3D PATH

D6.1

Title: Data Management Plan

CLEPIO BIOTECH

DATE : 2025.06.30

Document control page

Project name	**3DPATH**
Document	**D6.1 – Data Management Plan**
Type (distribution level)	**☒ Public** **☐ Sensitive (confidential as specified in the Grant Agreement)**
Due delivery date	**2025.06.30**
Date of delivery	**2025.06.30**
Status and version	**v1**
Deliverable responsible	**Clepio Biotech**
Author(s)	**Claudia Capitini** **Ludovico Silvestri**
Official reviewer(s)	**Taiana Deodato**

Table of content

Executive summary **3**

1. FAIR data **5**

1.1 **Making data findable, including provisions for metadata** 5

1.2 **Making data accessible** 6

1.3 **Making data interoperable** 7

2. Other research outputs **8**

3. Allocation of resources **8**

4. Data security **8**

5. Ethics **8**

Overview 8

Ethics-committee workflow – key steps & methods 9

Executive summary

3DPATH aims to create a 3D scanner suitable for clinical use by leveraging technological innovations in Light-Sheet Fluorescence Microscopy (LSFM) and data management analysis.

The project will generate an extensive quantitative dataset from more than 300 human 3D histology samples—significantly exceeding previous pilot studies, which were limited by earlier LSFM methods. As no comparable large-scale dataset currently exists, a formal sample numerosity calculation is not possible. Instead, the data generated will be curated and analysed to establish evidence-based estimates for the sample size required in subsequent validation phases, including those necessary for CE-IVD Class C certification.

Each sample analyzed will generate approximately 1 TB of uncompressed data and 50 GB of compressed data. Samples will be decompressed for visualization and compressed for storage.

All samples will be assigned a unique alphanumeric study ID (UUID-style). These IDs will be included in all filenames and metadata records and will be mapped unambiguously to the clinical metadata sheet. If subsets of data are released at a later stage, their corresponding internal IDs will be included in the accompanying metadata, allowing users to trace each file back to the parent dataset associated with a DOI. Eventually, each study ID will be linked with the ID provided by the partners.

In addition to imaging data and metadata, the project will also generate human annotations of relevant structures within the imaged data. These annotations will be used to train and evaluate AI-based image analysis models and will be stored in standard .tiff format. Beyond the annotations, the AI models themselves will also be a key outcome of the project and will be stored in standard formats depending on the library used (e.g., .keras for Keras).

Ethical clearance will be obtained for all human samples. Local approvals will cover both biobank specimens and the fresh tissue analyses planned for year 3. A material transfer agreement between the two partner institutions will govern the exchange of samples. All permissions will explicitly allow the resulting 3D histopathology datasets to be deposited in public repositories.

The entire 3DPATH project will be conducted in accordance with open science practices, with the goal of sharing data with the community to maximise the project's social impact. Data management will adhere to the FAIR principles (see section 2).

Table 1 summarizes the data types used in each work package and their corresponding formats. This table has been shared with all partners, and any

new data types or formats will be added as needed. Each work package leader is responsible for keeping the table up to date.

WP	Data Type	Data Format	Generated	Collected
WP 1	Specification documents, development reports, testing reports, usage guide, manuals, data from (optical) simulation experiments, data for optomechanical components	docx, pptx, pdf, html, dwg, stp, Zemax files	Yes	Yes
WP2	Specification documents, development reports, testing reports, code, code documentation, neural network models, ground truth data, usage guide, manuals, docker files and images.	docx, pptx, Adobe Portable Document Format (PDF/A, PDF) (.pdf), html, py, ipynb, c, cpp, TIFF (.tif, .tiff), ZARR (.json), csv, sty, yml, .pt, HDF/5, dockerfile, txt	Yes	No
WP3	Textual documents, tabular data, presentations, images and graphics, scanned images	docx, pptx, xlsx, csv, JPEG (.jpeg, .jpg, .jp2), PNG (.png), SVG (.svg), GIF (.gif), TIFF (.tif, .tiff), Adobe Portable Document Format (PDF/A, PDF) (.pdf), mrxs, ndpi, svs	Yes	No
WP4	Textual documents, tabular data, presentations, images and graphics	docx, pptx, xlsx, JPEG (.jpeg, .jpg, .jp2), PNG (.png), Adobe Portable Document Format (PDF/A, PDF) (.pdf)	Yes	Yes
WP5	Textual data, presentations, tables, images and graphics	docx, pptx, xlsx, csv, JPEG (.jpeg, .jpg, .jp2), GIF (.gif), TIFF (.tif, .tiff), PNG (.png), Adobe Portable Document Format (PDF/A, PDF) (.pdf)	Yes	Yes
WP6-7	Textual documents, tabular data, presentations, images and graphics, web-based content, video material	docx, pptx, xlsx, csv, JPEG (.jpeg, .jpg, .jp2), PNG (.png), SVG (.svg), GIF (.gif), TIFF (.tif, .tiff), PNG (.png), Adobe Portable Document Format (PDF/A, PDF) (.pdf), MP4 (.mp4)	Yes	Yes

Table 1 – Types and formats of data generated and collected during the project.

1. FAIR data

1.1 Making data findable, including provisions for metadata

All datasets will be assigned a unique ID, unambiguously linking imaging data with metadata. We will generate a rich metadata structure, including clinical information (e.g. pathology, first diagnosis, treatment), information on sample treatment (e.g., staining), and on imaging (e.g., wavelength used, pixel size). To maximize discovery, we will use standard metadata for clinical information (European Health Data Space, https://ehds2pilot.eu/package/excellence-in-data-quality/), for medical imaging (DICOM, https://www.dicomstandard.org/), and for microscopy (OME, https://docs.openmicroscopy.org/ome-model/5.6.1/). The use of standards will encompass both metadata field names and keywords. Our metadata will be saved in .xml file format, enabling searchability with standard libraries. Metadata will be created during sample processing. In detail, clinical metadata will be provided directly by the suppliers of the samples. Sample treatment and imaging metadata will be provided by the operator processing the sample. To this aim, a Standard Operative Procedure (SOP) will be provided to all project partners, assuring coherent use of metadata descriptors.

Files will be identified with the elements described in Table 2. The date format used is YYYY.MM.DD.

The version format follows the structure:

- **v#.#** (e.g., *v0.1, v0.2, v0.3*): used for **versions under review**. Each increment reflects a new iteration or revision before the official release.
- **v#** (e.g., *v1, v2, v3*): used for **released versions**. The transition from a decimal version (e.g., *v0.6*) to an integer version (e.g., *v1*) indicates that the document or dataset has been reviewed, validated, and officially published.

Data type	File name convention
Lab Book (if digital)	ProjectName_LabBook_Number_AuthorName_Date
Report	ProjectName_Report_Number_AuthorName_Date_Version
Presentation	ProjectName_PresentationTitle_AuthorName_Date_Version
Interview	ProjectName_Interview_IntervieweeName_Date_Version
Samples Management Table	ProjectName_SamplesTable_HospitalName_Date_Version
Video	ProjectName _VideoTitle_AuthorName_Date_Version
Microscope images	ProjectName _ImageType_SampleID_Date_Version

Table 2 – File names convention for data generated during the project.

1.2 Making data accessible

All data and metadata will be stored in a server accessible to project partners through a secure, SSH-authentication-based protocol (SFTP). This kind of authentication is on a personal basis, so we will maintain strict control on who is accessing the data and when. SSH keys to access the server will be distributed upon completion of a form assessing the identity of the SSH key holder, and specifying his/her duties in following the provisions of this Data Management Plan. We will maintain records of the logs of our SFTP and SSH servers for the entire duration of the project. Data will be kept accessible only to consortium members until publication of scientific papers, on non-peer-reviewed repositories (e.g., bioRxiv), or until all consortium members agree for their dissemination. The event occurring first will enable data sharing with the public. This embargo is to maximise the scientific impact of the research, and to guarantee that scientific merit of this research is acknowledged to the consortium generating the data. In any case, all data will be made publicly available before the end of the project.

For sharing with the public and the community, data and metadata will be made accessible by uploading them to an institutional repository suitable for big data, Zenodo (https://zenodo.org/). In addition, we are exploring agreements with more specialized institutional repositories, such as BigPicture (https://bigpicture.eu/) and EOSC4Cancer (https://eosc4cancer.eu/). In addition to data and metadata, we will also share the software tools needed to explore the data, together with instructions for use.

All the data and metadata will be retained by the coordinator for at least 3 years after the end of the project. In any case, Zenodo will guarantee long-term preservation of data.

All data and metadata will be accessible through CC BY-NC-SA license.

As clinical data will be fully anonymized, we do not foresee any specific need for a data access committee.

1.3 Making data interoperable

As specified in 1.2, we will make use of well accepted metadata standards. Since we are putting together different standards in a single metadata file, we will provide an open document containing a mapping of each metadata entry in our files to the original ontology. This mapping file will be published on GitHub, and a link to it will be provided in each metadata file generated.
Data will be stored using the NGFF Zarr format (https://ngff-zarr.readthedocs.io/). A link to the documentation of the standard NGFF Zarr library will be provided in each metadata file generated.

As specified in 1.2, all data will be made publicly available before the end of the project, under a CC BY-NC-SA license.

To further improve data reuse and interoperability, we will provide a white paper describing in detail the data and metadata structure, the main software tools for accessing both data and metadata, the sources of the data, their intended use and limitations imposed by the CC BY-NC-SA license. This white paper will be published on GitHub, together with the mapping file mentioned above, and a link to it will be provided in each metadata file.

Open & FAIR – All imaging data will follow open-science and FAIR principles, according to EOSC recommendations.

Scale – ~500 samples → ~750 TB raw (1 cm^3 @ 1 μm^3, 3 × 8-bit channels); lossy compression trims this to ≤ 40 TB.

Format & curation – Stored as NGFF Zarr with rich metadata on local servers; consortium members will review quality of both data and metadata prior to final publication.

Release workflow – Once approved, datasets (including AI-ready annotations) are published on OSF/EUCAIM with DOIs for maximum findability.

Licensing – Shared under CC-BY-NC-SA: free non-commercial use, no third-party commercialisation, same license for derivative work.

Ownership & costs – Each partner owns and funds storage for the data it generates (Clepio owns most; demo-site data belong to the respective institutions).

Commercial path – Clepio retains the right to commercialise its own datasets and will negotiate agreements to licence other partners' data, ensuring both

community access and a competitive edge for AI tool development (AI training).

2. Other research outputs

In addition to 3D imaging data and metadata, the project will also produce AI models for image segmentation and analysis. These models will be owned by the party developing them. Potential sharing of some models will be decided case by case. In case of sharing, the license could be CC-BY-NC-ND 4.0: free non-commercial use, no third-party commercialisation, no creation of derivative works.

Protocols for sample preparation might be shared on public repositories (e.g., https://www.protocols.io/). Before sharing, potential protection of the IP - e.g., with a patent - will be evaluated by the partner(s) developing the protocol.

3. Allocation of resources

The costs associated with the Data Management Plan are primarily related to data storage. These are estimated to range between €15,000 and €20,000.

The responsible for the Data Management Plan is Ludovico Silvestri, CSO at Clepio Biotech.

4. Data security

Local storage will be performed in high-end Network-Attached Storage (NAS) systems with disk redundancy (RAID 5 or 6). Long term preservation will be guaranteed by usage of the trusted repositories mentioned in 1.2.

5. Ethics

Overview

The study reuses > 300 biobanked blocks for 3D histopathology, so it must satisfy Swiss, Italian and EU regulations on human tissue research, data protection, and Good Clinical Practice. Ethical approval, informed consent, secure data handling, and a material-transfer framework ensure participant rights while enabling cross-site analysis.

Ethics-committee workflow – key steps & methods

1 **Regulatory framework** – This Data Management Plan ensures compliance with international, European, and national regulations and guidelines governing the ethical conduct of research and the protection of personal data. Specifically, the project adheres to:

- The Declaration of Helsinki, which sets ethical principles for medical research involving human subjects.
- The ICH-GCP (International Council for Harmonisation - Good Clinical Practice) guidelines, ensuring that the rights, safety, and well-being of trial participants are protected and that clinical trial data are credible.
- The General Data Protection Regulation (EU Regulation 2016/679, GDPR), which governs the processing of personal data within the EU.
- Italian data protection laws, including Legislative Decree 196/2003 (Personal Data Protection Code) as amended by Legislative Decree 101/2018, which harmonizes national law with the GDPR.
- Swiss data protection legislation, including the Swiss Federal Act on Data Protection (FADP) and related ordinances, particularly relevant to the management and use of samples and data sourced from Swiss biobanks.
- European Medicines Agency (EMA) guidelines on clinical trial data handling, protection of personal data, and transparency, particularly applicable to data sharing and publication requirements.

2 **Delegated responsibilities** – The Principal Investigator (PI) is designated as the data processing delegate. Authorized staff members involved in data processing activities are formally designated in writing. The PI is responsible for submitting annual status reports on the study to the Ethics Committee.

3 **General consent** – A baseline biobank consent is already in place for previously collected samples. When feasible, participants will be re-contacted and re-consented for new uses of their samples or data. Written consent will be obtained and documented prior to any new data use.

4 **Confidentiality & data protection** – Data will be pseudonymized before transfer and analysis. Access to personally identifiable information will be restricted to authorized personnel as designated by the PI, and

appropriate safeguards will be in place to ensure confidentiality. Cross-border data transfers will comply with EU and Swiss legal requirements.

5 **Material handling** – FFPE blocks will be retrieved under an Ethics Committee-approved biobank protocol. A Material Transfer Agreement (MTA) will govern the shipment of samples to the Clepio and their return to UBERN following LSFM analysis, when applicable.

6 **Participant rights** – Participants may withdraw consent at any time. Upon request, samples will be destroyed and personal data erased, unless anonymized research has already been completed using the materials.

7 **Record retention** – De-identified datasets originally from Italy will be stored for at least 7 years. Samples originally from Switzerland for a period of 10 years.