

BioSamples database: the global hub for sample metadata and multi-omics integration

Dipayan Gupta^{1,*}, Isuru Liyanage¹, Yasset Perez-Riverol¹, Enrique Sapena Ventura¹, Wei Kheng Teh¹, Tim Van Den Bossche^{2,3}, Senthilnathan Vijayaraja¹, Amnon Khen¹, Ugis Sarkans¹, Tony Burdett^{1,4}, Jeena Rajan^{1,*}

¹European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

²VIB-UGent Center for Medical Biotechnology, VIB, Ghent 9000, Belgium

³Department of Biomolecular Medicine, Ghent University, Ghent 9000, Belgium

⁴Present address: BioFAIR Hub, Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, United Kingdom

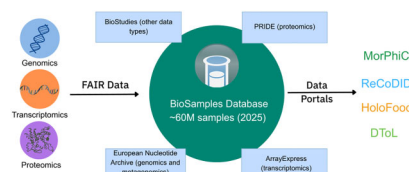
*To whom correspondence should be addressed. Email: dgupta@ebi.ac.uk

Correspondence may also be addressed to Jeena Rajan. Email: rajan@ebi.ac.uk

Abstract

The BioSamples database (<https://www.ebi.ac.uk/biosamples/>) at the European Bioinformatics Institute (EMBL-EBI) is a core infrastructure resource that provides a centralized platform for the storage, curation, and dissemination of biological sample metadata. BioSamples holds sample descriptions across diverse datasets, enhancing their reusability and integrability in accordance with the FAIR (findable, accessible, interoperable, and reusable) principles. BioSamples serves as the central place to store the sample metadata for multiple repositories and archives at EMBL-EBI, including the European Nucleotide Archive, ArrayExpress, and the European Phenome-Genome Archive. In this article, we outline technical updates made to support increased sample registration and provide a better user experience. We describe four use cases whereby different communities have leveraged BioSamples to connect multi-omics data and enable the creation of public data portals and data status trackers. BioSamples is freely available, and its content is distributed under the EMBL-EBI Terms of Use available at <https://www.ebi.ac.uk/about/terms-of-use>. The BioSamples code is available at <https://github.com/EBI/BioSamples/biosamples-v4> and <https://doi.org/10.5281/zenodo.17304021> and is distributed under the Apache 2.0 license.

Graphical abstract



Introduction

Registration of sample metadata at the EMBL-EBI BioSamples database [1] results in persistent identifiers (accessions) that can be used as part of submissions to other omics archives, such as the European Nucleotide Archive (ENA) [2], ArrayExpress [3], the European Phenome-Genome Archive (EGA) [4], PRIDE [5], and MetaboLights [6]. Reusing BioSamples accessions links the different data types through the metadata, providing context and enabling cross-functional analyses. In recent years, BioSamples has evolved into a central hub for coordinating sample metadata, extending beyond EMBL-EBI resources to a wider ecosystem of databases, repositories, and community-specific portals that present omics data. This expansion began in 2022 with ReSOLUTE [1], one of the first consortia to adopt BioSamples as the foundation of its sample metadata system, and has since accelerated with newer

initiatives such as MorPhIC [7] and the Darwin Tree of Life project [8]. While earlier versions of BioSamples primarily supported metadata from genomics and transcriptomics studies (e.g. RNA-seq experiments), the scope has broadened substantially. BioSamples now actively collaborates with the proteomics community, positioning itself as a large-scale multi-omics hub that connects diverse resources and enables integrative analyses across data modalities.

Since our previous NAR update in 2022 [1], the BioSamples database has seen significant growth, with the number of registered samples increasing three-fold to 60 million by 2025. In this article, we outline key infrastructure advancements, including the introduction of bulk submission capabilities to support large-scale sample metadata deposition and enhancements to the RESTful API that improve programmatic access and interoperability. Together, these developments es-

Received: September 15, 2025. Accepted: September 30, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

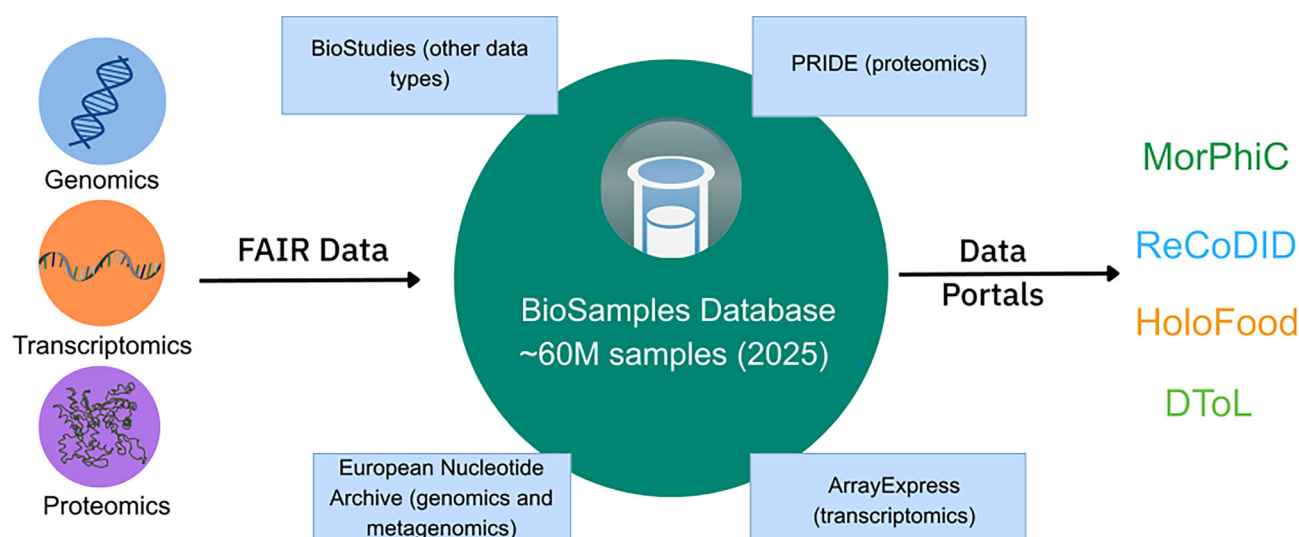


Figure 1. The BioSamples database links multi-omic data through its role as a central metadata hub, enabling unified dataset views via community data portals.

establish BioSamples as a key hub for managing omics sample metadata in the future. We also describe the new checklist editor and the growing ecosystem of databases, resources, and portals that rely on BioSamples as their central metadata hub (Fig. 1). Finally, we highlight the integration of BioSamples with non-genomics resources such as PRIDE (proteomics), underscoring its role as a foundation for multi-omics data management and reuse.

Technical improvements

The BioSamples team has implemented several technical enhancements to improve the database infrastructure and ensure the continued sustainability of BioSamples in response to increasing demand, as outlined below.

Bulk sample metadata submission

Originally, the BioSamples database only accepted single-sample submissions; this approach was adequate when submission rates were modest. However, growing demands from large-scale repositories such as the ENA and consortia like the EMBL TREC (Traversing European Coastlines) Initiative; <https://www.embl.org/about/info/trec/>, which have increasingly delegated sample management to BioSamples, have driven a sharp rise in sample submissions (see Fig. 2). To efficiently manage the growth in samples associated with large-scale experiments and maintain the infrastructure (RESTful API, submission system) responsiveness, BioSamples has recently introduced a bulk submission framework enabling users and partner archives to submit large batches of samples in a single request. This bulk submission approach significantly reduces the communication overhead and latency inherent in processing individual submissions sequentially. By batching multiple samples, it minimises the number of request-response cycles between BioSamples and collaborating archives like ENA, resulting in faster ingestion and validation workflows. Moreover, bulk submissions help alleviate network congestion and server load, enhancing scalability and robustness. The bulk submission mechanism is supported through well-defined RESTful APIs and standardised metadata formats, ensuring seamless integration with upstream

data providers and automated pipelines. This development is critical in supporting high-throughput projects and consortia that generate extensive biological sample data, enabling BioSamples to serve as a scalable central metadata hub for diverse omics data ecosystems (Fig. 1).

Drag and drop submission

To better support researchers, particularly those without programming experience, BioSamples introduced a simplified drag-and-drop interface. Submitting biological samples one by one is often slow and prone to errors, especially in studies involving large datasets or complex designs. The new drag-and-drop upload tool allows users to submit multiple samples at the same time using a structured, human-readable, tab-delimited ISA-tab [9] format (e.g. <https://www.ebi.ac.uk/biosamples/upload/downloadExampleFile>). This enables researchers to prepare their sample metadata offline in familiar spreadsheet tools, significantly reducing the time and effort required for submission.

The interface also supports the creation of relationships between samples, e.g. linking a processed sample back to its original tissue/material or grouping all samples from a particular study. This helps preserve the biological and experimental context without requiring technical knowledge of APIs or ontologies. By minimising manual input and validation errors, this tool lowers the barrier to entry for high-quality metadata submission, making BioSamples accessible to a broader research community, including those in smaller labs or collaborative projects with limited informatics support.

New checklist editor, JSON schema store, and biovalidator

The new checklist editor has been successfully developed and rolled out in 2025 by the BioSamples team, replacing the legacy ENA-maintained editor. This tool enables checklist maintainers to create and edit sample checklists along with their associated fields, offering a user-friendly interface and support for exporting checklists as TSV or JSON formats. Sample submitters can select an appropriate checklist, populate it with sample data, and submit it directly to either ENA

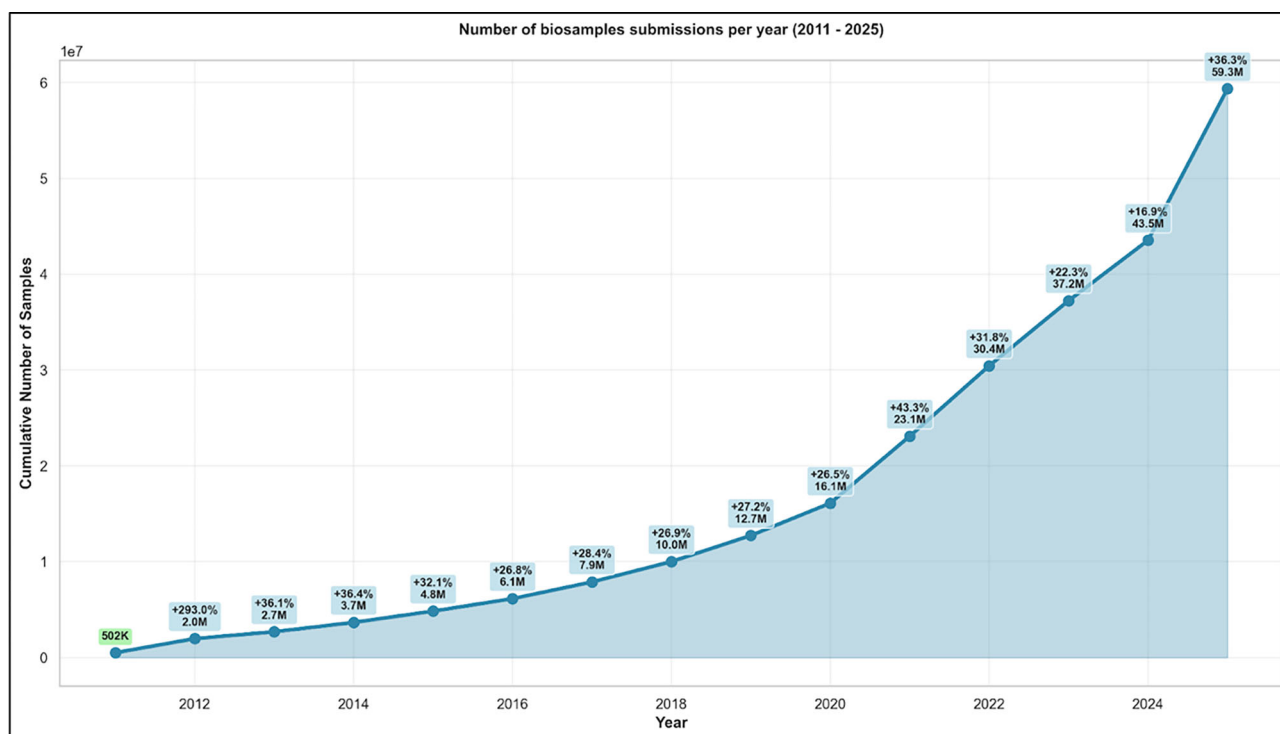


Figure 2. Number of BioSamples submissions per year.

or BioSamples, thereby streamlining the metadata submission process.

The primary goal of the new checklist editor rollout was to modernise checklist management by transitioning to a JSON-based system that is easier to maintain and extend and to enable checklist versioning. New checklists are created using the editor and stored in a centralised JSON Schema Store. Checklist versions are maintained for backward compatibility, ensuring that existing submissions remain valid even as new versions evolve.

When a sample needs to be validated, the corresponding checklist is retrieved from the schema store, and both the sample metadata and checklist schema are passed to the biovalidator [10] for validation. This integrated flow ensures consistent structure, improved data quality, and seamless interoperability with existing services, marking a significant step towards a more sustainable and flexible metadata infrastructure.

We have introduced new features and enhancements to biovalidator, including identifier validation, improved cache management, and support for the latest JSON Schema. These updates make the tool more adaptable for use across archives such as EGA at EMBL-EBI. Together with the checklist editor and JSON Schema Store, the biovalidator forms part of a growing ecosystem of validation tools designed to strengthen the FAIRness of life sciences metadata, as shown in Fig. 3.

ENA–BioSamples integration

As mentioned above, the BioSamples database now serves as the central hub for sample metadata management, fully integrating workflows previously managed by the ENA. This integration covers validation, accessioning, storage, and metadata exchange with the International Nucleotide Sequence Database Collaboration (INSDC) [11] partners, providing standardised pipelines and dedicated infrastructure that en-

sures consistent, high-quality, and non-redundant metadata across the international data-sharing ecosystem.

For researchers, sample registration is streamlined; users can submit samples directly to BioSamples (<https://www.ebi.ac.uk/biosamples/submit>), independent of subsequent sequence data submissions to ENA. This separation of roles establishes BioSamples as the authoritative source of sample metadata, while ENA focuses on sequencing data. It also enables sample metadata to be publicly released earlier than the associated sequence data, a practice many large consortia now adopt for early visibility, such as the EMBL TREC (Traversing European Coastlines) Initiative. With comprehensive APIs, support for bulk and programmatic submissions and full INSDC-compliant data exchange, BioSamples provides a scalable framework for long-term interoperability, extending beyond ENA to other omics archives, including proteomics and reinforcing FAIR principles [12] for findable, accessible, interoperable, and reusable metadata.

Centralising sample management in BioSamples further enhances interoperability between repositories; the BioSample accessions registered during ENA submissions can be referenced when additional data types, such as proteomics datasets in PRIDE, are generated for the same specimens, creating a scalable pathway for linking diverse multi-omics datasets.

BioSamples: a sample metadata hub

The deposition of rich metadata is essential for ensuring the findability, interoperability, and downstream reuse of data [12]. When datasets are deposited with minimal metadata, their utility is severely constrained; without adequate contextual and provenance information reliable analyses become challenging, and integration with other datasets is often impossible. The BioSamples database has evolved into the central repository for sample metadata across EMBL-EBI resources

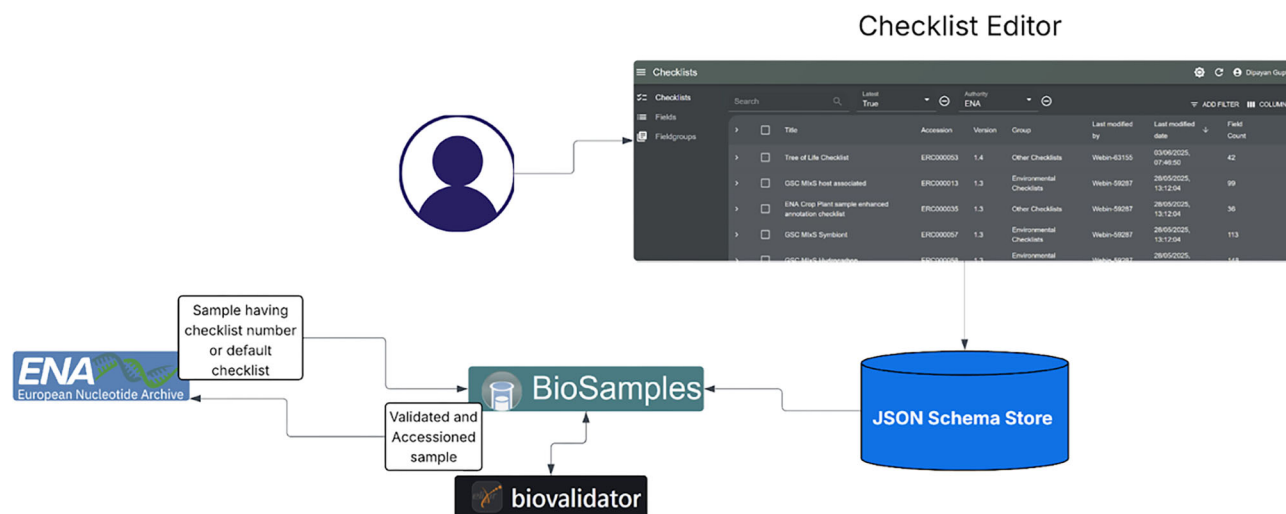


Figure 3. A new, user-friendly checklist editor has been introduced to streamline the internal management of checklists and JSON schemas. This enhances consistency and outsources the sample validation to BioSamples.

and is increasingly serving as a key hub for sample management within international consortia, community portals, and domain-specific databases. In this section, we describe several of these initiatives to illustrate how BioSamples can be leveraged and seamlessly integrated with omics resources, thereby enhancing the visibility, interoperability, and reusability of linked omics data.

PRIDE database: sample metadata annotations for proteomics data

In 2021, the proteomics community introduced SDRF-Proteomics [13], a tab-delimited format designed to systematically capture sample metadata in public proteomics archives. Beyond recording sample information, the format also enables the description of the full experimental design, including details such as sample preparation, labelling strategies, and instrument settings. Because of its flexibility and community-driven development, SDRF-Proteomics has been adopted by multiple software tools as the *de facto* standard for describing proteomics sample metadata [14,15]. The PRIDE database [5], the world's largest public repository of mass spectrometry-based proteomics data, incorporated SDRF-Proteomics as the standard way to record sample metadata during dataset submission. To ensure metadata completeness, PRIDE applies validation steps at submission based on community guidelines and standardised templates. For example, when cell line samples are used, the official cell line identifier must be provided. These community-maintained templates are integrated into PRIDE to guide researchers through the annotation process.

As part of this implementation, the first SDRF-annotated datasets submitted to PRIDE were also registered in BioSamples. Each sample annotated in SDRF received a BioSamples accession, and these identifiers were linked back into the SDRF files (e.g. <https://www.ebi.ac.uk/pride/archive/projects/PXD000652>). This created the first infrastructure to connect PRIDE datasets with BioSamples, enabling systematic cross-referencing of proteomics samples with other life science data resources. By 2025, >8000 proteomics samples have been deposited in BioSamples via PRIDE, positioning PRIDE as the largest non-genomic resource integrated with BioSamples.

This integration not only enhances the FAIRness (findability, accessibility, interoperability, and reusability) of proteomics data but also strengthens the links between proteomics and other omics domains, paving the way for more comprehensive multi-omics studies.

MorPhiC

The MorPhiC (Molecular Phenotypes of Null Alleles in Cells, <https://morphic.bio/>) project [7] is a collaborative initiative to systematically characterise gene function across human and model organism cell lines, using advanced perturbation and sequencing techniques. In one MorPhiC study, a panel of human cell lines was used to generate gene knockouts and over-expression clones for selected targets. These engineered cell lines were then differentiated, processed for library preparation, and sequenced to uncover their transcriptional profiles. To enhance the discoverability and reusability of MorPhiC datasets, curators developed automated extract, transform, and load (ETL) pipelines that parsed structured metadata from internal systems and programmatically registered each biomaterial in BioSamples using its REST APIs. The BioSamples curation API was further used to enrich sample records post-submission with cell line properties, process metadata, and ontology annotations. This API supports third-party and submitter curation, enabling ongoing improvements while retaining original submission records. Each curation event is stored with a curator ID and timestamp, ensuring traceability and version control. A public-facing, uncured view remains accessible for audit and comparison.

To align the sample metadata with community standards, BioSamples' built-in validation standards for transcriptomics samples were applied. BioSamples supports JSON Schema representations of community standards such as MIN-SEQE (<https://www.fged.org/projects/minseqe/>), enabling automated validation at the time of submission or curation. This ensured that key fields such as cell line, organism, protocol information, and sequencing type were present and consistent.

With these improvements, MorPhiC samples in BioSamples are richly annotated, ontology-linked, and readily searchable. Users can browse or query the project's sample metadata through the BioSamples portal. For exam-

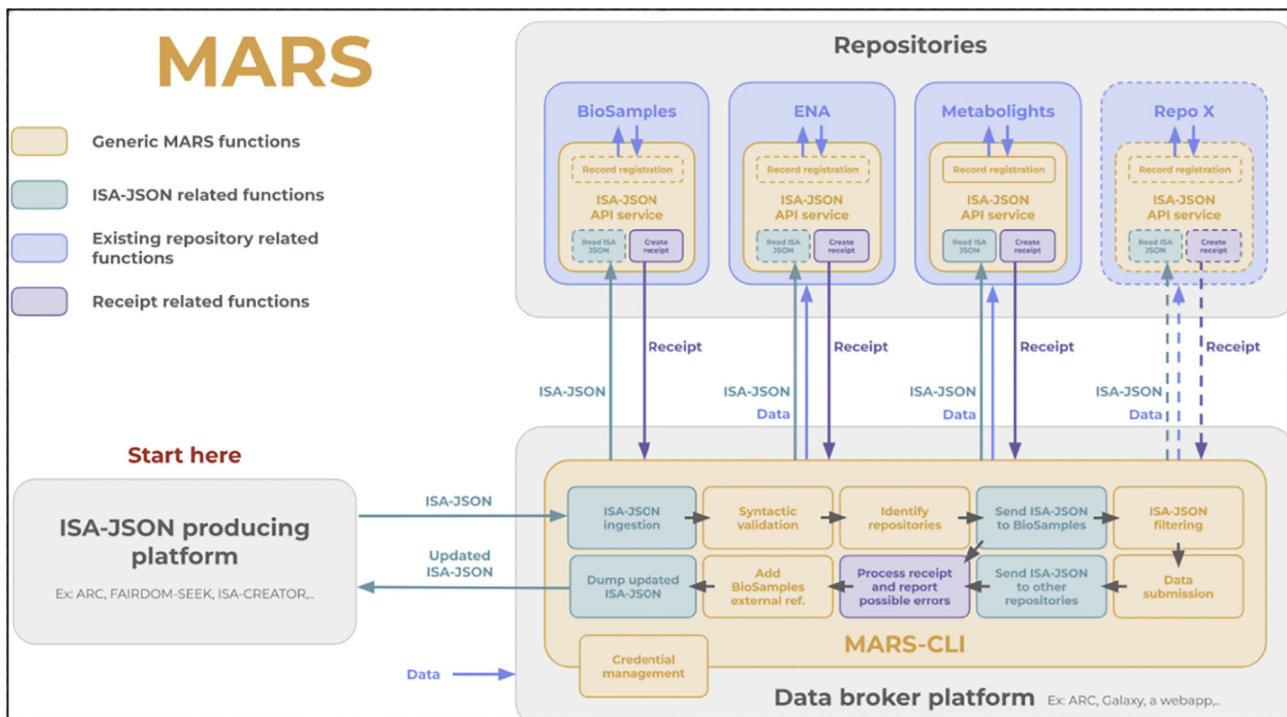


Figure 4. MARS data flow and use of BioSamples.

ple, the 1600 MorPhiC sample records can be accessed directly via <https://www.ebi.ac.uk/biosamples/samples?filter=attr:project:MorPhiC>. These enriched metadata records are available in multiple standard formats, including JSON and XML, making them suitable for integration into automated analysis pipelines and downstream discovery tools.

ReCoDID COVID-19 pilot study

The ReCoDID (Reconciliation of Cohort Data in Infectious Diseases, <https://recoqid.eu/>) COVID-19 pilot study is a research project focused on collecting and harmonising multi-omics data from 151 PCR-confirmed hospitalised COVID-19 patients. To manage and share their sample metadata effectively, the project uses BioSamples as the platform to register their biological samples. Each patient is represented as a top-level BioSample record with minimum non-identifiable metadata; this then links to child samples representing different data types such as SARS-CoV-2 genome sequences (hosted in ENA), clinical-epidemiology data (in EGA), and antibody assay results (available in ArrayExpress and BioStudies). This hierarchical organisation supports participant-level linkage across diverse datasets, enabling comprehensive integration without duplicating metadata. BioSamples provides the ReCoDID project with a standardised metadata repository that ensures findability and interoperability through controlled vocabularies, ontology annotations, and schema validation. The platform's RESTful APIs and relationship mapping allow programmatic access and browsing of linked samples, facilitating cross-repository data integration. There are currently 458 ReCoDID samples available in BioSamples: <https://www.ebi.ac.uk/biosamples/samples?filter=attr:project+name:ReCoDID%20COVID-19%20pilot%20study>. By leveraging BioSamples, the ReCoDID COVID-19 pilot study achieves consistent metadata registration and enhances

the accessibility and reusability of its multi-centre cohort data for the broader research community.

Darwin tree of life

The Darwin Tree of Life (DTOL; <https://portal.darwintreeoflife.org/>) [8] project aims to sequence the genomes of 70 000 species of eukaryotic organisms in Britain and Ireland. These genomic data are important for deepening our understanding of evolutionary processes, exploring the biology of organisms and ecosystems, supporting conservation efforts, and developing tools for medicine and biotechnology. All genome sequences generated by the project are made publicly available through the DTOL data portal (<https://portal.darwintreeoflife.org/data?0=phylogenyCurrentClass%20-%20kingdom>), which allows researchers to access and download both metadata and associated data. High-quality standardised contextual metadata were agreed upon early in the project (<https://doi.org/10.12688/wellcomeopenres.17605.1>). Metadata are collected and submitted to BioSamples by COPO, a meta(data) brokering system [16]. The metadata collected and submitted in the BioSamples repository enable the linking, filtering, and searching of data via the DTOL Portal. There are ~67 000 DTOL samples submitted to BioSamples as of August 2025; <https://www.ebi.ac.uk/biosamples/samples?filter=attr:project+name:DTOL>. Central to the portal is the submitted sample in BioSamples, which links raw and assembled sequences and associated data via sample relationships. A status tracker (<https://portal.darwintreeoflife.org/tracking?0=phylogenyCurrentClass%20-%20kingdom>) is also available to show the stage of the data generation workflow that a particular species is at. The DTOL data model has also been used for the European Reference Genome Atlas [17], whose data are available from the EMBL-EBI Biodiversity

portal (https://www.ebi.ac.uk/biodiversity/data_portal?0=ERGA&1=phylogenyCurrentClass%20-%20kingdom).

HoloFood

The HoloFood project (<https://www.holofood.eu>) [18] applies a multi-omic approach, including genomics, transcriptomics, and metabolomics, alongside comprehensive meta-data, to investigate how host–microbiome interactions affect animal production using different feed conditions. The aim is to improve animal nutrition, reduce environmental impact, and enhance animal welfare. All raw data, processed results, and associated metadata are publicly available through established open-access repositories, including the ENA, MGnify [19], MetaboLights [6], and BioSamples. To integrate data across these diverse domains, HoloFood makes extensive use of BioSamples, particularly its capacity to define sample relationships (e.g. ‘derived from’) and to present structured metadata tables for data types not easily stored elsewhere (e.g. histological and inflammatory marker data). BioSamples’ structured data tables allow the capture of detailed, domain-specific information in a machine-readable way. These tables are organised under a structured data section, where each entry has a type (e.g. Antimicrobial Resistance (AMR), Molecular Markers, Histology, Fatty Acids, Cell Viability, etc.) and an associated table of values. Each row in a table usually represents one measurement, observation, or tested parameter, with columns describing attributes such as the method, unit, interpretation, or result, e.g. <https://www.ebi.ac.uk/biosamples/samples/SAMEA112750425>. Leveraging the well-structured, richly annotated sample records in BioSamples enables the linkage and interoperability across different omics layers. This approach allows discoverability and cross-repository integration through a dedicated, user-friendly data portal (<https://www.holofooddata.org/>) [20] developed by the MGnify team. The portal makes use of the sample structure relationships in BioSamples to unify access to datasets across ENA, MGnify, and MetaboLights. By mapping derived data samples back to host organisms through consistent metadata relationships, the portal supports enhanced querying capabilities across all associated omics resources.

Discussion

We have outlined several technical improvements above designed to help researchers make better use of BioSamples. In support of the sample services of large archives such as ENA and PRIDE, BioSamples is continuously evolving to handle increasing demands. To further address the increasing volume and complexity of metadata in BioSamples, we are modernising our infrastructure. The BioSamples platform is being migrated to Kubernetes to improve scalability, reliability, and manageability. In parallel, we are upgrading our search functionality by moving from Solr to Elasticsearch, delivering faster, more flexible, and more robust metadata querying. Together, these upgrades will ensure BioSamples can continue to efficiently meet the growing needs of the life sciences community. In addition, the BioSamples documentation is being updated based on user feedback collected over the past year and migrated to a new platform, making it easier for users to find guidance on submission and retrieval of samples.

In response to feedback from consortia, BioSamples is extending the types of relationships between samples to better

capture experimental design. Newly introduced relationship categories, such as ‘negative control’ and ‘positive control’, are being added to the existing framework.

BioSamples is also enhancing its role as an authoritative platform for multi-omics and life sciences data by enabling external references to additional resources such as the BioImage Archive (BIA) [21]. This integration will facilitate connection of diverse data types - genomics, transcriptomics, proteomics, metabolomics, and imaging - across multiple EBI archives, improving interoperability and allowing researchers to navigate and access related datasets more seamlessly through a unified platform.

Finally, BioSamples is collaborating with the MARS (Multi-omics Adapter for Repository Submissions) initiative to streamline the submission of multi-omics data to multiple specialised repositories. MARS facilitates the use of the ISA-JSON format (<https://isa-specs.readthedocs.io/en/latest/isajson.html>) to standardise metadata exchange, enabling seamless data submission across various platforms, as shown in Fig. 4. Through this integration, BioSamples will serve as a main hub, linking sample metadata with other EBI archives, thereby enhancing the interoperability and accessibility of multi-omics datasets. This will contribute to better tools to enable researchers and data managers to submit, link, and manage their multi-omics experimental data more easily, aligning with the evolving needs of the research landscape.

Acknowledgements

We thank the following for collaborations: Darwin Tree of Life Consortium, Recodid Consortium, MorPhiC Consortium, Holofood Consortium, MGnify Team at EMBL-EBI, and Metaproteomics Initiative (doi: 10.1186/s40168-021-01176-w).

Author contributions: Dipayan Gupta (Software [lead], Writing—original draft [equal]), Isuru Liyanage (Software [equal], Writing—review & editing [supporting]), Yas-set Perez-Riverol (Resources [supporting], Writing—original draft [supporting], Writing—review & editing [supporting]), Enrique Sapena (Writing—review & editing [supporting]), Wei Kheng Teh (Resources [supporting], Writing—review & editing [supporting]), Tim Van Den Bossche (Resources [supporting], Writing—review & editing [supporting]), Senthilnathan Vijayaraja (Software [supporting]), Amnon Khen (Software [supporting]), Ugis Sarkans (Writing—review & editing [supporting]), Tony Burdett (Conceptualisation [lead], Funding acquisition [lead], Resources [lead], Supervision [lead], Writing—review & editing [supporting]), and Jeena Rajan (Project administration [lead], Resources [lead], Supervision [lead], Writing—original draft [lead])

Conflict of interest

None declared.

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council (grant number BB/X001911/1), in part by the Wellcome Trust (208391/Z/17/Z, 223745/Z/21/Z), in part made possible by grant number 2024-350543 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community

Foundation, and the projects received funding from the European Union's Horizon Europe research and innovation programme under grant agreement numbers 101094353 and 101059915. Additionally, the work was supported by ELIXIR, the research infrastructure for life science data, and by the European Molecular Biology Laboratory. T.V.D.B. acknowledges funding from the Research Foundation Flanders (FWO) (V439325N, 1286824N). Funding to pay the Open Access publication charges for this article was provided by European Molecular Biology Laboratory.

Data availability

The BioSamples database is freely available, and its content is distributed under the EMBL-EBI Terms of Use available at <https://www.ebi.ac.uk/about/terms-of-use>. The BioSamples code is available at <https://github.com/EBIBioSamples/biosamples-v4> and <https://doi.org/10.5281/zenodo.17304021> and distributed under the Apache 2.0 license.

References

- Courtot M, Gupta D, Liyanage I *et al*. BioSamples database: fAIRer samples metadata to accelerate research data management. *Nucleic Acids Res* 2022;50:D1500–D1507. <https://doi.org/10.1093/nar/gkab1046>
- O'Cathail C, Ahamed A, Burgin J *et al*. The European Nucleotide Archive in 2024. *Nucleic Acids Res* 2025;53:D49–D55. <https://doi.org/10.1093/nar/gkae975>
- Sarkans U, Fullgrabe A, Ali A *et al*. From ArrayExpress to BioStudies. *Nucleic Acids Res* 2021;49:D1502–D1506. <https://doi.org/10.1093/nar/gkaa1062>
- Freeberg MA, Fromont LA, D'Altri T *et al*. The European Genome-phenome Archive in 2021. *Nucleic Acids Res* 2022;50:D980–D987. <https://doi.org/10.1093/nar/gkab1059>
- Perez-Riverol Y, Bandla C, Kundu DJ *et al*. The PRIDE database at 20 years: 2025 update. *Nucleic Acids Res* 2025;53:D543–D553. <https://doi.org/10.1093/nar/gkae1011>
- Yurekten O, Payne T, Tejera N *et al*. MetaboLights: open data repository for metabolomics. *Nucleic Acids Res* 2024;52:D640–D646. <https://doi.org/10.1093/nar/gkad1045>
- Adli M, Przybyla L, Burdett T *et al*. MorPhiC Consortium: towards functional characterization of all human genes. *Nature* 2025;638:351–9. <https://doi.org/10.1038/s41586-024-08243-w>
- Darwin Tree of Life Project Consortium. Sequence locally, think globally: the Darwin Tree of Life Project. *Proc Natl Acad Sci USA* 2022;119:e2115642118.
- Rocca-Serra P, Brandizi M, Maguire E *et al*. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 2010;26:2354–6. <https://doi.org/10.1093/bioinformatics/btq415>
- Liyanage I, Burdett T, Driesbeke B *et al*. ELIXIR biovalidator for semantic validation of life science metadata. *Bioinformatics* 2022;38:3141–2. <https://doi.org/10.1093/bioinformatics/btac195>
- Karsch-Mizrachi I, Arita M, Burdett T *et al*. The international nucleotide sequence database collaboration (INSDC): enhancing global participation. *Nucleic Acids Res* 2025;53:D62–D66.
- Wilkinson MD, Dumontier M, Aalbersberg IJ *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. <https://doi.org/10.1038/sdata.2016.18>
- Dai C, Fullgrabe A, Pfeuffer J *et al*. A proteomics sample metadata representation for multiomics integration and big data analysis. *Nat Commun* 2021;12:5854. <https://doi.org/10.1038/s41467-021-26111-3>
- Dai C, Pfeuffer J, Wang H *et al*. quantms: a cloud-based pipeline for quantitative proteomics enables the reanalysis of public proteomics data. *Nat Methods* 2024;21:1603–7. <https://doi.org/10.1038/s41592-024-02343-1>
- Claeys T, Van Den Bossche T, Perez-Riverol Y *et al*. lesSDRF is more: maximizing the value of proteomics data through streamlined metadata annotation. *Nat Commun* 2023;14:6743. <https://doi.org/10.1038/s41467-023-42543-5>
- Shaw F, Minotto A, McTaggart S *et al*. COPO—managing sample metadata for biodiversity: considerations from the Darwin Tree of Life project. *Wellcome Open Res* 2022;7:279. <https://doi.org/10.12688/wellcomeopenres.18499.1>
- Mc Cartney AM, Formenti G, Mouton A *et al*. The European Reference Genome Atlas: piloting a decentralised approach to equitable biodiversity genomics. *npj Biodivers* 2024;3:28. <https://doi.org/10.1038/s44185-024-00054-6>
- Tous N, Marcos S, Goodarzi Boroojeni F *et al*. Novel strategies to improve chicken performance and welfare by unveiling host–microbiota interactions through hologenomics. *Front Physiol* 2022;13:884925. <https://doi.org/10.3389/fphys.2022.884925>
- Richardson L, Allen B, Baldi G *et al*. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res* 2023;51:D753–D759. <https://doi.org/10.1093/nar/gkac1080>
- Rogers AB, Kale V, Baldi G *et al*. HoloFood Data Portal: holo-omic datasets for analysing host–microbiota interactions in animal production. *Database* 2025;2025:baae112. <https://doi.org/10.1093/database/baae112>
- Hartley M, Kleywegt GJ, Patwardhan A *et al*. The BioImage Archive—building a home for life-sciences microscopy data. *J Mol Biol* 2022;434:167505. <https://doi.org/10.1016/j.jmb.2022.167505>