DELIVERABLE 7.1 - DATA MANAGEMENT PLAN (DMP)

VERSION 2.0

BioNT

NOVEMBER 21st 2025

biont-training.eu

contact@biont-training.eu

Co-funded by
the European Union

About this report

Grant Agreement	101100604 - DIGITAL-2022-TRAINING-02
Action Acronym	BioNT
Action Title	Bio Network for Training
Deliverable	7.1 Data Management Plan
Work package	7 Project management and coordination
Dissemination level	Public
Authors	R. Alves, S. Di Giorgio, T. Müller, E. Ortega, L. Paladin, I. Paredes Cisneros, S. Razick, approved by all BioNT partners
Delivery date	2025-11-21

Consortium members

Acronym	Partner
EMBL	EUROPEAN MOLECULAR BIOLOGY LABORATORY
BIOBYTE	BIOBYTE SOLUTIONS GMBH
HPCNOW	HPC NOW CONSULTING SL
UO	UNIVERSITETET I OSLO
UB	UNIVERSITAT DE BARCELONA
ZBMED	INFORMATION CENTRE FOR LIFE SCIENCE
RIcapacity	RIcapacity
ALU-FR	ALBERT-LUDWIGS-UNIVERSITAET FREIBURG
EPFL	ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE

Table of Contents

History of changes

Version	Publication Date	Changes
2.0	2025-11-21	• The consortium partners were updated as per the last amendment: Simula Consulting was replaced by RIcapacity • The reference to mailing lists for participants was removed, as the consortium decided to design a pipeline

		granting them anonymity. Mailing lists were clarified. • The process for pseudo-anonymisation of participants' data was described. • Publication of the project's deliverables was mentioned. • Formats for translations were specified. • The Lhumos platform was introduced, and the project's planned usage of it was described. • Updated data retention practices by different platforms, including Zoom, Google Workspace and CECAM's course management. • Clarified the purpose and use of mailing lists for the entire project and the community event. • Aligned text with current communication practices on chat and social media platforms, as well as email. • Introduced the Zenodo BioNT community to facilitate documents uploaded to this platform. • Expanded list of formats to reflect those used to date. • Handling of personal data is now more explicit on what data is kept, for how long and on which platforms. • Removed use of GitLab for hosting training materials. GitHub is now the primary platform with all content under the BioNT organisation. • An elaborate translation process, detailing the AI-assisted and curation steps. • The training material section was significantly reworked to reflect the use of Galaxy Project and The Carpentries tooling, document structure and, where applicable, infrastructure. • Descriptions about metadata were embedded in the corresponding data-type sections. • The use of GitLab for project management and Git and Google Workspace for document management and versioning is now aligned with project practices. • Long-term data preservation is now streamlined to follow licensing and release of outputs as per EMBL's Open Science policy. • Data storage and resource use now reflect an estimate of the storage needed for all project outputs, including lesson materials and media such as recordings. • Clarified practices and execution of the DMP by project members. • Corrected the work package the DMP belongs to
1.0	2023-04-28	• First version

Project Overview

The BioNT consortium is dedicated to providing a comprehensive training program and fostering a community for digital skills relevant to the biotechnology industry and biomedical sector. With a curriculum tailored for both beginners and advanced professionals, BioNT aims to equip individuals with the necessary expertise in handling, processing, and visualising biological data, as well as utilising computational biology tools. Leveraging the consortium's strong background in digital literacy training and extensive network of collaborations, BioNT is poised to professionalise life sciences data management, processing, and analysis skills.

Data Management Plan

This document describes the Data Management Plan (DMP) for the BioNT project. It provides an overview of the data management policies, how they will be implemented throughout the project and how these apply to all types of data generated in the Action. This DMP is a living document: it will be updated with details or relevant changes throughout the project duration or whenever necessary. The format of the plan and its content have been defined taking into account the guidelines on Data Management provided by the European Commission and the FAIR principles.

The objects collected and managed during the implementation of the action will fall under two primary categories:

- Data
- Software

The current document includes two separate sections for the two categories. Additionally, it includes sections on storage and backup, as well as access and security, applicable to both Data and Software. Finally, legal aspects and the resources needed for the implementation of this DMP are described in the last two sections.

Data

This section describes data which will be used in the projects, including how it will be managed and stored. It explains how the data lifecycle will be made compliant with the FAIR principles:

- *Findable: Will a persistent identifier be provided? Will metadata be rich enough (e.g. relevant keywords) and made available in such a way as to allow discoverability (i.e. can the metadata be harvested and indexed)?*
- *Accessible: If access restrictions will be implemented, how will they be minimised to ensure as much access as possible to the data? Will specific software be required to access/read the data/metadata and will this software be made available?*

- *Interoperable: Describe data and metadata vocabularies, ontologies, standards, formats or methodologies you will follow to make your data interoperable in order to allow data exchange and re-use, and in particular to facilitate re-combinations with different datasets from different origins.*
- *Re-usable: Describe how you will provide documentation and software needed to reproduce/validate data analysis and facilitate data re-use.*

Data types and formats

This section includes details on the types (e.g. images, video, numeric, text) and formats (e.g. TIFF, tab-delimited text, hdf5) of data to be produced, including data produced through processing and analysis of other data.

BioNT will generate a range of data types and formats, which will be carefully managed and preserved throughout the project lifecycle. Specifically, the project will produce the following data:

1. Course participant information: Data such as participant names, contact details, and demographic information, used to manage course registration, certificate generation, and to be anonymised and aggregated to measure training impact in BioNT reports. Temporary data, such as IP addresses and account identifiers, may be stored to allow virtual participation (e.g. Zoom), but are automatically deleted after 90 days.

2. Mailing lists: Generated to facilitate internal communication and coordination among project contributors. Mailing lists created for this aim include staff of partner organisations in the consortium. Additional mailing lists were created to support the BioNT Community event and deleted once no longer needed.

3. Training materials: Documents, presentations, and other materials used to deliver training and educational content to participants.

4. Images and illustrations: Visuals used during the workshops and as part of the training material, as well as in the course registration platform, website and other dissemination and communication activities.

5. Input data used during the workshops: Datasets used in the workshop practicals, to exercise and test data analysis and other acquired skills.

6. Video recordings of lessons: Video recording footage of training sessions and lessons for documentation purposes, and to provide participants with access to the content after the event.

7. Video recordings of presentations and sessions, as well as other media created during the CarpentryConnect and BioNT Community event 2024. To be edited following the data privacy agreement for the event and shared only with participants of the same session(s). A group photo was used on the BioNT website with consent from its participants.

8. Internal meeting notes: Recorded during consortium meetings.
9. Chat logs: Regular communication between partners across three platforms (Matrix, Slack, Mattermost). Generated during online training sessions, for interactions between trainers and helpers.
10. Survey data: Data from surveys designed to assess the effectiveness of the training program and to gather feedback from participants. This data is pseudo-anonymised as described below.
11. Website content: Descriptions of the training program, participant testimonials, and other relevant information about the consortium, its activities and outputs.
12. Presentations delivered in conferences: Presentations and posters to showcase the training program at conferences and other events.
13. Public deliverables of the project, displayed on the project website and some are also archived on Zenodo, in the community: zenodo.org/communities/biont_eu-project.

To ensure that all of this data is properly managed and preserved, the project team will follow best practices for data management and will use appropriate data formats and standards. Project data will be stored in different, backed-up locations, depending on their use and applicable policies and regulations. These locations are described in the *data storage platforms* section below. Additionally, the project team will work to ensure that all data is properly documented, annotated, and labelled to facilitate its reuse and interpretation by others.

BioNT will use open and non-proprietary formats to promote data accessibility and interoperability. Followingly, a tabular description of the formats used in the project:

Data type	**Format(s) used**
Course participant information	Spreadsheet (CSV), structured text (JSON, XML)
Mailing lists	Spreadsheet (CSV), mailing lists.
Emails	
Information about entities for advertisement	Spreadsheet (CSV), structured text (JSON, XML, HTML)
Training materials	PDF, HTML, Markdown, Slides (PPT/PPTX, ODP HTML)

Images and illustrations	SVG, PNG, JPEG, PDF
Input data used during workshops	Spreadsheet (CSV), structured text (JSON, XML), other open formats specific to tools and software used (for biological sequences handling: FASTA, FASTQ, GFF/GTF, BED, BAM/SAM)
Video recordings of lessons	MP4, Ogg, WebM
Internal meeting notes	Text files (Google Docs), Markdown
Zoom chat logs	Text file (TXT), JSON
Zoom recordings and captions	MP4 (Lhumos), SRT
Survey data	Spreadsheet (CSV), online survey tools (LimeSurvey, Google Forms)
Website content	HTML, CSS, JavaScript, Markdown, PDF
Presentations delivered at conferences and meetings	PDF, HTML, Markdown, Google Slides, Slides (PPT/PPTX/ODP)
Translation of training materials	PO, Markdown, HTML

Data storage platforms

To accommodate the diverse requirements across the project's life cycle, and to share content with different audiences, BioNT will use multiple platforms to host and share data. These platforms are described below:

- The pipeline for managing registrations to and feedback from workshops is based on two platforms. The CECAM platform, hosted at EPFL, is used to manage the applicant's information and communication. In parallel, the LimeSurvey platform hosted on EMBL servers is used to collect pre- and post-workshop information through pseudo-anonymised surveys. The survey data is linked to the applicant's data only via a unique identifier, provided in the CECAM registration process, as well

as in the EMBL-based survey. This ensured that only workshop organisers could link the two information sources and confirm that applicants completed their feedback. Trainers, helpers, and anyone else who might be involved in the participants' selection only access their anonymised surveys and the unique identifier, but cannot match it with personal data. All participant data (as well as most of the data related to each workshop instance, e.g. the dates) is therefore hosted on the CECAM platform at the care of EPFL. More information about this is included in the section Legal aspects.

- To access the virtual workshop, participants join a Zoom webinar call. Zoom was chosen as the delivery platform due to its webinar mode that allows participants to remain unidentified while also allowing recordings and exchanges between trainers and helpers. Temporary data, such as IP address and user-defined alias, is stored in the platform, which allows the organiser to track their participation. This data is deleted automatically after 90 days. Towards trainers, helpers and other workshop participants, they can hide their identity by editing the displayed name.

- Working documents that do not contain personal data of course participants and require synchronous online editing, such as meeting notes, will be stored and edited in a shared drive on EMBL's Google Workspace. All partners will have access to this drive, and external collaborators have access to specific subfolders linked to the collaborative activity. All files in a Google Workspace shared drive are owned by the organisation (EMBL), allowing the project management team to control and monitor the data, as well as to delete it if needed. Data is kept for 30 days after deletion. More information about the usage of this platform can be found in the section Document storage practices.

- For backup, versioning, and project monitoring purposes, documents storing important, long-lasting information (e.g. process documentation) will be generated, live edited and stored in a Markdown-based format and transferred to a repository in GitLab or GitHub. These projects and repositories (only accessible to the project partners) also serve to monitor the project progression through issues, tasks, boards and milestones. Training materials in Markdown and HTML formats will be hosted publicly on GitHub, where they are expected to remain beyond the lifetime of the BioNT project. Where there was a need to store the data in a repository physically hosted in Europe, the entire repository was stored on ZB MED's GitLab instance. More information about the usage of this platform can be found in the section Document storage practices.

- GitLab.com or GitHub.com will also be used to collaboratively design and develop training materials, with one repository for each course and language, where applicable. The structure of these folders is detailed in the Lesson folder structure section. Training and accessory lesson materials (when available in the format of presentations or slides) will also be deposited, after the course delivery, in one repository that will provide a digital object identifier (DOI) to them (e.g. Zenodo), to enable systematic/combined search based on the deposited metadata.

- Video and audio materials generated during the training program will be uploaded to the Lhumos platform, a training platform that features video recordings of lectures, slide navigation, access to support material, code repositories, and exercises. Justification of this choice is available in the section Storage and back-up.
- The BioNT website includes relevant information and data about the project. An extensive description can be found in the section Website. In addition, more information about the website content will be included in the Dissemination and Communication Plan.
- In line with the call for proposals, all project events' information will be sent to the Digital Skills and Jobs Platform. To provide potential participants with relevant information about the courses, each entry in the platform will follow the recommendations and guidelines provided by the Platform's team. More information about the usage of the Digital Skills and Jobs Platform will be included in the Dissemination and Communication Plan.
- Finally, while not directly related to data storage, it is important to note that information about the project will be disseminated through various BioNT communication channels, such as social media. Further details on the usage and practices of these channels will be provided in the BioNT Dissemination and Communication Plan.

The table below refers to the Data Type presented in the section Data types and format, and indicates which platform(s) will be used for each.

Data type	**Platform(s) used**
Course participant information	members.cecam.org
Mailing lists	Mailing list software
Information about entities for advertisement	Google Workspace (continuous update), Social networks (initially Twitter/X, later LinkedIn)
Training materials	GitLab / GitHub, Zenodo, DeepL Translate
Images and illustrations	members.cecam.org, Google Workspace, GitLab / GitHub, Website, Canva, Zenodo
Input data used during workshops	GitLab / GitHub

Video recordings of lessons / Captions	Lhumos, Infomaniak kDrive (with EPFL)
Internal meeting notes	Google Workspace (while editing), GitLab (for storage)
Survey data	GitLab, Google Workspace
Website content	GitLab and biont-training.eu website server (at ZB MED)
Presentations delivered at conferences and meetings	Google Workspace (while editing), GitLab (for storage), Canva, Zenodo

Reuse of existing data

This section lists existing data sets that will be used and specifies the terms of use (e.g. licence, collaboration with the data producing group). When reusing public data, it provides links to the source.

Training materials

The training material for this project has been initiated in The Carpentries or other established training communities and is at different stages of development. The Carpentries is an open and inclusive community that creates and shares teaching materials for data science and related skills. The training materials are designed to be compatible with self-learning and to be adopted by the community of trainers, as they include extensive textual descriptions. This approach is crucial for the sustainability of training projects, and all other sources of training materials considered in BioNT, such as the Galaxy Project (specifically Galaxy Training Project) and CodeRefinery.

New training material has been generated using the new lesson infrastructure developed in The Carpentries, called The Carpentries Workbench. The Workbench provides a modern, modular, and flexible platform for developing and producing Carpentries-style lessons. Should the training materials not be directly generated in the Workbench, they will be designed in a format compatible with a later adoption by the Workbench, hence in Markdown-based text, assisted by presentation slide decks where necessary.

Existing lessons from other communities will also be integrated, including the Galaxy Training Network and CodeRefinery materials. These lessons will be personalised based on the partners' expertise and stakeholder requests. The Galaxy Training Network is a collaboration between the Galaxy Project and the ELIXIR training community that develops

and shares training materials for bioinformatics tools and workflows using the Galaxy platform. CodeRefinery is a project that provides sustainable and collaborative software development training for researchers in academia and industry.

All the training materials in these projects are openly available under permissive licences, such as Creative Commons, and can be accessed via their respective websites. Links to these materials will be provided on the BioNT's website and in any other relevant documentation. The version information and dates of the materials that are used will be tracked to ensure reproducibility and transparency of the work, as the materials are continuously evolving.

Additionally, all BioNT training materials will be versioned, including the new and personalised ones, using GitLab / GitHub to keep track of changes and facilitate collaboration within the project team and with the wider community. The materials will be released under a Creative Commons Attribution (CC-BY) licence, which allows others to use, remix, and build upon BioNT's work, provided they give appropriate credit to the original authors.

Following, a table including the link to the existing training materials that will be adapted for the project's purpose, for each BioNT course. The reference to the specific versions will be added at the start of the specific course design phase.

Training workshop	**Source of existing training material**
Bioinformatics introduction	• training.galaxyproject.org
Introduction to programming languages	• carpentries.org/workshops-curricula • training.galaxyproject.org
Command-line and cluster computing	• hpc-carpentry.org/ carpentries-incubator.github.io/hpc-intro/ • carpentries-incubator.github.io/workflows-snakemake/ • pawseysc.github.io/singularity-containers/
Open and FAIR principles and Data management	• carpentries-incubator.github.io/fair-bio-practice
Instructor training	• elixir-europe.org/platforms/training/train-the-trainer • carpentries.github.io/instructor-training/
Software development best practices	• osulp.github.io/git-advanced/ • carpentries-incubator.github.io/python-testing/ • coderefinery.org/lessons/
Machine learning and Artificial intelligence	• carpentries-incubator.github.io/deep-learning-intro

Data structure

This section describes how the data will be organised and managed during the project. For filesystem-based data management, a description of the directory structure and naming conventions is provided, together with quality control procedures for data content, structures and conventions.

Document storage practices

To ensure that data is organised and managed effectively during the project, a set of file naming conventions has been established. The conventions will apply to all platforms where data and documents will be stored (e.g. GitLab, Google Workspace and BioNT website). Each file will have a short filename describing its content in a meaningful way. To avoid spaces in filenames, underscores will be used to separate words. Where needed and to ensure that each final document is uniquely identifiable, filenames will include a timestamp and/or a version number. Dates used will follow the ISO 8601 standard (YYYY-MM-DD, such as 2023-04-23).

Using GitLab / GitHub and Git version control will allow tracking all changes to the data and ensure that each version is securely backed up. When needed, READMEs, links to relevant documents and a table of contents will be provided as project documentation. Details about the data organisation, directory structure and naming conventions will be included where relevant.

A Google Workspace drive (BioNT-DIGITAL-2022-TRAINING-02) will be created and managed by the coordinating organisation.

Lesson folder structure

As lessons will be developed using tools by the Galaxy Project and The Carpentries, BioNT content will follow any conventions in use by those projects. All the work will be tracked with Git and stored centrally on GitHub / GitLab repositories. Lesson content will additionally be rendered and offered in the form of a website for easy consultation and use. This content will be linked from the BioNT website and other applicable contexts (e.g. next to lesson recordings).

Website

BioNT has a dedicated website to provide information about the training materials and related events. The website has been developed using modern web technologies to ensure compatibility with a wide range of devices, screen resolutions and browsers. To improve the findability of the training materials on the web, Bioschemas will be implemented. Bioschemas are structured metadata annotations for web pages that provide machine-readable information about the content of the website. This markup follows the Schema.org vocabulary, which is widely used by search engines, digital assistants, and other web services. By using Bioschemas, the visibility and discoverability of the training

materials will be enhanced, facilitating their integration with other web-based resources and platforms.

For the entire duration of the project, the website will be hosted and maintained in a private virtual machine at ZB MED, under the registered domain "biont-training.eu". More details about the website content will be available in the Dissemination and Communication Plan for BioNT.

The website includes a download section, where users can access the training materials and related resources, as well as a news section, where the latest developments and events are announced. Links to the project's social media accounts are present in the footer and other locations. In addition, the BioNT website provides information about project partners, their expertise, and their roles in the project.

Metadata and documentation

This section outlines the methods for documenting and tracking data, as well as how to access and link relevant documentation to the corresponding data. The documentation process will adhere to community metadata standards and will encompass all necessary information for effective discovery, interpretation, and re-use of the data and training content.

Summary of relevant points previously mentioned

To track and document the data, GitLab / GitHub repositories are used for version control and documentation. Each lesson generated has its own repository, with a corresponding webpage for ease of access. Within each repository, a README file is included, containing all relevant information about the lesson, such as the methodology used to collect the data, data processing and analysis steps. Additionally, a contribution guide and automatic sanity and style checks are included to increase accessibility and ensure the quality of contributions.

Documentation

Relevant documentation to enable the interpretation and reuse of the training materials is present both in the material itself, as instructor guidelines, and in this document. Documents are versioned, and an implicit history of changes is kept for each file (as for this one). Lesson materials are versioned with Git as described in the Software section below.

Software

This section provides an overview of the software that will be utilised for generating, processing, and analysing data in the project. It also outlines the management of the code used in the project, including how it will be made available and the level of long-term support that is being considered. Additionally, it addresses the FAIR principles for data publication.

In the BioNT project, all code used for training purposes during the lessons has been managed using Git, a version control system. All partners, during the development of the training materials, store their code in a code repository and, when applicable, centrally hosted on platforms such as GitLab and GitHub. The project management team ensures that the code repositories comply with the requirements of FAIR software. The compatibility with long-term support will be guaranteed by following the best practices in the software development community, including regular updates, bug fixes, and documentation.

Storage and backup

This section describes where the data and metadata will be stored, including the backup strategy.

Training materials are stored on GitHub, EMBL's Google Workspace (training presentations) and Zenodo, once final. Git repositories stored in GitHub live both in the central location and as "clones" or copies on the personal computers of collaborators. While not strictly a backup, this distributed way of working adds a protection layer against data loss or damage. Presentations in the Google Workspace are versioned and can be restored within a 30-day window if deleted.

The EMBL Google Workspace platform is also used as a cloud storage and collaborative editing tool for meeting agendas, documents and other files that may require collaborative access and tracked live editing. Restoring or comparing old versions is possible at any time.

Metadata is stored on multiple platforms to ensure accessibility in the different phases of usage. Zenodo provides a final archival for some of the material (presentations), whereas GitHub and the BioNT website include other machine-actionable metadata descriptions, such as JSON-LD BioSchemas descriptions. Lhumos is described below.

The BioNT website content is stored on ZB MED's GitLab server and rendered to a locally hosted virtual machine that serves the biont-training.eu website. As all content lives in the Git repository and website content is not directly editable, no backup of the VM is required.

Mailing lists are stored and managed through an internal list server, modifiable only from the internal EMBL network. This ensures data security and limits edit access to members of the organisation while still allowing for emails to circulate through the mailing list itself.

Video and audio materials generated during the training program are stored on the Lhumos platform (lhumos.org), an open-access platform for scientific media. The Lhumos platform automatically extracts video scrubbing information and allows organising different videos in lessons and curricula under the BioNT space. Videos will be metadata-annotated, which will ensure that the videos are discoverable through widely used search engines and accessible to the scientific community.

Access and security

This section describes who will access the data and how (in particular, if access is needed for external collaborators) and what security measures are in place. The project will deal with sensitive information (personal data of participants), hence it references documents or agreements related to data access and sharing, and describes the risks and mitigation steps.

Access to the administrative/management data will be granted to all project partners and relevant external collaborators, depending on their involvement and need for the data. The data will be accessed through the GitLab platform and the Google Workspace cloud storage, with appropriate permissions and access controls in place to ensure data security and integrity.

For external collaborators, a Data Sharing Agreement (DSA) will be signed to ensure compliance with relevant data protection regulations and to outline the terms and conditions for data access and use. In cases where sensitive data is being handled, the DSA will include specific clauses to protect the privacy and confidentiality of the data, and to describe the risks and mitigation steps associated with data access and sharing.

All partners and external collaborators will be required to follow appropriate data security measures, including the use of secure passwords and/or two-factor authentication. All partners are aware of this document and any data privacy and confidentiality restrictions that the project and its member organisations abide by. In addition, regular backups of the data are performed as described above to ensure data preservation and recovery in case of data loss or corruption. Any security incidents or breaches will be reported to the project officer and relevant authorities, as appropriate.

The training materials, in line with the Open Science model that the project was designed on, will be released under the Creative Commons Attribution (CC-BY) licence to facilitate sharing, re-use, and adaptation of the content.

Legal aspects

This section highlights relevant legal aspects regarding the usage of sensitive data, and describes what data will be publicly released, when and under which licence.

Handling of personal/sensitive data

Potential participants will need to provide personal data (demographic data) when applying to courses; however, no sensitive data will be required for participant selection. Therefore, the management of personal, but not sensitive data, is described in this paragraph.

Personal and demographic data will be collected and stored through the platform members.cecam.org, provided by the project partner EPFL-CECAM. All data is hosted on

EPFL-CECAM internal servers. The platform is designed to manage the course registration process, as well as the selection of and communication to participants for the courses, and therefore already includes most functionalities required by BioNT. Details about their terms and conditions can be found here: members.cecam.org/terms-and-conditions.

In summary, to protect participants' data, the project will:

- Obtain user consent to process personal data,
- Inform users about the existing mechanism to respond to data subject requests, i.e. requests for access, correction, deletion, etc.
- When needed, establish a retention period for personal data to be stored; all personal data will be kept on the CECAM platform for an undetermined amount of time, but deletion can be requested at any time.
- Anonymise the course feedback data and the demographic information needed to measure the project's success in reports to protect the participants' privacy.

In addition, all project members who have access to the personal data agreed to abide by the present document and all its terms, ultimately serving as internal regulation.

Data ownership and intellectual property

This section details who the data owner is, i.e. who has the right to control access. It covers ownership and IP matters, and explains whether intellectual property rights are affected, which ones and how they will be dealt with.

The ownership of the training materials will be shared among project partners. The Carpentries incubator will not claim intellectual property rights unless the modified materials are pushed back to The Carpentries Incubator (for later adoption by The Carpentries training community). Therefore, the project team will use The Carpentries Workbench to develop the training materials and host them on an internal server. The use of the Workbench does not imply The Carpentries' intellectual property rights.

The project team will ensure that all materials are appropriately licensed. Training materials will be published under the Creative Commons Attribution (CC-BY) licence, which allows for sharing and adaptation as long as proper attribution is given to the original creators. Any intellectual property rights related to the data generated through the project will be shared among the project partners, with no one partner having exclusive ownership.

Data publication and long-term preservation

This section describes how the project will comply with the partners' policies and how data for preservation will be selected. Additionally, it explains if and what data must be retained or destroyed for contractual, legal or regulatory purposes.

The DMP for BioNT involves several aspects related to the storage, access, preservation, and sharing of project data.

All BioNT partners will work to ensure that all data and training material generated are “FAIR” by ensuring that the data is Findable, Accessible, Interoperable, and Reusable. This includes providing website metadata to allow discoverability, following data and metadata vocabularies, ontologies, standards, formats, or methodologies in the field to make training datasets interoperable, providing documentation and software needed to reproduce and validate the code used in the training materials, and finally, to minimise accessibility barriers to the training materials and video recordings.

The project data will be preserved for at least 10 years, complying with the requirement of the Open Science policy of the institution coordinating the project, EMBL. The data that must be retained or destroyed for contractual, legal, or regulatory purposes will be identified and handled appropriately.

The project will publicly release all the data and training materials generated. While there are no plans for any research paper as an output of the project, if any of the partners choose to work on one, there is no reason to withhold project data until the time of publication. The project will use the Creative Commons Attribution (CC-BY) licence for all publicly released data.

In conclusion, the BioNT project has put in place a robust data management plan that takes into consideration the storage, access, preservation, and sharing of project data. The project team will ensure that all data and training materials generated are FAIR and will comply with EMBL's requirements for data preservation. The data and training material generated by the project will be publicly released, and the data owner is the BioNT project. The project will follow appropriate measures to ensure the intellectual property rights of all partners involved.

Responsibilities

This section identifies (naming individuals if possible) who are the project stakeholders and their responsibilities.

The project stakeholders and their responsibilities, as well as the coordination of data management responsibilities across partners, are specified in the proposal. Each task, deliverable, milestone and work package is assigned to a specific partner and, in particular, to the team working in the partner entity and assigned to the BioNT project, along with their responsibilities and required resources.

Updates and changes to the DMP will be the responsibility of the team in the coordinating institution, EMBL, who will communicate with the project stakeholders as needed. The project manager, reachable via contact@biont-training.eu, is responsible for coordinating data management responsibilities across partners and informing the project stakeholders of any relevant changes or updates. The project manager is also the reference for this document, and can be contacted for any queries related to it.

Resources

This section identifies the resources needed to implement this DMP and which budget covers the associated costs.

The following resources will be needed to implement this data management plan:

- Storage: Currently, the storage needed for the training datasets, software code, and training materials for all BioNT training repositories maintained in GitHub / GitLab is estimated to be in the 2-4 GB range. This estimation includes intermediate files needed to generate the final version of the training material and its translations. Additional storage will be needed for long-term preservation of project data (e.g. survey results) through GitLab, as well as for input test datasets to be used in the workshops. Video recordings and other media are expected to require around 200 GB of storage. This estimate is, however, likely to differ from the final number as video compression algorithms improve. The storage capacity estimation will be continually assessed as the project progresses, and updated information will be added to this section.
- Technical requirements: High-bandwidth Internet access is needed to facilitate data transfer and collaboration among project partners. In addition, access to high-performance computing resources for the relevant courses is required. To ensure that project data is only accessible to authorised personnel, access control mechanisms are in place. Considering the project's commitment to open source tools and formats, no purchase of specialised software has been anticipated. More detailed technical requirements for each course are included in the corresponding course reports.
- Training needs: All project partners and the coordinating team have adequate expertise to train the project staff on matters concerning this DMP, and will approach the project management team in case of doubts.