



## **D6.3 - Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis**



<b>Deliverable number:</b>	D6.3
<b>Work package:</b>	WP6 – Terrestrial Biosphere
<b>Intermediate Objective:</b>	Release of deliverable D6.3
<b>Deliverable type:</b>	<input checked="" type="checkbox"/> Document, report
	<input type="checkbox"/> Websites, patent filings, videos, etc.
	<input type="checkbox"/> Other: please specify .....
<b>Dissemination level:</b>	<input checked="" type="checkbox"/> Public
	<input type="checkbox"/> Restricted
<b>Estimated delivery (bimester):</b>	B11
<b>Actual delivery date:</b>	31/08/2024
<b>Author(s) (Partner-OU):</b>	AUTHORS: Caterina Catalano, Anna Paola Casazza, Aldo Ceriotti, Paola Cremonesi, Barbara Menin, Emanuela Pedrazzini, Pietro Roversi, Stefano Santabarbara, Bianca Castiglioni (CNR-IBBA)
<b>Reviewed by:</b>	ITINERIS Executive Board
<b>Note:</b>	

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 infrastructures”

## Table of contents

### SUMMARY

1.	<i>INTRODUCTION</i> .....	5
2.	<i>GLOSSARY</i> .....	5
3.	<i>ACTIVITIES</i> .....	6
3.1	<b>Functional and structural characterization of macromolecules</b> .....	6
3.1.1	<b>Optimization and functional characterization of recombinant protein expression in plant host</b> .....	6
3.1.2	<b>Protein and peptide production in prokaryotic and eukaryotic systems</b> .....	7
3.1.3	<b>Measurement of molecular mass distribution in solution using mass photometry</b> ..	7
3.1.4	<b>Structural characterization of macromolecules using X-ray crystallography or Cryo-EM</b> .....	8
3.2	<b>Metagenomics</b> .....	9
3.3	<b>Biomolecule production</b> .....	10
4.	<i>DATASETS AND ASSETS</i> .....	13
4.1.	<b>Description of datasets and assets</b> .....	13
4.1.1	<b>Functional and structural characterization of macromolecules</b> .....	13
4.1.1.1	<b>Optimization and functional characterization of recombinant protein expression in plant host</b> .....	13
4.1.1.2	<b>Protein and peptide production in prokaryotic and eukaryotic systems</b> .....	14
4.1.1.3	<b>Measurement of molecular mass distribution in solution using mass photometry</b>	15
4.1.1.4	<b>Structural characterization of macromolecules using X-ray crystallography or Cryo-EM</b> .....	15
4.1.2	<b>Metagenomics</b> .....	16
4.1.3	<b>Biomolecule production</b> .....	17
4.2	<b>Data and metadata format standards</b> .....	18
4.2.1	<b>Functional and structural characterization of macromolecules</b> .....	18
4.2.2	<b>Metagenomics</b> .....	19
4.2.3	<b>Biomolecule production</b> .....	19
4.3	<b>Data size</b> .....	20
4.3.1	<b>Functional and structural characterization of macromolecules</b> .....	20
4.3.1.1	<b>Optimization and functional characterization of recombinant protein expression in plant hosts</b> .....	20
4.3.1.2	<b>Protein and peptide production in prokaryotic and eukaryotic systems</b> .....	20
4.3.1.3	<b>Structural characterization of macromolecules using X-ray crystallography or Cryo-EM</b> .....	21

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

4.3.2 Metagenomics.....	21
4.3.3 Biomolecule production.....	22
4.5 Data storage.....	23
4.6. Sharing assets .....	23
4.7. Archiving and preservation .....	24
4.7.1 Publications .....	24
4.7.2 Data archives and data papers.....	24
4.7.3 Miscellaneous assets and official documents archiving .....	25
5. <i>INTERACTIONS WITH OTHER WPS OF ITINERIS</i> .....	26
5.1 WP2. Access and Fairness.....	26
5.2 WP3. Training.....	26

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

## 1. INTRODUCTION

Activity 6.12 “Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development” is in charge to the OU CNR-IBBA Milano. With the aim to reach the common main objective of creating a ITINERIS hub, the data acquisition and storage standards draw inspiration from the well-established principles and metadata standards of the IBISBA platform.

Biotechnology is considered a key enabling technology in the development of a low-impact circular economy, where the integration of biological and industrial process will allow a sustainable exploitation of renewable resources, thus safeguarding the environment. This requires the development of large-scale processes that must be able to deliver the desired goods in a robust and economically sustainable way. The development of such processes requires different skills and coordinated efforts to achieve the desired goals within a reasonable span of time. This activity will be focused on the potentiation of a research infrastructure dedicated to the development of pipelines for production of enzymes and other bioproducts through sustainable industrial processes. This is normally achieved by a trial-and-error process which may involve a long series of steps, including the identification of bioparts, their functional and structural characterization, metabolic engineering, choice of expression/production system, and process scale-up. While all these activities are time-consuming and labor-intensive, digitalization of individual steps or parts of the pipeline has an enormous potential to optimize and accelerate bioprocess development. The present activity will generate an integrated infrastructure where automation, miniaturization and data science will be thoroughly exploited to develop new paradigms in industrial biotechnology. This will require state-of-the-art equipment to generate high quality data which can be fed into innovative software and workflows to accelerate bioprocess development. The activity will sustain the potentiation and digitalization of dedicated platforms for biomolecule/microorganism discovery, characterization, and engineering, which will be fully integrated in the IBISBA European infrastructure, thus providing translational R&D services to a large community of industrial biotechnology stakeholders.

## 2. GLOSSARY

**ASSET:** Any (tangible or intangible) output of the action such as data, knowledge or information whatever its form or nature, whether it can be protected or not — that is generated in the action, as well as any rights attached to it, including intellectual property rights.

**DATASET:** A dataset is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity.

**IBISBA:** IBISBA is an Engineering Biology Research infrastructure that supports the growth of Industrial Biotechnology, combining biotechnological processes with chemical processes, translating early scientific results into prototypes, innovation and, ultimately, industrially workable solutions. IBISBA is unique in that it brings together, within a coordinated network infrastructure, services that cover the different steps in R&D project pipelines, from bioprocess conceptualization and design all the way through to pilot phase testing.

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

### 3. ACTIVITIES

In accordance with the principles derived from the IBISBA network, the ITINERIS project places a strong emphasis on ensuring seamless access to IBISBA data within its own data management standards. Building upon the foundation laid by the IBISBA Hub, the ITINERIS data management framework is designed to facilitate easy access to a diverse range of research assets pertinent to large-scale biotechnological processes. The standards encompass cataloging data, models, SOPs, samples, organisms, and publications, organized using the ISA (Investigation, Study, Assay/Analysis) recommendation for a comprehensive perspective. Embracing flexibility in content location, the ITINERIS data infrastructure supports both locally uploaded data and references to external repositories, including local and national e-infrastructures, and public archives recommended by ELIXIR. Moreover, the ITINERIS standards incorporate best practices for data set naming and the assignment of Digital Object Identifiers (DOIs) for published data assets, following the DOI format embedded in the IBISBA Hub. IBISBA obtains them from CrossRef. By explicitly considering ease of access to IBISBA data, ITINERIS endeavors to establish a data management framework that seamlessly integrates with the broader IBISBA network, fostering collaboration and knowledge-sharing across both projects.

#### 3.1 *Functional and structural characterization of macromolecules*

The activities for macromolecular functional and structural characterization are:

- Optimization and functional characterization of recombinant protein expression in plant hosts;
- Protein and peptide production in prokaryotic and eukaryotic systems;
- Measurement of molecular mass distribution in solution using mass photometry;
- Structural characterization of macromolecules using X-ray crystallography or Cryo-EM.

##### 3.1.1 *Optimization and functional characterization of recombinant protein expression in plant host*

The platform offers design strategies to produce high levels of recombinant proteins and peptides in plants as bioreactors, accompanied by an *in silico* module. The known structural features of the protein of interest (POI) are used for *in silico* modelling of the POI, and the choices of vector, target sequence, codon-usage, expression system and purification strategy are tailored to the required POI properties (*e.g.* stability, purity, or alleviation of allergenic/toxic features).

Possible protein expression strategies are tested by transient expression in plant systems. *In vivo* tests of the engineered POI are carried out using rapid, reliable transient expression systems (plant cell protoplasts, leaf agroinfiltration, yeast or bacterial expression), depending on the POI features. POI stability (half-life) and subcellular localization as well as co- and post-translational modifications are monitored by microscopy and biochemical/biophysical analysis. Optimization of recombinant POI(s) sequence(s), expression system and expression strategy is carried out by testing for POI yield, its intracellular localization/stability and monitoring host physiology and/or transcriptomics of the expression host during POI expression (for plants, this is performed in different tissues).

The best protein engineering strategy will be used for stable production of POI in plants. Vectors for transient and permanent expression containing the POI of interest are provided to the customers as the final product. For details see <https://www.youtube.com/watch?v=Lj3v6G6YTY>.

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

For protein engineering a file containing the following information is required:

- Origin of the protein (species; for plants also ecotype, landrace or cultivar);
- Protein function;
- Protein toxicity;
- DNA and protein sequence;
- Known co-translational and post-translational modifications (if available).

### 3.1.2 Protein and peptide production in prokaryotic and eukaryotic systems

Our platform offers specialized services for the purification of recombinant proteins across various expression systems, employing rigorous techniques to achieve high yields and purity. We focus on three distinct expression platforms:

1. **Bacterial Expression Systems**: leveraging *Escherichia coli* (*E. coli*) and other bacterial hosts for protein expression; employing affinity chromatography, ion exchange chromatography, and size exclusion chromatography for efficient purification; optimization of conditions for high-throughput and cost-effective protein isolation.

2. **Mammalian Cell Expression Systems**: utilizing mammalian cell lines for the expression of complex proteins; applying protein A/G chromatography, ion exchange, and hydrophobic interaction chromatography for precision purification; emphasizing the preservation of post-translational modifications and ensuring bioactivity.

3. **Plant Expression Systems**: harnessing plant-based expression for sustainable protein production; implementing differential centrifugation, membrane filtration, and chromatography for the extraction of recombinant proteins from plant tissues; addressing challenges associated with plant-specific compounds and optimizing protocols accordingly.

Our technical expertise ensures the customization of purification strategies based on the inherent properties of each expression system. These services are tailored to meet the demands of research, biopharmaceutical development, and industrial applications, providing researchers and biotechnologists with a reliable resource for advancing their protein-related projects.

### 3.1.3 Measurement of molecular mass distribution in solution using mass photometry

Mass photometry (MP) is a cutting-edge technique based on light-scattering enabling detection of individual, unlabeled molecules in dilute solutions. Its broad measurement range make it an essential technique for researchers studying complex biological molecules.

The technique offers precise insights into the molecular mass distribution of various biomolecules. The core of the MP instrument is a mass photometer, a sophisticated device that measures the light scattering of individual molecules as they absorb onto a glass microscope slide. This allows for the detection of single molecules without the need for labels, providing a direct and accurate analysis of molecular masses.

The MP instrument boasts an impressive measurement range, accurately determining molecular masses from 40 kDa to 5 MDa. This wide range makes it an invaluable tool for studying a variety of biomolecules, including proteins, nucleic acids, membrane proteins, and aggregates, as well as large protein complexes. Its versatility in application underscores its significance in molecular biology and biochemistry research.

To conduct a measurement using MP, a straightforward procedure is followed that ensures accurate results. Beginning with sample preparation, all materials and samples are prepared in advanced to be ready for analysis. Once prepared, the sample is introduced to the mass

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

photometer, where the light scattering of individual particles is measured. The data collected during this phase are then analyzed to calculate the molecular mass distribution. This is achieved using calibrants of known mass, which serve as a reference to ensure the accuracy of the results.

### *3.1.4 Structural characterization of macromolecules using X-ray crystallography or Cryo-EM*

**Structural Biology:** Our platform offers comprehensive support for structure determination of macromolecules, utilizing advanced X-ray crystallography and/or Cryo-EM tools and protocols.

**Structure Determination from X-ray Diffraction:** we combine commercially available crystallization screens, a state-of-the-art crystallization robot, precision instruments for crystal handling, and well-established cryoprotection methods. These resources ensure high-quality crystal formation and preservation. We collect X-ray diffraction data through partnerships with leading synchrotron facilities. Every step of the process—from sample preparation and shipping to data processing and analysis—is managed to produce reliable structural data, utilizing both experimental phasing and molecular replacement methods. The process is further supported by advanced computational tools and hardware, ensuring accurate and efficient determination of macromolecular structures.

For structures requiring experimental phasing, we utilize the Global Phasing software suite, including SHARP and Solomon, implemented within the automated pipeline autoSHARP. This approach is particularly valuable for novel structures or when anomalous data has been collected, such as from samples containing anomalous scatterers. The autoSHARP pipeline streamlines the process of identifying phase information from experimental data, allowing us to efficiently solve the phase problem and generate initial electron density maps, which are then refined iteratively to produce accurate structural models.

For cases where a homologous structure is available, we use molecular replacement tools such as CCP4 Molrep or Phaser. These tools enable rapid structure determination by fitting a known model into the experimental data. Molecular replacement is particularly useful for well-characterized protein families or when high-quality diffraction data are available, allowing us to produce accurate models without the need for experimental phasing.

Once the initial structural model is obtained—whether through experimental phasing or molecular replacement—it undergoes refinement; we use Global Phasing autoBUSTER and Phenix refine to ensure that the final model fits the experimental data with minimal errors. This refinement process involves optimizing atomic coordinates and adjusting the model to improve its fit with the electron density, which is crucial for producing a high-quality structure that accurately represents the macromolecule.

All structure determination calculations are performed on our two-GPU server, which provides the necessary computational power for handling refinement and phasing tasks. This server also stores all project-related files, ensuring seamless data management throughout the structure determination process. The complete structural determination protocol, along with the final coordinates in PDB format, is archived and securely stored, ensuring that all aspects of the process are traceable and reproducible. Once the structure is finalized, the protocol and PDB files are shared with the user, providing both the refined model and the details of the procedures used to determine it.

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

Cryo-EM: for Cryo-EM, the preparation of macromolecular samples at our facility starts with purification and vitrification steps. The final purification step for Cryo-EM samples is always size exclusion chromatography (SEC), performed on our ÄKTA Micro system. This system is optimized to minimize sample dilution, eliminating the need for post-SEC concentration steps. The purification protocol is customized for each sample in advance, ensuring the highest purity and concentration levels are achieved before vitrification.

For vitrification and data collection, we collaborate with free-access external facilities such as the one at the Milano Human Technopole, University of Milano Facilities, and the FLOCEN facility in Florence. The vitrification process is carried out using a Vitrobot, ensuring the rapid freezing of samples into a vitreous state. We work closely with facility personnel to optimize grid preparation, with preliminary screening often involving negatively stained samples to assess grid quality before Cryo-EM data collection. Our facility relies on advanced instrumentation at partner facilities and integrates high-performance computational tools to achieve high-resolution structural data.

### 3.2 *Metagenomics*

The infrastructure provides scientific and technical support for the metagenomic and/or metabarcoding analyses of different environmental samples (soil, water, gut, food, etc).

A fully operational genome facility with two MiSeq Illumina Next Generation Integrated systems, Oxford Nanopore GridION for Real-Time DNA/RNA Sequencing, Droplet Digital PCR, QuantStudio Real-Time PCRs, computer farm equipment for data analysis and storage is available.

Metabarcoding analyses consists in advanced microbial community analysis through next generation sequencing/massively parallel approaches directly from environmental samples. Metabarcoding using targeted amplicons such as 16S rRNA genes is the most popular method for characterizing microbial communities of interest. Other amplicon targets can be used for other groups of microbes, such as ITS for fungi, 18S rRNA for eukaryotes, or RdRP for RNA viruses.

Shotgun metagenomics is the approach used to analyse all the genes present in a microbial community, providing not just phylogenetic analysis but also insight into the functional capabilities of each species within the community.

Long read metagenomics has the potential to combine the best properties of both shotgun and metabarcoding approaches. The long reads produced mean that amplification is not needed because whole target regions (e.g., 16S rRNA genes) are frequently recovered intact. It is also often possible to recover intact genomes of microorganisms and plasmids, potentially gathering valuable phylogenetic and virulence data.

Metagenomics analyses include:

1) DNA extraction and quality control: DNA is extracted directly from samples using mostly commercial kit with magnetic beads to specifically capture DNA while excluding organic inhibitors. This type of kit has a good balance of DNA yield and quality, as demonstrated on a variety of environmental sample types. An internal protocol based on the use of silica particles, with a chaotropic agent for cell lysis is used for some types of food samples. The qualitative analysis of the extracted DNA is performed by spectrophotometer (Nanodrop ND-1000).

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

- 2) Library preparation: For shotgun metagenomics, whole DNA is prepared for sequencing using the Illumina Nextera XT DNA library prep kit. Samples are barcoded and mixed together for sequencing. For metabarcoding libraries are prepared following the 16S Metagenomic Sequencing Library Preparation Protocol (Illumina, San Diego, CA, USA).
- 3) Shotgun metagenomics: The libraries are sequenced using Illumina technologies for paired end reads. The number of samples multiplexed depends on the sample type. Bioinformatic analysis make it possible taxonomic identification, gene findings and annotation, metabolic reconstruction.
- 4) 16S rRNA and other target genes sequencing by Next generation sequencing using Illumina technology (Metabarcoding): The libraries obtained are quantified by Real-Time PCR with KAPA Library Quantification Kits (Kapa Biosystems, Inc., MA, USA), pooled in equimolar proportion and sequenced by MiSeq (Illumina) instrument with 2×250-base paired-end runs. Bioinformatics analyses allow microorganisms (bacteria, yeast, fungi) identification.
- 5) Long read metagenomics sequencing in real time by Oxford Nanopore GridION (Metagenomics). Bioinformatic analysis make it possible taxonomic identification, gene findings and annotation, metabolic reconstruction.

#### Sequencing Platform technologies:

Reliable and high quality NGS data is a key component in metagenomics research. Illumina Technologies is based on short-read technologies carrying out sequencing by synthesis. Single-stranded DNA-binding proteins are used for amplification, followed by the addition of fluorescent-labelled deoxynucleoside triphosphates to bridge the amplified DNA template.

Oxford Nanopore Technologies offer real-time, long-read, direct, and large-scale sequencing of DNA or RNA. The GridION system sequences DNA or RNA by Nanopore reader proteins embedded in an electrically resistant membrane. Ultra-long reads span repeat regions in complicated genomes easily and enhance the accuracy of genome assembly and large structural variation detection enormously.

The advancement of modern sequencing technologies has revealed the existence of many yet unknown species and highlighted the tremendous complexity of microbiomes. Nevertheless, cultivation-based studies are essential because they facilitate proper taxonomic description of novel microbes and open avenues for downstream functional studies. For this purpose, our infrastructure is able to efficiently separate single cells from complex microbial communities into 96-well plates using a bacterial single cell sorter enable of directly generating clonal cultures (B.SIGHT™, Cytena).

### **3.3 Biomolecule production**

The biomolecule production platform is designed with a view to allow the implementation of sustainable and efficient biotechnological processes for the synthesis of high-value biomolecules in wild-type, metabolically engineered microbial strains and plant cell lines.

The platform is equipped with all the necessary instrumentation for i) the isolation and characterization of microbial strains, ii) the optimization of biomass accumulation of the - including a high-processivity microbioreactor system for rapid screening of up to 48 different strains/growth conditions, iii) the scale up of photoautotrophic processes, and also includes iv) preparative and analytical instruments for the extraction, analysis and quantification of the obtained metabolites.

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

The platform primarily aims at optimizing biotechnological processes based on photosynthetic microorganisms and is equipped for the validation and scale up (lab-scale photobioreactors) of the production processes.

The biomolecules platform offers the following expertise:

1) *Design, generation, validation and characterization of metabolically engineered microbial strains for the production of high-value metabolites.* This task is developed by means of recombinant DNA technologies to enhance the metabolic flux of the cells toward the production of target compounds or achieve heterologous production of non-endogenous metabolites of interest.

2) *Optimization of key parameters of microbial growth processes (e.g. pH, T, dissolved nutrients, light if photoautotrophic.) in batch assays and bioreactor systems.* Process optimisation is a key step in the development of sustainable and efficient biotechnological processes. The metabolite platform boasts cutting-edge instrumentation for high-throughput screening of microbial strains and plant cell lines in a microbioreactor system, the BioLector XT (Beckman Coulter). The Biolector XT is a microfluidic-scale reactor capable of working in both batch and fed-batch modes for simultaneous screening of up to 48 different strains or culture conditions, featuring versatile modules for working with light, anaerobic conditions or modulated oxygen levels. This setup drastically reduces the time required for the optimisation of both photosynthetic and fermentation processes, by real-time evaluation of critical cultivation parameters. These include monitoring biomass, fluorescence, pH and dissolved oxygen (DO) in the liquid phase, making it suitable for both aerobic and anaerobic cultures. Key features of the BioLector XT include:

- online, pre-calibrated optical sensors for accurate measurements;
- an optional microfluidic module for precise pH control and nutrient feeding;
- an innovative gassing head designed specifically for anaerobic experiments;
- working volume ranging from 800 to 2400  $\mu\text{L}$ .

Applications of this advanced microbioreactor are broad, encompassing process development and optimization, as well as conducting scale-down studies.

In addition, for lab-scale screening experiments of phototrophic organisms, such as algae, cyanobacteria and plant cells, the platform is equipped with a Multi-Cultivator MC 1000-OD reactor system (PSI), equipped with a gas mixer and a cooling unit. This system is specifically designed to enable rapid comparison studies of various organisms, mutants or cultivation conditions. Notably, the system can be applied for the cultivation of also other type of microorganisms, for example heterotrophic bacteria and yeast grown in suspension without the light setting. Key features of the Multi-Cultivator MC 1000-OD include:

- Cultivation of autotrophic and heterotrophic microorganisms grown in suspension;
- Parallel or multi-variant screening experiments in 8 independent vessels;
- Independent illumination and OD monitoring (680 and 720 nm) for each vessel;
- Color illumination in multi- or mix- version (up to 8 colors in each cultivation slot);
- Uniform temperature and aeration control for all vessels;
- Online software control.

This comprehensive approach ensures efficient and effective production of valuable biomolecules, leveraging the full potential of microbial metabolic capabilities.

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

3) *Extraction, chemical analysis and quantification of metabolites produced.*) the platform is equipped with a cell homogenizer for the efficient disruption of different cell types, necessary for the recovery of the metabolites of interest. Also available are benchtop ultra/centrifuge systems for sample preparation. Regarding the analysis of the metabolites produced, the platform is equipped with the necessary instrumentation for fine characterization based on chromatographic analysis and optical spectroscopy. Preparative and analytical HPLC instrumentation is available for purification, analysis and quantification of target compounds. while in regard to the optical characterization, the platform includes a spectrofluorimeter (UV-Vis) and a customized fluorimeter equipped with a CDD camera for higher resolution ( $\lambda > 600$  nm) for the study of molecules which emission extends up to the near-infrared.

4) *Process scale-up at lab-scale in photobioreactor system (available only for biotechnological processes based on photosynthetic microorganisms).* For scaling up to TRL3 of algal biotechnological processes, the platform is equipped with a 25L horizontal tubular photobioreactor, namely the Lgem Lab-25 tubular glass PBR. The unit comprises one vertical tubular helix made from transparent glass pipes, a glass circulation vessel and a dimmable LED illumination system.

## 4. DATASETS AND ASSETS

In accordance with the principles derived from the IBISBA network, the ITINERIS project places a strong emphasis on ensuring seamless access to IBISBA data within its own data management standards. Building upon the foundation laid by the IBISBA Hub, the ITINERIS data management framework is designed to facilitate easy access to a diverse range of research assets pertinent to large-scale biotechnological processes. The standards encompass cataloging data, models, SOPs, samples, organisms, and publications, organized using the ISA (Investigation, Study, Assay/Analysis) recommendation for a comprehensive perspective. Embracing flexibility in content location, the ITINERIS data infrastructure supports both locally uploaded data and references to external repositories, including local and national e-infrastructures, and public archives recommended by ELIXIR. Moreover, the ITINERIS standards incorporate best practices for data set naming and the assignment of Digital Object Identifiers (DOIs) for published data assets, following the DOI format embedded in the IBISBA Hub. IBISBA obtains them from CrossRef. By explicitly considering ease of access to IBISBA data, ITINERIS endeavors to establish a data management framework that seamlessly integrates with the broader IBISBA network, fostering collaboration and knowledge-sharing across both projects.

### 4.1. Description of datasets and assets

Data and assets generated within the IBISBA platform within the ITINERIS framework will mainly come from experiments belonging to its four strands; additionally, data will come from desk studies, minutes of meetings, brochures and visual media. Some data will report on experimental analyses and other assets will take the form of experimental workflows. Finally, some assets will result from external services. These may be linked to personal data and thus subject to GDPR.

#### 4.1.1 Functional and structural characterization of macromolecules

##### 4.1.1.1 Optimization and functional characterization of recombinant protein expression in plant host

The optimization and functional characterization of recombinant protein/peptide expression in plant hosts involve a variety of experimental techniques and analyses, generating several types of datasets to assess the efficiency, yield, and functionality of the expressed proteins. The datasets typically provided by such studies are:

1. Genetic Constructs and Sequence Data: sequences of the recombinant genes or genetic constructs used for expression in plant hosts. This includes the coding sequence of the target protein, promoters, terminators, and any other regulatory elements.
2. Plant Transformation and Transgene Integration Data: data confirming successful transformation of plant hosts with the genetic constructs.
3. Expression Analysis Data: quantitative and qualitative data on the expression levels of the recombinant protein in plant tissues. This includes mRNA expression levels (RT-PCR), protein expression levels (SDS-PAGE, Protein blotting), and protein quantification (ELISA), often across different plant tissues or developmental stages.
4. Cell Biology Data: data from cell biology experiments assessing the protein localization by microscopy or subcellular fractionation by isopycnic ultracentrifugation on sucrose gradient
5. Biochemical and Biophysical Assay Data: data from biochemical and biophysical assays assessing the polymerization grade, N or O- glycosylation, enzymatic activity, binding

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

affinity, stability, protein-protein interactions, and other functional properties of the recombinant protein. This may include velocity ultracentrifugation on sucrose gradient; enzyme assays, Co-immunoprecipitation; binding assays (e.g., surface plasmon resonance), thermal stability assays, and other relevant biochemical assays.

6. Protein Purification and Characterization Data: data on the purification and characterization of the recombinant protein. This includes SDS-PAGE gels showing purity and molecular weight, mass spectrometry for protein identification and confirmation, and spectroscopic analyses for structural and functional characterization.
7. Plant Phenotypic and Growth Data: data on the growth characteristics and phenotypic traits of the plant hosts expressing the recombinant protein. This includes growth rates, biomass accumulation, and any observable phenotypic changes associated with recombinant protein expression.
8. Statistical and Computational Analysis Outputs: results from statistical analyses of experimental data, such as significance tests, correlation analyses, and data normalization. Computational modeling outputs may include structural predictions, molecular dynamics simulations, or sequence analysis results.
9. Experimental Metadata: metadata documenting experimental details, including plant species and cultivars used, growth conditions (e.g., light intensity, temperature, humidity), sampling time points, and any treatments or experimental variables.

#### **4.1.1.2 Protein and peptide production in prokaryotic and eukaryotic systems**

The production of proteins and peptides in prokaryotic (such as bacteria) or eukaryotic mammalian cells involves a range of experimental techniques and analyses. The datasets generated from such studies typically include:

1. Genetic Constructs and Sequence Data: sequences of the genes or genetic constructs used for protein or peptide production. This includes the coding sequence of the protein or peptide of interest, promoters, terminators, and any other regulatory elements.
2. Transformation or Transfection Confirmation Data: data confirming successful transformation (in prokaryotic systems) or transfection (in eukaryotic systems) with the genetic constructs. This includes PCR validation of the presence and integrity of the recombinant DNA.
3. Expression Analysis Data: quantitative and qualitative data on the expression levels of the protein or peptide of interest. This includes protein expression levels (Western blotting), protein quantification (ELISA), and possibly mRNA expression levels (RT-PCR) in the case of eukaryotic systems.
4. Protein or Peptide Purification Data: data on the purification of the expressed protein or peptide, showing purity and molecular weight (SDS-PAGE), as well as chromatographic profiles demonstrating separation and yield.
5. Biochemical and Biophysical Assay Data: data from biochemical and biophysical assays assessing the enzymatic activity, binding affinity, stability, molecular mass distribution in solution and other functional properties of the produced protein or peptide. This may include enzyme assays, binding assays (e.g., SPR), thermal stability assays, and other relevant biochemical assays.

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

6. Structural Characterization Data: data from structural analyses to determine the three-dimensional structure of the protein or peptide. This includes NMR spectra showing chemical shifts and NOE patterns, or X-ray crystallography data in PDB format.
7. Experimental Metadata: metadata documenting experimental details, including the type of expression system used (prokaryotic or eukaryotic), culture conditions, inducer concentrations (if applicable), sampling time points, and any experimental variables or treatments applied.

#### *4.1.1.3 Measurement of molecular mass distribution in solution using mass photometry*

The output data from a mass photometer typically includes the following:

1. Mass Distribution: A histogram or plot showing the distribution of molecular masses in the sample. This provides insights into the heterogeneity of the molecules.
2. Mass Values: Precise mass measurements for individual molecules. These values are obtained without labeling or modifying the sample.
3. Intensity vs. Mass Curve: A graph showing the interference signal (intensity) as a function of molecular mass. The slope of this curve allows accurate mass determination.
4. Single Molecule Information: Since mass photometry detects individual molecules, you get information about each molecule's mass.

#### *4.1.1.4 Structural characterization of macromolecules using X-ray crystallography or Cryo-EM*

The structural characterization of macromolecules using X-ray crystallography or Cryo-EM (Cryo-Electron Microscopy) involves several types of datasets that provide detailed insights into the three-dimensional structures of biological molecules. The datasets typically generated in such studies are:

##### 1. X-ray Crystallography Data:

- Diffraction Images: these are the raw diffraction images obtained from the X-ray diffraction experiments. They capture the scattering of X-rays by the crystallized macromolecule and are processed to generate structure factors.
- Structure Factors: structure factors are derived from the diffraction images and contain the amplitudes and phases of the diffracted waves. They are essential for determining the electron density map and subsequently the atomic model of the macromolecule.
- Refinement Parameters and Models: these files contain the refined atomic coordinates and associated parameters (e.g., B-factors) that describe the structure of the macromolecule. They are derived from iterative refinement against the experimental data.

##### 2. Cryo-EM Data:

- Micrographs: micrographs are the raw images obtained from the Cryo-EM experiments. They contain images of the macromolecule particles embedded in vitreous ice.

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

- **Particle Picks:** these files contain coordinates of the particle positions manually picked or automatically detected from the micrographs. They are used for subsequent image processing steps.
- **2D Class Averages:** class averages represent averaged images of particles assigned to similar classes based on their two-dimensional projections. They help assess the homogeneity of the particle population.
- **3D Density Maps:** these maps represent the three-dimensional electron density distribution of the macromolecule obtained through Cryo-EM reconstruction. They are crucial for visualizing the overall shape and structure of the molecule.
- **Atomic Models:** similar to X-ray crystallography, atomic models derived from Cryo-EM data represent the refined atomic coordinates of the macromolecule, fitted into the Cryo-EM map.

3. **Validation and Quality Control Data:** reports summarizing the validation metrics of the atomic models, such as resolution, R-factor, and stereochemical quality assessments (e.g., Ramachandran plot).

4. **Experimental Metadata:** metadata documenting experimental details, including sample preparation methods (crystallization conditions for X-ray crystallography or specimen preparation for Cryo-EM), data collection parameters (e.g., exposure times, radiation dose for Cryo-EM), and processing parameters.

#### 4.1.2 *Metagenomics*

The metagenomic platform includes a coordinated system for bacterial single-cell isolation and DNA library preparation, along with a high-throughput sequencing system, and generates several types of datasets. These datasets provide a comprehensive view of the bacterial community's genetic and functional landscape, enabling detailed analysis of microbial diversity, gene content, and potential functional capabilities.

The datasets from such a platform are:

1. **Raw Sequencing Reads:** these files contain the raw sequencing reads generated by the high-throughput sequencing system, including nucleotide sequences and associated quality scores.
2. **Quality Control Reports:** these reports provide an assessment of the quality of the sequencing reads, including metrics such as read length distribution, base quality scores, GC content, and the presence of adapter sequences.
3. **Preprocessed Reads:** these files contain reads that have been preprocessed to remove low-quality bases, adapter sequences, and other contaminants. This step often includes trimming and filtering of sequences based on quality thresholds.
4. **Assembled Genomes or Contigs:** after assembling the sequencing reads, the resulting contigs or draft genomes are provided. These represent the reconstructed sequences of bacterial genomes or large genomic fragments.
5. **Binned and Taxonomic Assignments:** these files include information on the taxonomic classification of the assembled sequences, often based on sequence similarity to known

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

databases. They may also include binning results, where sequences are grouped into bins representing different bacterial species or strains.

6. Genetic and Functional Annotations: annotations of genes and other functional elements within the assembled sequences, often including gene predictions, protein coding sequences, and functional annotations (e.g., using tools like Prokka or RAST).
7. Metadata: metadata associated with the samples, including information on the source of the samples, isolation conditions, and any experimental treatments applied.
8. Single-Cell Data: data specific to single-cell isolation, including information on the identity of single cells, barcodes used for indexing, and cell-specific sequencing reads.
9. Assembly and Binning Statistics: statistics summarizing the assembly and binning process, such as the number of contigs, N50 value, total assembled length, and completeness and contamination estimates for each bin.
10. Functional Profiles: functional profiling data, such as the abundance of different functional genes or pathways within the metagenome, often generated using tools like HUMAN or KEGG.

#### 4.1.3 Biomolecule production

For biomolecule production, using a high-throughput microbioreactor, bacterial and algal cultivation systems and a related equipment, several types of datasets are expected. These datasets encompass various aspects of the cultivation process, including environmental conditions, biological responses, and product yields. Here's an overview of the datasets:

1. Environmental Monitoring Data: data on the environmental conditions within the microbioreactor or cultivation systems, including temperature, pH, dissolved oxygen, CO<sub>2</sub> concentration/consumption, light intensity (for algal cultures), and agitation speed. This data is typically collected in real-time or at regular intervals.
2. Biomass and Growth Data: measurements of biomass concentration over time, such as optical density (OD), cell counts, dry weight, and chlorophyll content (for algal cultures). This data helps track the growth kinetics of the cultures.
3. Nutrient and Metabolite Concentrations: concentrations of key nutrients (e.g. carbon source, nitrate, phosphate) and metabolites (e.g. specific biomolecules of interest released from the cells) in the culture medium. This data can be collected through HPLC or other analytical techniques.
4. Genomic and Transcriptomic Data: data from high-throughput sequencing of DNA or RNA to analyze the genetic and transcriptomic profiles of the microbial or algal cultures. This includes raw sequencing reads, assembled genomes, gene expression profiles, and annotations.
5. Product Yield and Purity Data: quantitative data on the yield and purity of the target biomolecules produced, such as high-value metabolites (e.g. carotenoids, phenylpropanoids, etc.) or precursors of biofuels and bioplastics. These data are often obtained through HPLC, or specific assays.
6. Image Data: microscopic images of the cultures to monitor cell morphology, aggregation, and other phenotypic characteristics, cell counts and detection of potential contaminants. These data are often analyzed using image analysis software.

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

7. Experimental Metadata: metadata describing the experimental setup, including details about the microbial or algal strains used, culture medium composition, inoculation density, and any experimental treatments applied.
8. Quality Control and Calibration Data: data related to the calibration and validation of sensors and instruments used in the cultivation systems, ensuring accurate and reliable measurements.

## 4.2 Data and metadata format standards

Within the context of the ITINERIS project, the data acquisition and storage standards draw inspiration from the well-established principles and metadata standards of the IBISBA platform. In alignment with the SEEK platform that underpins the IBISBAHub, metadata adheres to minimal recommendations ensuring comprehensive documentation. The standards include essential metadata such as Uploader Name, IBISBA ID, Project name, Upload Date, Creation Date, Title, and Version Number. Notably, the Uploader Name serves as the recognized representative of the organization(s) that owns the data/assets, with the assumption that the named person has the rights to upload data and set sharing permissions. The IBISBA ID uniquely identifies individuals within the IBISBA Hub environment, utilizing ORCID for academics. The IBISBA project serves as the default project name, and assets are tracked based on their Upload Date, distinguishing versions through a Version Number incrementation system. To credit co-workers and collaborators, the metadata includes a field for the names of other people involved. Emphasizing ease of access for the ITINERIS project, the integration of IBISBA data is facilitated through adherence to these standards, ensuring a seamless exchange of information and collaborative utilization of the shared infrastructure.

All data and metadata of IBBA's RI are generated in accordance with FAIR principles, using the recommended file formats (accepted standards).

### 4.2.1 Functional and structural characterization of macromolecules

#### File Formats

The data are generated using the following recommended file formats:

- Genetic Constructs and Sequence Data: FASTA, GenBank or similar
- Plant Transformation and Transgene Integration Data: PCR results (CSV), sequencing data (chromatograms, FASTA)
- Transformation or Transfection Confirmation Data: PCR results (CSV), sequencing data (chromatograms, FASTA)
- Expression Analysis Data: CSV, TIFF, JPEG
- Protein or Peptide Purification Data: TIFF, JPEG, CSV
- Protein Characterization Data: MGF, mzML, UV-Vis spectra
- Structural Characterization Data: Bruker, Varian, etc., PDB files
- Cell Biology data: TIFF, CSV
- Biochemical and Biophysical Assay Data: TIFF, JPEG, CSV
- Plant Phenotypic and Growth Data: CSV, TIFF, JPEG
- Cellular Localization and Interaction Data: TIFF, JPEG, CSV
- Statistical and Computational Analysis Outputs: CSV, PDF
- Expression Analysis Data: TIFF, JPEG, CSV

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

Mass photometry Data:

- A list of events in CSV format.
- A Discover MP-generated report in PDF, CSV, or JSON formats.
- The movie in MP4 format.
- Scores, or data on brightness, motion, sharpness, saturation, and signal, can be exported as an .h5 file.
- Raw frames in .h5 format.

X-ray Crystallography Data:

- Diffraction Images: HDF5, CBF, MTZ, CIF.
- Structure Factors: CIF, MTZ
- Refinement Parameters and Models: CIF, PDB (Protein Data Bank) files
- Cryo-EM Data:
- Micrographs: MRC
- Particle Picks: STAR, CSV
- 2D Class Averages: MRC
- 3D Density Maps: MRC, CCP4
- Atomic Models: CIF, PDB files.

Validation and Quality Control Data Reports: PDF, text files.

Experimental Metadata: CSV, JSON, or database file

#### 4.2.2 *Metagenomics*

The metagenomic data are generated using the following recommended file formats:

- Raw Sequencing Reads: FASTQ files
- Quality Control Reports: FastQC
- Preprocessed Reads: FASTQ or FASTA files
- Assembled Genomes or Contigs: FASTA files
- Binning and Taxonomic Assignment: Text or tab-delimited files
- Gene and Functional Annotations: GFF, GTF, or tab-delimited files
- Metadata: Tab-delimited or JSON files
- Single-Cell Data: Tab-delimited files or specialized formats like HDF5
- Assembly and Binning Statistics: Text or tab-delimited files
- Functional Profile: Tab-delimited files

#### 4.2.3 *Biomolecule production*

The data on biomolecule production are generated using the following recommended file formats:

- Environmental Monitoring Data: CSV, JSON
- Biomass and Growth Data: CSV
- Nutrient and Metabolite Concentrations: CSV
- Genomic and Transcriptomic Data: FASTQ, BAM
- Product Yield and Purity Data: CSV
- Image Data: TIFF, JPEG, or similar image file formats

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

- Experimental Metadata: CSV, JSON
- Quality Control and Calibration Data: CSV

### 4.3 Data size

#### 4.3.1 Functional and structural characterization of macromolecules

##### 4.3.1.1 Optimization and functional characterization of recombinant protein expression in plant hosts

For a comprehensive study involving multiple genetic constructs, plant species, and extensive characterization, the volume of data generated in the optimization and functional characterization of recombinant protein expression in plant hosts can vary based on the scale and complexity of the experiments. This estimation considers the accumulation of data across various stages of the experimental workflow, including genetic design, transformation, expression analysis, protein purification, functional characterization, and statistical analysis.

- Genetic Constructs and Sequence Data: Small to moderate (MBs to GBs).
- Plant Transformation and Transgene Integration Data: Small to moderate (MBs to GBs).
- Expression Analysis Data: Moderate (MBs to GBs).
- Protein Purification and Characterization Data: Moderate to large (GBs).
- Cell Biology data: Moderate to large (GBs).
- Biochemical and Biophysical Assay Data: Moderate to large (GBs).
- Plant Phenotypic and Growth Data: Moderate (MBs to GBs).
- Statistical and Computational Analysis Outputs: Small to moderate (MBs to GBs).
- Experimental Metadata: Small (MBs).

##### 4.3.1.2 Protein and peptide production in prokaryotic and eukaryotic systems

The volume of data generated in protein and peptide production studies can vary based on the complexity and scale of the experiments. For a comprehensive study involving multiple expression systems (prokaryotic and eukaryotic), genetic constructs, and extensive characterization, the total data volume could range from several hundred megabytes (MB) to several gigabytes (GB). Studies focusing on structural characterization may generate larger datasets, particularly in the case of NMR or X-ray crystallography data. This estimation considers data accumulation across various stages of the experimental workflow, including genetic design, expression confirmation, protein purification, functional and structural characterization, and experimental metadata documentation.

- Genetic Constructs and Sequence Data: Small to moderate (MBs to GBs).
- Transformation or Transfection Confirmation Data: Small to moderate (MBs to GBs).
- Expression Analysis Data: Moderate (MBs to GBs).
- Protein or Peptide Purification Data: Moderate (MBs to GBs).
- Biochemical and Biophysical Assay Data: Moderate to large (GBs).
- Structural Characterization Data: Large (GBs to TBs), especially for Cryo-EM.
- Experimental Metadata: Small (MBs).

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

#### 4.3.1.3 Structural characterization of macromolecules using X-ray crystallography or Cryo-EM

The volume of data generated in structural characterization studies using X-ray crystallography or Cryo-EM can vary widely based on the size and complexity of the macromolecule, the resolution desired, and the number of experimental conditions. For a comprehensive structural characterization study involving both X-ray crystallography and Cryo-EM, the total data volume can range from several hundred gigabytes (GB) to several terabytes (TB). Cryo-EM datasets, particularly micrographs and 3D density maps, tend to be larger due to the nature of the technique and the resolution requirements. X-ray crystallography datasets are smaller in comparison but still significant in size, especially when considering raw diffraction images and refined atomic models. Here's a rough estimation:

- X-ray Crystallography:
    - Diffraction Images: Hundreds of megabytes (MB) to gigabytes (GB) per dataset.
    - Structure Factors: Typically, small, few megabytes (MB) per dataset.
    - Refined Models: Several megabytes (MB) per dataset.
  - Cryo-EM:
    - Micrographs: Hundreds of gigabytes (GB) to terabytes (TB) per dataset.
    - Particle Picks: Few megabytes (MB) to gigabytes (GB) per dataset.
    - 2D Class Averages: Gigabytes (GB) per dataset.
    - 3D Density Maps: Gigabytes (GB) to tens of gigabytes (GB) per dataset.
    - Atomic Models: Several megabytes (MB) to gigabytes (GB) per dataset.
- Validation and Quality Control Data: Small to moderate, few megabytes (MB) per dataset.  
Experimental Metadata: Small to moderate, few megabytes (MB) per dataset.

#### 4.3.2 Metagenomics

- Sequencing output. For high-complexity samples, sequencing 10-20 Gb per sample is common, leading to 10-100 GB of data per sample.
- Multiple Samples: For a study with 10-100 samples, the total data size can range from 100 GB to several terabytes (TB).
- Quality Control Reports:
  - Single Sample: These are generally small, around 1-10 MB per sample.
  - Multiple Samples: For 100 samples, this amounts to around 100 MB to 1 GB.
- Preprocessed Reads:
  - Single for a metagenomic study with 100 samples, the total data volume can range widely from a few terabytes to tens of terabytes. A rough estimate for such a study could be:
    - Lower Bound: Around 1-2 TB (for a lower sequencing depth and simpler analyses).
    - Upper Bound: 50-100 TB (for a higher sequencing depth, single-cell data, and complex analyses).
- This estimation accounts for the accumulation of data across different stages of the metagenomic workflow, including raw and processed reads, assemblies, annotations, and associated metadata. Below is a more detailed rough estimate of the data size for the different stages of the process:
  - Raw Sequencing Reads:

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

- Single Sample: Depending on the sequencing platform (Illumina or Oxford Nanopore) and depth, a single metagenomic sample can generate between 10 GB and 100 GB of raw sequencing data. For example, Illumina platforms typically generate around 1-5 GB per gigabase (Gb) of Sample: Slightly smaller than raw reads, typically 80-90% of the original size, so 8-90 GB per sample.

- Multiple Samples: For 100 samples, this could be around 800 GB to 9 TB.

- Assembled Genomes or Contigs:

- Single Sample: Depending on the complexity, around 1-10 GB per sample.

- Multiple Samples: For 100 samples, this could be around 100 GB to 1 TB.

- Binning and Taxonomic Assignment:

- Single Sample: These files are usually small, around 10-100 MB per sample.

- Multiple Samples: For 100 samples, this could be around 1-10 GB.

- Gene and Functional Annotations:

- Single Sample: Typically, around 100-500 MB per sample.

- Multiple Samples: For 100 samples, this could be around 10-50 GB.

- Metadata:

- Single Sample: Generally small, around 1 MB per sample.

- Multiple Samples: For 100 samples, this would be around 100 MB.

- Single-Cell Data:

- Single Sample: Highly variable, but for a single-cell sequencing project, data per cell can range from 50 MB to 1 GB. For 1000 cells, this can range from 50 GB to 1 TB.

- Multiple Samples: For 100 samples each containing data for 1000 cells, this could be around 5 TB to 100 TB.

- Assembly and Binning Statistics:

- Single Sample: Small, around 1-10 MB per sample.

- Multiple Samples: For 100 samples, this could be around 100 MB to 1 GB.

- Functional Profile:

- Single Sample: Around 10-100 MB per sample.

- Multiple Samples: For 100 samples, this could be around 1-10 GB.

### 4.3.3 Biomolecule production

The volume of data generated from biomolecule production can vary based on the scale and complexity of the experiments:

- Environmental Monitoring Data: Small to moderate, depending on the frequency of measurements (KBs to MBs).

- Biomass and Growth Data: Moderate, especially with high-frequency sampling (KBs to MBs).

- Nutrient and Metabolite Concentrations: Moderate, depending on the number of analytes and sampling frequency (KBs to MBs)

- Genomic and Transcriptomic Data: Large, particularly with high-throughput sequencing (GBs to tens of TBs).

- Product Yield and Purity Data: Moderate, based on the number of assays (tens to hundreds of MBs).

- Image Data: Large, especially with high-resolution imaging and frequent captures (few GBs).

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

- Experimental Metadata: Small (a few MBs).
- Quality Control and Calibration Data: Small to moderate (KBs to MBs).

#### 4.5 Data storage

The amount of data produced by most of our platforms is not large: typically, a dataset size is in the range between a few Mbytes to Gbytes, except for the Cryo-EM structural determination and the metagenomic analyses, that can generate Tbytes of data. It is important to balance the collection of sufficient data for analysis and the need not overload the system with redundant information. The RI's server has a 1000 GB disk, that will be implemented with another 12000 GB disk with a data transfer rate of 255 MiB/s.

Metagenomic data will be stored in a dedicated cloud infrastructure located in Aruba's data center. Our core infrastructure is composed of two physically managed servers with a total of 40TB HDD storage, 8TB SSD storage, 40 CPUs and 192GB RAM. A managed physical firewall will oversee protecting our data and granting access to our users. Additional computational requirements can be met by relying on the cloud, either by adding other physical instances to our core or by adding virtual machines for the time strictly required to fulfill our needs. Raw data generated during analyses is currently backed up at our institute daily.

Scripts transfer the data periodically from the PCs handling the instruments to the ITINERIS Server which will be the node from which the ITINERIS Hub will receive them. We shall open the server to the Hub's requests following ITINERIS specifications. In particular, we shall expose our metadata catalogs through a CSW endpoint compliant with OGC specifications. We shall either set up a CSW server using GeoNetwork or a CKAN one. Our endpoint will support search (GetRecords) and retrieval (GetRecordById) operations to allow the Metadata Hub to perform the harvesting process. We shall communicate to the ITINERIS Metadata Hub the information needed to connect to our endpoint, including URL, access credentials (if required), and any specific search parameters.

#### 4.6. Sharing assets

Data and assets from the platform will be accessible to ITINERIS. If the data and assets belong to IBISBA, the IBISBA Data Management plan defines the standards for sharing data and assets generated within its scope. These data and assets will be linked to specific project beneficiaries and cataloged on the designated IBISBA data management platform. The primary repository for IBISBA assets is the project's internal collaborative platform. While access to the IBISBA data management platform is public, contributors are required to register for user access. Participants submitting data and assets to the platform will have the option to regulate access based on preferences. IBISBA asset owners can choose to: 1. Establish full access restrictions, creating a private asset, subject to compliance with project terms and conditions. 2. Limit access to project beneficiaries, determining whether they can view, download, edit, or manage the asset. 3. Make the asset public, with the option for viewing only or viewing and downloading. 4. Implement partial access, allowing specific project sections to remain private while others are shared with beneficiaries or made completely public. Adhering to the consortium agreement for IBISBA,

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

restricted access data (levels 1 and 2) will be subjected to an embargo period, lasting up to 48 months after the official project end date or until data publication, whichever is shorter. Asset owners may request a shortened embargo period, with such requests communicated to the IBISBA management team. Any alterations to the embargo period's duration are contingent on validation by the General Assembly, requiring official written requests before the end of the IBISBA project and decisions put to a vote.

## ***4.7. Archiving and preservation***

### ***4.7.1 Publications***

Aligned with our commitment to advancing the digitalization of bioprocess analysis and accelerating bioprocess development, we embrace best practices for disseminating research reports in compliance with the H2020 rules. Our preference lies with journals offering an open-access (Gold) option for publication. Alternatively, we opt for journals permitting self-archiving (green), with a specific inclination towards stable archives like bioRxiv.org. These platforms, serving as free online repositories, play a crucial role in the enduring preservation of scientific reports.

Every publication stemming from our project will undergo systematic uploading to OpenAIRE, the open-access repository dedicated to EU projects. Leveraging the robust capabilities of the SEEK platform, which implements the IBISBA data management framework, we streamline the publication process by generating Digital Object Identifiers (DOIs) for snapshots and packaging data in a format conducive to depositing in long-term storage platforms, including Zenodo (a component of OpenAIRE). This holistic approach not only ensures the accessibility of publications but also guarantees their longevity through archiving in reputable repositories.

In emphasizing our commitment to the adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, we actively contribute to the broader landscape of open science. Through these practices, our aim is to not only facilitate the digitalization of bioprocess analysis but also to expedite bioprocess development by fostering transparent, accessible, and reusable research outputs.

### ***4.7.2 Data archives and data papers***

With the overarching goal of fostering the digitalization of bioprocess analysis and accelerating bioprocess development, our platform places a strong emphasis on the publication of comprehensive datasets associated with research publications. Project beneficiaries are actively encouraged to disseminate these datasets, either as supplementary data accompanying publications or as standalone data papers. This can be seamlessly achieved through diverse platforms tailored for such purposes, including open-access journals that readily accept data papers.

Aligned with our commitment to advancing bioprocess development, the recommended platforms, integrated with the IBISBA data management framework, include OpenAire and Mendeley data (<https://data.mendeley.com/>). These platforms are strategically configured to

*D6.3 Standards for data acquisition and storage, developed and tested to facilitate digitalization of bioprocess analysis and accelerate bioprocess development.*

automatically harvest compatible datasets from the IBISBA data management platform, ensuring a streamlined and efficient process. Moreover, project beneficiaries possess the flexibility to leverage other open-source web applications, such as Dataverse ([dataverse.org](https://dataverse.org)), specifically designed for sharing, preserving, citing, exploring, and analyzing research data.

In pursuit of the deposition of public data archives conducive to bioprocess advancement, IBISBA utilizes relevant datatypes through established repositories. ELIXIR, the EU Research Infrastructure, maintains an exhaustive list of Core Data Resources and recommended Deposition Databases ([www.elixireurope.org/platforms/data](https://www.elixireurope.org/platforms/data)). Leveraging the robust underlying IBISBA data management framework, powered by the SEEK platform, datasets deposited in ELIXIR public archives can be seamlessly organized and cataloged alongside other project datasets, models, and Standard Operating Procedures (SOPs). This strategic approach ensures comprehensive and integrated access to research outputs, contributing significantly to the overarching goal of expediting bioprocess development through effective digitalization practices.

#### *4.7.3 Miscellaneous assets and official documents archiving*

In line with our commitment to facilitating the digitalization of bioprocess analysis and expediting bioprocess development, the platform addresses the archiving of project reports, encompassing deliverables and milestones. Our project management team will take proactive steps to ensure the preservation of these documents by registering them on the IBISBA data management platform (IBISBA Data Hub) and assigning a Digital Object Identifier (DOI) to each document.

In pursuit of our objective to accelerate bioprocess development, we recognize the significance of preserving key project documents beyond the project's lifespan. To achieve this, we are dedicated to exploring effective archiving solutions. Specifically, for documents classified as public at the project's conclusion, an optimal choice is the utilization of institutional open archives such as <https://hal.inrae.fr/>, which serves as the INRAE instance of the French National multidisciplinary open archive for research works belonging to public and private research operators.

For documents with restricted access, our approach remains meticulous. These documents will be securely stored on local servers until conditions permit their seamless deposition in an open archive, aligning with our commitment to transparent, accessible, and responsible project-related asset management. This strategic approach not only adheres to regulatory standards but also supports our overarching goal of advancing bioprocess development through effective digitalization and streamlined archiving practices.

## 5. INTERACTIONS WITH OTHER WPS OF ITINERIS

In the contest of the ITINERIS project, our activity is strongly correlated with other WPs and in particular with WP2 for improving FAIRness and WP3 for the training activities.

### 5.1 WP2. Access and Fairness

Within the ITINERIS project, the described activities follow WP2 guidelines for the FAIR (Findable, Accessible, Interoperable, and Reusable) principles and to ensure alignment with the ITINERIS strategy for the access procedures.

### 5.2 WP3. Training

Our activity is connected to WP3 for the common training system needed to ensure continuous formation to the current RIs personnel.