nar/gkx965

[8] Richardson, L *et. al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* 2023 Jan; 51(D1): D753–D759 https://doi.org/10.1093/nar/gkac1080

[9] Mirzayi C *et al.* Reporting guidelines for human microbiome research: the STORMS checklist. *Nat Med. 2021* 27(11):1885-1892. doi: 10.1038/s41591-021-01552-x

[10] Kelliher, J. *et. al* Microbiome data management in action workshop: Atlanta, GA, USA, June 12–13, 2024 *Environmental Microbiome 2025* ***20*** *(40) doi.org/10.1186/s40793-025-00702-9*