BioNT - Network for Training

Deliverable 7.1 | Data Management Plan

Co-funded by the European Union

Version 1.0

28 April 2023

Grant Agreement number	101100604 - DIGITAL-2022-TRAINING-02
Action Acronym	BioNT
Action Title	Bio Network for Training
Deliverable number and title	7.1 Data Management Plan
Work package number and title	6 Community and capacity building
Dissemination level	Public
Authors	S. Di Giorgio, T. Müller, E. Ortega, L. Paladin, I. Paredes Cisneros, S. Razick, approved by all BioNT partners
Delivery date	28.04.2023

DMP version 1.0

Project partners

1. EUROPEAN MOLECULAR BIOLOGY LABORATORY (EMBL)
2. BIOBYTE SOLUTIONS GMBH
3. HPC NOW CONSULTING SL
4. UNIVERSITETET I OSLO
5. UNIVERSITAT DE BARCELONA
6. INFORMATION CENTRE FOR LIFE SCIENCE (ZBMED)
7. SIMULA CONSULTING
8. ALBERT-LUDWIGS-UNIVERSITAET FREIBURG
9. ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE (EPFL)

HISTORY OF CHANGES		
Version	Publication Date	Changes
1.0	2023-04-28	• First version

Table of Contents

Data Management Plan

This document describes the Data Management Plan for the BioNT project. It provides an overview of the data management policies that will be encoded and how they will be implemented throughout the project, with regards to all types of data generated in the Action. This Data Management Plan is a living document: it will be updated with details or relevant changes whenever necessary along the project duration. 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.

Project Abstract

BioNT - BIO Network for Training - is an international consortium of academic entities and SMEs with the aim to provide a high-quality **training program and community** for digital skills relevant to the **biotechnology industry and biomedical sector**. The basic curriculum of the program aims at employees or job seekers with limited prior experience in handling, processing and visualising biological data and/or the usage of computational biology tools. The advanced curriculum is targeted at Data Stewards, System Administrators, and other professionals who hold leading roles in the management and deployment of computational resources at their workplaces. It is building upon the strong expertise in digital literacy training of its members as well as on its wide network of collaborations and other initiatives aimed at **professionalising Life Sciences data management**, processing and analysis skills. Standardisation and state-of-the-art technological implementation will reinforce the training and outreach activities plan. The program builds on short dedicated remote workshops as well as training recordings and will generate specific training material as **Open Educational Resources** which will be made available in a dedicated web platform. The consortium structure and high connectivity to relevant networks will foster **project sustainability** and efficient knowledge delivery to the stakeholders throughout the entire duration of the project and beyond, all this under a well defined and transparent management system.

Introduction to the Data Management Plan

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 main sections for the two categories. Additionally, it includes a section regarding the **Storage and back-up** and a section about **Access and security**, both applicable to both Data and Software. Finally, **Legal aspects** and the **Resources** needed for the implementation of this Data Management Plan 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 format

This section includes details on the types (e.g. images, 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, for the participants selection and to be aggregated to measure training impact in BioNT reports.
2. Mailing lists: Generated to facilitate communication with participants and other stakeholders throughout the project.
3. Information about target audience for course advertisement: Details of the organisations and individuals who will be targeted in the project's outreach and marketing activities.
4. Training materials: Documents, presentations, and other materials used to deliver training and educational content to participants.
5. 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.

6. Input data used during the workshops: Datasets used in the workshop practicals, to exercise and test data analysis and processing skills acquired.
7. 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.
8. Internal meeting notes: Recorded during consortia / event preparation meetings.
9. Chat logs: Generated during the online training sessions, to be used for documentation purposes.
10. Survey data: Data from surveys designed to assess the effectiveness of the training program and to gather feedback from participants.
11. Website content: Descriptions of the training program, participant testimonials, and other relevant information.
12. Presentations delivered in conferences: Presentations to showcase the training program at conferences and other events.

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. All data will be stored in secure, backed-up locations, and will be made available to other researchers and stakeholders in accordance with applicable policies and regulations. 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), email list (Mailing list software such as Mailchimp)
Information about entities for advertisement	Spreadsheet (CSV), structured text (JSON, XML)
Training materials	PDF, HTML, Markdown
Images and illustrations	SVG, PNG

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)
Zoom chat logs	Text file (TXT), JSON
Survey data	Spreadsheet (CSV), online survey tools (SurveyMonkey, Google Forms)
Website content	HTML, CSS, JavaScript
Presentations delivered in conferences and meetings	PDF, HTML, Markdown, Google Slides

The description of the Data Type “Input data used during workshop” will be expanded in following versions of this document. Compressed folders (e.g. ZIP or TAR.GZ) might be used, mainly for sharing purposes during the training, but an uncompressed version of the data will always be archived for backup.

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:

- 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 **Google Workspace**. All partners will have access to this drive. 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. More information about the usage of this platform can be found in section Document storage practices.

- For backup, versioning, and project monitoring purposes, after the live editing session the documents will be migrated to a Markdown-based format and transferred to a repository in **GitLab.com**. The repository in GitLab.com (only accessible to the project partners) also serves to monitor the project progression through issues, tasks,

boards and milestones. While an extensive description of the project management tool is beyond the scope of this document, this will be thoroughly detailed in the BioNT Dissemination and Communication Plan. Should there be a need to store the data in a repository physically-hosted in Europe, the migration of the entire repository to a GitLab instance hosted by a BioNT partner, such as EMBL or ZB MED, will be considered. The repositories' content and functionalities are entirely compatible across public or institutional GitLab platforms, making the migration a straightforward process. The migration of the repository is currently deemed unnecessary as it will not contain any personal data of project participants. More information about the usage of this platform can be found in section Document storage practices.

- GitLab.com (or GitHub.com, in the case of the training developed in the Galaxy Network framework) will also be used to collaboratively design and develop training materials, with one repository for each course. The structure of these folders is detailed in section Lesson folder structure. Lesson materials, after the course delivery, will also be deposited 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.

- The course registration, participant selection and result notification, will be managed through the platform **members.cecam.org**, managed by the project partner EPFL. This implies that all participant data (as well as most of the data related to each workshop instance, e.g. the dates) will be hosted there. More information about this is included in the section Legal aspects.

- Video and audio materials generated during the training program will be uploaded to the **TIB AV-Portal**, an open-access platform for scientific videos. Justification of this choice is available in the section Storage and back-up.

- The **BioNT website** will also include 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 will be advertised on the **Digital Skills and Jobs Platform**. To provide potential participants with relevant information about the courses, each entry in the platform will consist of a brief course overview, participant eligibility criteria, enrollment instructions, and contact details of a designated point of contact. 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 a mailing list software. 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	Mailing list software, Google Workspace (continuous update)
Training materials	GitLab / GitHub
Images and illustrations	members.cecam.org, Google Workspace, GitLab / GitHub, Website
Input data used during workshops	GitLab / GitHub
Video recordings of lessons	TIB AV-Portal
Internal meeting notes	Google Workspace (while editing), GitLab (for storage)
Zoom chat logs	GitLab (for storage)
Survey data	GitLab
Website content	Website server
Presentations delivered in conferences and meetings	Google Workspace (while editing), GitLab (for storage)

Reuse of existing data

This section lists existing data sets that will be used and specify the terms of use (e.g. licence, collaboration with the data producing group). When re-using 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 Galaxy and CodeRefinery, follow the same approach.

Currently, it is expected that most of the new training material will be 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 delivering 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.

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 in 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.

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, 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 be applicable to all platforms where data and documents will be stored (e.g. GitLab and Google Workspace).

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. To ensure that each version is uniquely identifiable, each file will begin with the date that it was created, following the ISO 8601 standard (e.g., 2023-04-03), and will end with a version number identifier, if necessary.

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, links to relevant documents will be provided in the project documentation, with further details about the data organisation, including the directory structure and naming conventions.

Lesson folder structure

Following the previous section, a folder structure with clear names and descriptions will be used to organise the training materials and data produced during the project in the GitLab repository. Making these folders homogeneous in terms of content is particularly important since they will be publicly shared. The structure below, with the given sub-folders, will be used for all the lessons that will be generated in the project:

- Training_materials: Containing all the documents related to the training program, including manuals, presentations, and handouts. This folder will also include scripts used by the course participants and instructor along the course. Access to this folder will be granted to all project members and collaborators. The e-learning versions of the training materials will be made public (under the licence CC-BY) and openly-accessible through the project website.

- Survey_data: Containing all the anonymised data generated during the training program, including attendance records, surveys, and evaluations. Access to this folder will be granted to project members responsible for data collection and analysis, as well as project managers.

- Survey_code: Linking/including all the code and scripts used to analyse the survey data and other data sources not included in the training materials. Access to this folder will be granted to project members responsible for data analysis and code development.

- Survey_results: Containing all the results and outputs generated from the analysis of the survey's data. Access to this folder will be granted to project members responsible for data analysis and project managers.

- Admin: Containing all administrative and management documents related to the specific course (e.g. notes from the preparation meetings among BioNT partners). Access to this folder will be granted to all BioNT partners.

This structure is subject to change during the course of the project, and additional folders and subfolders may be added as needed to maintain an organised and clear file structure.

Website

BioNT will have a dedicated website to provide information about the training materials and related events. The website will be developed using modern web technologies to ensure compatibility with a wide range of devices 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, IP address: 134.95.56.207 and under the registered domain “biont-training.eu”. This is a secure server that will be regularly backed up to prevent any data loss. More details about the website content will be available in the Dissemination and Communication Plan for BioNT.

The website will include 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 will be announced. It will also include links to the project's social media accounts, with updates and information about the training program. In addition, the BioNT website will provide information about the 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 will be used for version control and documentation. Each lesson generated will have its own repository, with a corresponding webpage for ease of access. Within each repository, a README file will be included, containing all relevant information about the lesson, such as methodology used to collect the data, and data processing and analysis steps. Additionally, a contribution guide will be included to increase accessibility and ensure the quality of contributions. The community metadata standards, such as Schema.org, will be used to improve the findability of BioNT’s training materials on the web.

Metadata

To ensure that the data are discoverable and interpretable, metadata that describes the contents of the training materials will be provided. The metadata will follow community metadata standards, such as Dublin Core or DataCite, and will include the following information:

- Descriptive title
- Name of the responsible organisation
- Relevant dates
- A brief description, including the purpose, scope, and methods used
- Format
- Licence

Metadata will be included in the README files in each repository on GitLab. This metadata will help users to discover, understand and reuse the training materials.

Documentation

All relevant information will be documented to enable interpretation and reuse of the training materials. The documentation will be updated each time a new version of any official document is created, as an explicit history of changes will be kept for each file (as for this one). The general guidelines on how to navigate the project folders (also documenting the project structure) will be included in the README files in each repository on GitLab and Google Workspace.

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 will be managed using Git, a version control system. All partners, during the development of the training materials, will store their code in a code repository, such as GitLab. The project management team will supervise 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. Some courses will also include workflows, generated for training purposes. These courses will be deposited in open platforms such as WorkflowHub. The project team will ensure that the workflows are compliant with the FAIR principles and have persistent identifiers assigned to them.

In summary, the BioNT project will be managed effectively, and the project's software will meet the FAIR principles for software publication. Code repositories will be used, complying with these principles. Further details about the specific tools used will be added to this Data Management Plan when the exact software needs of each course are clarified (at the beginning of each course design phase).

Storage and back-up

This section describes where the data and metadata will be stored, including the back-up strategy.

All the training materials, data, and metadata will be stored on multiple platforms to ensure accessibility in the different phases of usage (while collaboratively editing them or while storing them for backup). The EMBL Google Workspace platform will be used as a cloud storage and collaborative editing tool for meeting agendas documents, presentations and other files that may require easy access and tracked live editing.

The GitLab repository will be used to store these types of documents once in the backup phase, as well as the training materials, data, and metadata related to the BioNT training. The mailing lists will be managed through an internally managed list server, which will ensure data security and also that all project partners have access to all communications.

As anticipated, workbench materials, such as code and scripts, will also be maintained in GitLab, which will make it more accessible and open to external contributors. This will ensure that the training materials and analysis workflows are open and accessible to the wider community, which will promote collaboration and improve the quality of the training program.

Video and audio materials generated during the training program will be uploaded to the TIB AV-Portal (https://av.tib.eu/), which is an open-access platform for scientific videos. The TIB AV-Portal automatically assigns a DOI to each video, which will ensure that the video can be easily cited and referenced. The metadata for each video will be generated according to the website guidelines, which will ensure that the videos are discoverable and accessible to the scientific community.

In addition, for each official version of the lessons, persistent identifiers such as DOI will be applied to the entire lesson materials folders, including code (e.g. by using Zenodo), to ensure its long-term accessibility.

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, encryption, and secure communication channels. In addition, regular backups of the data will be performed 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 their internal server. 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 that BioNT will need. 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 deleted afterwards,
- 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, this document serves as internal agreement with all the project members that have access to the personal data and its terms are agreed among all project's members.

Data ownership and intellectual property

This section details who is the data owner, 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 the 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 from 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. The 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 data management plan for BioNT involves several aspects related to the storage, access, preservation, and sharing of project data.

Firstly, and as already introduced, all data and metadata generated by the project will be stored in multiple platforms to ensure their availability and accessibility depending on the phase of their life-cycle. These platforms include Google Workspace, as a collaborative editing platform, and Gitlab, used for the documents backup, versioning, project monitoring and where the training material will be maintained (more details are included in the previous paragraphs). The video and audio material will be uploaded to the platform av.tib.eu, and a DOI will be automatically assigned to it. The textual training material generated will also be deposited with a DOI e.g. in Zenodo. The repositories and archives will be provided with persistent identifiers to ensure the long findability of the material.

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 dataset interoperable, providing documentation and software needed to reproduce and validate the code used in the training materials, and finally to minimise access restrictions 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.

The data owner is the BioNT project. Intellectual property rights are not affected when using the workbench provided by The Carpentries. If the modified materials are pushed back to The Carpentries incubator, The Carpentries will claim intellectual property. The project will host the workbench into an internal server to avoid the claim of intellectual property by The Carpentries, and eventually consider adoption of the materials by The Carpentry community only at the end of the project duration, as a measure to ensure sustainability.

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 Data Management Plan will be the responsibility of the team in the coordinating institution, EMBL, who will communicate with the project stakeholders as needed. The project manager Isabela Paredes Cisneros (isabela.paredes@embl.de) is responsible for coordinating data management responsibilities across partners and informing the project stakeholders of any relevant change or update. The project manager is also the contact reference for this document, and can be contacted for any queries related to it.

Resources

This section identifies the resources needed to implement this Data Management Plan 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 dataset, software code, and training materials for all BioNT training repositories maintained in GitLab is estimated to be of at least 2 GB. This estimation is based on the model of the CodeRefinery repositories, similar to BioNT in terms of materials style. Following the data

management practices described in this document, the lesson materials data storage is foreseen to have a small footprint. 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. 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 will be needed to facilitate data transfer and collaboration among project partners. In addition, access to high-performance computing (HPC) resources for the relevant courses might be required. To ensure that project data is only accessible to authorised personnel, access control mechanisms will be needed. Considering the project commitment to open source tools and formats, no purchase of specialised software has been anticipated. More detailed technical requirements of each course will be added to this document at later stages.

- Data steward/data scientist: The expertise of the project's technical coordinator at EMBL (data scientist), with the support of the Data Management Coordinator (a Data Sciences position/office at EMBL), will provide ongoing support for data management tasks, including the implementation of the data management plan, metadata creation, data curation, and long-term preservation. This person will work in the project management team to ensure that data is managed effectively throughout the project lifecycle.

- Training needs: All project partners will need training in data and software management best practices to ensure that they can effectively implement the data management plan. The coordinating team includes a sufficient level of expertise to train the project staff on this, and will do so on occasion of the Supervisory Boards meetings or on demand organising 1:1 sessions, if needed.