H.F.R.I.
Hellenic Foundation for Research & Innovation

D1.2 – Data Management Plan

BINGO

Brain Imagined-Speech Communication

Greece 2.0
NATIONAL RECOVERY AND RESILIENCE PLAN

Funded by the European Union
NextGenerationEU

Funded by the European Union
NextGenerationEU

Greece 2.0
NATIONAL RECOVERY AND RESILIENCE PLAN

H.F.R.I.
Hellenic Foundation for Research & Innovation

The research project is implemented in the framework of H.F.R.I call “Basic research Financing (Horizontal support of all Sciences)” under the National Recovery and Resilience Plan “Greece 2.0” funded by the European Union – NextGenerationEU (H.F.R.I. Project Number: 15986).

Dissemination level:	Confidential (CO)
Contractual date of delivery:	Month 6, 27/05/2024
Actual date of delivery:	Month 6, 24/05/2024
Work Package:	WP1 – Project Management
Task:	T1.1 Administrative and financial management
Type:	Report
Approval Status:	Final
Version:	v1.0 (will be updated on M21)
Number of pages:	19
Filename:	D1.2_DataManagementPlan_v1.docx

Executive Summary: BINGO is deeply engaged in the collection and analysis of EEG data to decode imagined speech, contributing to advancements in Brain-Computer Interfaces (BCIs). To achieve this efficiently, the project implements structured data acquisition and evaluation protocols.The present document outlines the data generation and management processes involved in BINGO's experimental studies and analyses. It provides a detailed description of how EEGs are collected, at what stages, for what specific purposes, along with the methods employed for data processing. Additionally, the report defines the protocols for data flow, backup procedures, and secure storage, ensuring compliance with ethical and legal frameworks. Given the sensitive nature of neural data, the document also establishes measures to safeguard participant privacy and maintain strict data protection standards. Furthermore, an assessment of BINGO's alignment with FAIR (Findability, Accessibility, Interoperability, and Reusability) principles is conducted to enhance data transparency and reusability. As a living document, the Data Management Plan (DMP) will be continuously reviewed and updated throughout the project's duration to reflect technological advancements, regulatory changes, and evolving scientific needs.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

HISTORY

Version	Date	Reason	Revised by
v0.1	18/01/2024	Table of Contents	Fotis P. Kalaganis
v0.2	02/05/2024	Input in different sections	Fotis P. Kalaganis
v1.0	24/05/2024	Final Draft	Spiros Nikolopoulos

AUTHOR LIST

Organization	Name	Contact Information
CERTH	Fotis Kalaganis	fkalaganis@iti.gr
CERTH	Kostas Georgiadis	kostas.georgiadis@iti.gr
CERTH	Spiros Nikolopoulos	nikolopo@iti.gr
CERTH	Ioannis Kompatsiaris	ikom@iti.gr

ABBREVIATIONS AND ACRONYMS

BCI	**Brain Computer Interface**
DMP	Data Management Plan
DOI	Digital Object Identifier
EEG	ElectroEncephaloGram
GDPR	General Data Protection Regulation
LSL	Lab Streaming Layer
SOP	Standard Operating Procedure

Contents

INTRODUCTION

BINGO project's **Data Management Plan (DMP)** serves as a structured framework to oversee the complete data lifecycle, covering all data collected, processed, and generated throughout the project. Given BINGO's focus on decoding imagined speech from EEG signals, the DMP is essential for ensuring data integrity, accessibility, security, and compliance with ethical and legal standards. At the outset, the DMP establishes a rigorous approach to data collection, emphasizing high-quality **EEG recordings, standardization of experimental protocols, and adherence to neuroscientific best practices**. It defines the **types of neural and metadata** to be acquired, the **signal processing techniques**, and the responsible research teams overseeing data acquisition. This foundation is critical for ensuring **consistent and reproducible** research outcomes. As the project progresses, the DMP will define **structured pipelines** for **data preprocessing, feature extraction, and machine learning-driven analysis** of neural signals. It will outline protocols for **data storage, backup, and cybersecurity measures** to prevent loss or unauthorized access. Furthermore, **metadata annotation and documentation standards** will be implemented to facilitate collaboration among project members and ensure that future researchers can efficiently interpret and utilize the data.

In later stages, the DMP will guide **data dissemination and sharing** while balancing **open science principles with privacy considerations**. Given the **sensitive nature of neural data**, special attention will be given to **anonymization techniques, informed consent policies, and secure data-sharing mechanisms**. The plan will also include strategies for **long-term data preservation**, ensuring that the datasets and benchmarking frameworks produced by BINGO remain accessible for future advancements in Brain-Computer Interfaces (BCIs) and neurotechnology. As a **dynamic document**, the DMP will evolve alongside the project, continuously adapting to technological advancements, regulatory changes, and emerging scientific needs. Ultimately, this plan will reinforce BINGO's **scientific credibility, reproducibility, and long-term impact**, fostering advancements in **speech imagery decoding and broader neural engineering applications**.

DATA SUMMARY

TYPE OF STUDY

The BINGO project is a **basic science study** focused on investigating the **neural mechanisms of imagined speech** using **EEG-based Brain-Computer Interfaces (BCIs)**. The study is structured around an **iterative experimental protocol**, wherein **electrophysiological data** (EEG signals) will be **collected, analyzed, and refined** through an adaptive approach. This process ensures that the **imagined speech decoding models** developed in BINGO are **scientifically robust** and **incrementally improved** throughout the project's duration. BINGO's methodology follows a **controlled experimental design**, where participants will perform **imagined speech tasks** under specific conditions. EEG data will be collected **in multiple trials**, using **predefined stimulus prompts** (phonemes, syllables and words) while participants (N = 30) imagine pronouncing them. The collected EEG signals will be analyzed using **machine learning models**, with an emphasis on **interpretable neural decoding techniques**.

Data collection will follow a **two-tiered structure**:

- **Primary Data Collection:** Raw EEG signals recorded during the imagined speech experiments, synchronized via the **Lab Streaming Layer (LSL)**.
- **Secondary Data Collection:** Metadata such as participant demographics, behavioral responses, and task performance, alongside neurophysiological insights from existing datasets for comparative analysis.

The BINGO study ensures compliance with **ethical guidelines and data privacy regulations**, particularly concerning **human participant research and neural data confidentiality**.

TYPES OF DATA COLLECTED

BINGO project will generate and collect multiple **types of data** at different stages of the study.

Data Category	Description	Acquisition Method
Raw EEG Data	Time-series recordings of brain activity during imagined speech tasks.	EEG acquisition system (Wearable Sensing DSI-24)
Preprocessed EEG Data	Artifact-reduced and filtered EEG signals.	Signal processing pipelines
Event Markers	Time-stamped markers indicating stimulus presentation and task phases.	Lab Streaming Layer (LSL) synchronization
Task Performance Data	Participant responses (e.g., reaction times, behavioral metrics).	Experimental task software
Demographic Information	Participant details (age, gender, handedness, language proficiency).	Digital consent forms & questionnaires
Cognitive & Linguistic Data	Linguistic and cognitive baseline assessments.	Standardized neuropsychological tests
Metadata	Experimental protocol settings, EEG electrode placement details, session logs.	Study documentation

Data collection follows **standardized protocols** to ensure reproducibility and comparability across study sites. The datasets will be used to train and validate **imagined speech decoding algorithms**, with a focus on **incremental vocabulary learning and cross-linguistic neural analysis**. All collected data will be securely stored, anonymized, and handled in compliance with **GDPR** and **FAIR principles**, ensuring accessibility, interoperability, and long-term usability.

DATA QUALITY ASSESSMENT

Data quality is a cornerstone of the **BINGO** project’s Data Management Plan, ensuring the integrity, reliability, and fitness for purpose of the data collected throughout the project's lifecycle. The methodology for assessing data quality will adopt systematic approaches that identify errors, inconsistencies, and inaccuracies in datasets. By following well-established standards and best practices, BINGO project ensures the usability and trustworthiness of the data used in the project, leading to robust, data-driven outcomes. While there is no one-size-fits-all solution for addressing data quality concerns, the project will provide a comprehensive methodology that team members can adapt to their specific data contexts. The practices and approaches will evolve as the project progresses, underscoring the project’s commitment to continuous improvement and adapting to new challenges that may arise over time.

DATA QUALITY FRAMEWORK

DATA COLLECTION AND IDENTIFICATION

Achieving the objectives of BINGO requires gathering diverse datasets from a variety of sources. Since the project involves complex biomedical research, data collection must be approached holistically. It is essential to consider data collection in conjunction with other aspects of the project to ensure the data's suitability for its intended purposes. Focusing on collecting high-quality data initially is a proactive approach to mitigate risks associated with unusable or unsuitable data later in the project. The first step in the BINGO project's data collection process involves identifying potential data sources. Additionally, **open datasets** will also be considered and evaluated. To streamline the evaluation of data quality and suitability for the BINGO project, team members will be required to provide the following details during the cataloging process:

- Dataset Name or ID
- Dataset Description (brief overview)
- Links to the dataset
- Dataset Owner/Provider
- Type/Format of the data
- Access Options (public, restricted, etc.)
- Usage and Access Restrictions (if any)
- Time restrictions (if applicable)
- Privacy, Legal, and Ethics Concerns
- Value to the BINGO Project
- Suggested by
- Comments

DATA QUALITY ASSESSMENT

Data quality assessment is essential to ensure that the data collected and used in the BINGO project is accurate, reliable, and suitable for analysis. The following methodology is recommended for assessing the quality of data throughout the project:

- **Understand the Data**: Begin by thoroughly understanding the data. This includes familiarising the project team with the data sources, available documentation, and the context in which the data was collected.
- **Assess Internal Validity**: Evaluate the internal consistency and reliability of the dataset. This includes checking for any internal contradictions or discrepancies.
- **Assess External Validity**: Evaluate the generalisability of the data—how applicable the data is to broader contexts or populations.
- **Assess Missing Data**: Identify any missing data points and determine the impact this has on the completeness and accuracy of the dataset.
- **Assess Duplicate Records**: Identify and resolve any duplicate records that may skew analysis or lead to inaccurate conclusions.
- **Assess Standardisation and Normalisation**: Review the data for consistency in format, dates, measurement units, and other variables. The project will assess whether the data aligns with standard health codifications and mappings (e.g., **ICD-10**, **HL7**, **FHIR**).
- **Assess Data Against Clinical Suggestions**: Collaborate with clinical team members to check data integrity through range checks, format checks, and other clinical checks to ensure the dataset's reliability and accuracy.

- **Assess Data Mapping Capabilities**: Evaluate the ability to integrate and merge datasets from different sources. This includes assessing whether datasets align in terms of format, units, and structure, and whether they can be merged for comprehensive analysis.

This quality assessment process will be implemented consistently across all datasets collected during the project and will guide the continuous monitoring of data quality throughout the project's lifespan. The methodology will adapt based on the unique characteristics of each dataset and the specific research requirements of the BINGO project.

DATA COLLECTED FROM PARTICIPANTS

The BINGO project involves **systematic data collection** from participants to analyze the neural processes of imagined speech. The collected data encompasses **neurophysiological signals, cognitive assessments, and demographic information**, ensuring a **comprehensive approach** to understanding and decoding imagined speech from 30 participants.

1. Neurophysiological Data (EEG Recordings)

- **EEG Signals:** High-resolution brain activity recordings during imagined speech tasks.
- **Electrode Placement:** EEG electrodes positioned according to the **10-20 system**, with a focus on **Broca's and Wernicke's areas**, which are crucial for speech processing.
- **Event-Related Data:** Markers indicating stimulus presentation, participant response timing, and experimental conditions.
- **Signal Preprocessing:** Filtering, artifact removal (e.g., eye blinks, muscle activity), and feature extraction.

2. Cognitive & Behavioral Data

- **Baseline Cognitive Assessments:** Standardized neuropsychological tests to evaluate **language processing, working memory, and executive function**.
- **Task Performance Data:** Reaction times, accuracy rates, and error patterns during the imagined speech tasks.
- **Linguistic Assessments:** Vocabulary range, bilingual proficiency, and phoneme recognition abilities.

3. Demographic & Participant Metadata

- **Basic Demographics:** Age, gender, handedness (left/right), native language, and multilingual proficiency.
- **Medical & Neurological History:** Information regarding past or current neurological conditions that may influence EEG recordings.
- **Participant IDs & Session Logs:** Anonymized participant identifiers and timestamps of data collection sessions.

4. Experimental Metadata

- **Session Configuration:** Experimental protocol parameters (e.g., stimulus type, cue modality).
- **Environmental Factors:** Room conditions (e.g., noise levels, lighting) that may influence EEG signals.
- **EEG System Calibration Logs:** Data ensuring consistency across different recording sessions.

All participant data will be **pseudonymized and securely stored**, adhering to **GDPR and ethical research guidelines**. The data will be used to train and refine **machine learning models for imagined speech decoding**, ultimately contributing to **the development of robust Brain-Computer Interface (BCI) applications**.

5. Format and Scale of the Data

The data collected from participants in the BINGO project will be structured and organized to ensure ease of analysis, reproducibility, and scalability. Below is a description of the **format** and **scale** of the data generated throughout the study:

1. Data Format

- **EEG Data:**
  - **Format:** The raw EEG signals will be stored in **EEG-specific file formats** (e.g., **.edf**, **.bdf**, or **.fif** for compatibility with software like **EEGLAB** or **MNE-Python**). Preprocessed data may also be exported to **.csv** or **.txt** formats for further analysis or sharing.
  - **Sampling Rate:** Data will be recorded at high sampling rates (e.g., **500 Hz to 1000 Hz**) to capture the necessary temporal precision of neural activity.
  - **Channels:** The EEG recordings will involve **multiple electrode channels** (typically 8-19 channels) based on the **10-20 system**, focusing on regions like Broca's and Wernicke's areas. Specific channels might also focus on the occipital and parietal regions for cue analysis.
- **Behavioral and Cognitive Data:**
  - **Format:** Cognitive assessments, questionnaires, and task performance will be stored in structured formats such as **.csv**, **.xls**, or **.json** for easy access and further processing.
  - **Data Types:** This data will include numerical values (e.g., task accuracy, reaction times), categorical data (e.g., participant demographics, linguistic background), and textual data (e.g., participant responses to questionnaires).
- **Medical History and Demographic Data:**
  - **Format:** This data will be stored as **structured tables** in **.csv**, **.xls**, or **.json** formats, ensuring ease of integration with other datasets for participant analysis.
  - **Data Types:** Medical history will include categorical variables (e.g., previous neurological conditions, handedness), while demographic data will include age, gender, and linguistic background in numerical or categorical form.

2. Scale of the Data

- **Participant Scale:**
  - The project will involve 30 participants.
  - This results in the collection of 30 **EEG samples**.
- **Data Volume per Participant:**
  - **EEG Data:** Each participant will provide data across multiple sessions, with each session generating a large volume of raw EEG data (e.g., hundreds of megabytes per session depending on session length and the number of channels).
  - **Cognitive and Behavioral Data:** This data will be smaller in scale, typically generating tens to hundreds of kilobytes per participant, depending on the number of assessments and the complexity of the questionnaires.
  - **Medical History & Demographic Data:** The volume will be relatively small per participant but necessary for the overall participant profiling and subgroup selection (e.g., less than 10 KB per participant for these fields).

3. Data Integrity & Security

- All data will be securely stored in encrypted formats, ensuring compliance with **GDPR** and other relevant data protection regulations.
- The data will be hosted on secure, centralized servers with strict access control policies to protect participant confidentiality.
- Regular backups and integrity checks will be implemented to ensure the preservation of data over the course of the project.

4. Data Access & Sharing

- While the data will be made available to internal research team members for analysis, certain aspects of the data, such as **EEG signal processing algorithms** and **de-identified datasets**, may be shared publicly via platforms like **OpenNeuro** or the project’s **public repository** after appropriate anonymization.
- A **data management and sharing policy** will be in place to ensure that shared data is **well-documented**, **adequately anonymized**, and **used ethically** by external collaborators and stakeholders.

FAIR DATA

MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g., persistent and unique identifiers such as Digital Object Identifiers)?

Yes, all datasets generated and used in the BINGO project will be accompanied by metadata and assigned a unique identifier. New data produced during the study, such as EEG recordings, behavioral assessments, and imaging data, will be uploaded to appropriate data repositories, and each dataset will be assigned a **persistent identifier** such as a **Digital Object Identifier (DOI)** to ensure its discoverability, identification, and long-term access. This approach ensures that each dataset can be traced, accessed, and properly attributed in future research.

What naming conventions do you follow?

The naming conventions for datasets and files will be **standardized** and finalized as part of the ongoing project work. Initially, a consistent framework will be established for naming conventions related to EEG data, participant information, and associated metadata. These conventions will be documented in subsequent versions of the DMP, ensuring uniformity in naming across datasets and making it easier for external collaborators and systems to access the data.

Will search keywords be provided that optimize possibilities for re-use?

Yes, search keywords will be included in the metadata for each dataset. These keywords will be carefullyselected to optimize discoverability and reuse of the data. The keywords will be derived from the project's primary objectives and the types of data collected, such as "imagined speech," "EEG," "Brain-Computer Interface," "cognitive neuroimaging," and relevant cognitive domains (e.g., "speech perception," "phoneme decoding"). This strategy will enhance the likelihood of the datasets being found by other researchers working in related fields, thus maximizing their reuse potential.

Do you provide clear version numbers?

Yes, version control will be implemented for all datasets, ensuring transparency and traceability throughout the project. **Version numbers** will be applied as follows:

- **Initial drafts** of datasets will be labeled as **v0.1**.
- For minor revisions made before official submission, version numbers will increment in decimal format (e.g., **v0.1**, **v0.2**).
- **Final submissions** and any significant revisions will have the version number incremented by one (e.g., **v1.0**, **v2.0**), marking major milestones in the project.

These versioning practices will ensure that researchers are always able to track changes to datasets and access the most up-to-date information.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

In order to enhance the **findability**, **identifiability**, and **reusability** of the BINGO project's data, detailed metadata will be created for every dataset. The metadata will be structured to include both **descriptive metadata** and **administrative metadata**, allowing future researchers to interpret, cite, and use the data effectively. Key metadata elements will include:

- **Descriptive Metadata:**
    - Information about the data itself, including data type (e.g., EEG signals, behavioral data), the nature of the experiment (e.g., imagined speech task), and specific details such as the experimental protocol used, participant demographics, task performance, and context.
- **Administrative Metadata:**
    - Information about the data collection process, including the **date** and **time** of data collection, **location** of the study, **software** used for data processing, **file formats**, and **storage details**.
- **Bibliographic Metadata:**
    - For each dataset, the associated publication references or relevant research papers related to the data will be included. These references will help situate the data within the broader scientific context and allow researchers to cite the data correctly.

To ensure that the metadata is comprehensive and follows **international best practices**, **standard metadata schemas** will be employed. These include:

- **DataCite Metadata Schema** (https://datacite.org/): This schema is widely used for metadata related to datasets and will be applied to ensure the datasets are discoverable and can be cited properly.
- **ISA (Investigation/Study/Assay) Metadata Schema** (https://isa-specs.readthedocs.io/en/latest/isatab.html): This schema will be used to organize and describe experimental workflows, including the types of data collected and the methods used.
- **Fair Genome Metadata Model** (https://github.com/fairgenomes/fairgenomes-semantic-model): For data related to participant genomics and other bioinformatics, this schema will be applied to ensure compliance with **FAIR** principles (Findable, Accessible, Interoperable, Reusable) and enhance the overall quality of the metadata.
- **BIDS format of Nature journal (https://www.nature.com/articles/s41597-019-0104-8)**

This metadata will be made available in formats that are both **human-readable** and **machine-readable** (e.g., **JSON**, **CSV**, **XML**) to facilitate the sharing, discovery, and reuse of the datasets across different platforms. The metadata will also comply with **FAIR** principles, ensuring that it is openly accessible and reusable for the broader scientific community.

In sum, the BINGO project will adopt a comprehensive approach to metadata creation, ensuring that the data produced is **well-documented**, **easily discoverable**, and fully compliant with international standards for data sharing and reuse. This approach will not only facilitate internal use within the project but will also ensure that the data can contribute to advancing research in the broader field of **BCIs** and **neurotechnology**.

MAKING DATA INTEROPERABLE

Are the data produced in the project interoperable, allowing data exchange and reuse between researchers, institutions, organizations, countries, etc.?

Yes, ensuring the interoperability of data across research team members and beyond is a central focus of the BINGO project. Data collected during the project will be made available in widely used **standard formats** to facilitate easy exchange and reuse across researchers, institutions, and countries. The project will prioritize formats that are compatible with **open software applications** to maximize the potential for data re-combination and integration with other datasets from various origins. Once the project concludes, the shared datasets and associated metadata will be provided in these formats.

What data and metadata vocabularies, standards, or methodologies will you follow to make your data interoperable?
The BINGO project will utilize a combination of **standard metadata schemas** to ensure data interoperability across platforms and research domains:

- **DataCite Metadata Schema**: This will serve as the core metadata standard for capturing general information about the datasets, including details about the data's origin, authorship, and date of creation. DataCite ensures that each dataset is uniquely identifiable and can be reliably cited in academic publications.
- **ISA Framework (Investigation/Study/Assay)**: The ISA framework will be used to capture study-related details, including the experimental protocols, procedures, and assays. This framework helps to provide a detailed overview of the study design, ensuring that data is well-organized and contextualized for future use.
- **FAIR Genome Semantic Metadata Model**: This model will be integrated to capture any additional metadata related to genomics or other bioinformatics data that cannot be fully addressed by DataCite or ISA. The FAIR genome model ensures compliance with the **FAIR principles** (Findable, Accessible, Interoperable, and Reusable) and enhances the overall reusability of datasets across different research contexts.

These three schemas—**DataCite**, **ISA**, and **FAIR Genome**—will be combined into a structured and comprehensive metadata framework that addresses the complexities of the BINGO study, particularly the diverse data types, such as **EEG**, **neuroimaging**, and **cognitive assessments**. Additionally, the project will deliver a **common data model** and a **data catalogue** (an extension of **ADataViewer**). This catalogue will provide an organized overview of all data used within the project, including summary statistics and metadata information, ensuring that the datasets are easily accessible and interpretable by external users.

Will you be using standard vocabularies for all data types present in your dataset, to allow inter-disciplinary interoperability?
Yes, standard vocabularies will be used for all data types generated in the BINGO project to ensure **inter-disciplinary interoperability**. This approach ensures that datasets are understandable and usable by researchers from various fields, such as **neuroscience**, **cognitive science**, and **engineering**, while maintaining consistency across data types.

In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?
Yes, if the need arises to use uncommon or project-specific ontologies or vocabularies, the BINGO project will provide mappings to more widely used ontologies. For example, **DataCite** and **ISA** are standard metadata schemas, while the **FAIR genome semantic metadata model** is already mapped to well-established ontologies, such as **NCIT** (National Cancer Institute Thesaurus), **LOINC** (Logical Observation Identifiers Names and Codes), and **EFO** (Experimental Factor Ontology).
If a new metadata element is introduced that requires a project-specific vocabulary or ontology, it will be mapped to an appropriate, widely accepted ontology term using the **Ontology Lookup Service** by **EBI (European Bioinformatics Institute)**, ensuring continued compliance with international standards for data interoperability. This approach will help maintain the **long-term usability** of the data and ensure that it can be combined, cross-referenced, and integrated with data from other studies, further promoting the **reusability** and **re-combination** of the data in future research.

INCREASE DATA RE-USE (THROUGH CLARIFYING LICENCES)

How will the data be licensed to permit the widest re-use possible?

The licensing terms for data re-use in the BINGO project will be decided by the **Data Access Committee** in collaboration with the project team members. The aim is to adopt a licensing framework that maximizes reusability while respecting ethical, legal, and privacy considerations. The specific licensing model, such as **Creative Commons** or **Open Data Commons**, will be selected to facilitate broad access and reuse of the data, in accordance with the FAIR principles.

When will the data be made available for re-use?

The BINGO project will make data available for re-use as soon as possible after the completion of the relevant analyses, taking into account any necessary embargo periods for publishing results or seeking patents. If an embargo period is requested, the length and rationale will be clearly defined and communicated. This will ensure that external researchers can request access to the data after an appropriate delay, allowing sufficient time for the research team to publish and protect any intellectual property. The exact embargo duration will be determined as the project progresses and will be updated in the next version of the **Data Management Plan (DMP)**.

Are the data produced and/or used in the project usable by third parties, in particular after the end of the project?

The re-use of data after the end of the BINGO project will be possible, though certain data types may have restrictions. This may include data that is subject to participant consent, ethical constraints, or privacy regulations (e.g., sensitive health data). Decisions regarding the re-use of these data will be made in consultation with the **Data Access Committee**, and any restrictions will be clearly explained. The aim is to ensure the widest possible re-use, while maintaining compliance with ethical and legal requirements. Details about restricted data and the conditions for their use will be outlined in future versions of the **DMP**.

How long is it intended that the data remains re-usable?

The intended duration for data reusability will be determined as the project progresses. It is expected that data will remain publicly available and re-usable long-term, consistent with the **FAIR principles**. Further details about the long-term reusability of the data, including plans for data archiving and preservation, will be outlined in future updates of the **DMP**.

Are data quality assurance processes described?

At this stage of the BINGO project, detailed data quality assurance processes have not yet been fully defined. However, a **Quality Control Standard Operating Procedure (SOP)** will be developed and will include relevant work instructions. Data quality assurance will be aligned with existing standards and procedures for the different types of data collected (e.g., cognitive assessments, neuroimaging, biological samples). The **CERTH** will be responsible for coordinating data quality.

ALLOCATION OF RESOURCES

What are the costs for making data FAIR in your project?

At this stage, no specific costs for making data FAIR have been identified. However, the costs related to ensuring that the BINGO project’s data is findable, accessible, interoperable, and reusable (FAIR) are considered to be part of the project's regular operations. This includes the development of necessary infrastructure and personnel resources.

How will these be covered?

The costs associated with making data FAIR will be covered through the **HFRI** grant, as per the conditions of the Grant Agreement. These costs are included in the project's **Work Packages**, particularly those related to **data management**. The allocation of funds for these activities has been incorporated into the overall project budget.

Who will be responsible for data management in your project?

CERTH will be responsible for overseeing the management of all project data. This includes data collection, storage, curation, metadata creation, and ensuring compliance with the **FAIR principles** throughout the project's lifecycle.

Are the resources for long-term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

The **long-term preservation** of data is an integral part of the BINGO project's **data management** strategy. Resources for making the data available and reusable beyond the project's duration are covered by the project's personnel resources. No additional costs for long-term preservation have been foreseen at this stage.

ETHICAL ASPECTS

Are there any ethical or legal issues that can have an impact on data sharing?

Ethical and legal considerations will play a central role in the BINGO project's data sharing practices. All data shared will be subject to ethical review and approval by relevant ethics committees, ensuring compliance with national and international regulations. Ethical deliverables have already been generated, which will outline the project's ethical approach to data collection, processing, sharing, and storage. The project will adhere to all ethical standards concerning participant consent, privacy, and data protection. Any data shared will be anonymised or pseudonymised where appropriate, and access to data will be restricted based on ethical review findings and consent granted by the participants.

Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?

Yes, informed consent for data sharing and long-term preservation will be explicitly included in the questionnaires and study documentation that participants will review. The informed consent process will cover all aspects of data collection, use, sharing, and storage, in accordance with **General Data Protection Regulation (GDPR)** and other relevant data protection laws. BINGO will ensure that all participants are made fully aware of their rights regarding data protection, including the ability to withdraw consent at any time. Consent will be obtained through a secure platform, and participants will receive detailed information about how their data will be used, stored, and shared for future research purposes.

What is the scope and treatment of ethical issues?

The scope of ethical issues includes, but is not limited to, **informed consent**, **data protection**, **privacy**, and **potential risks to participants**. Ethical considerations will be integrated throughout the project lifecycle, from initial data collection to long-term data storage. Any data that contains personal or sensitive information will be carefully handled and only shared under strict conditions and following explicit participant consent. Data will be pseudonymised or anonymised wherever possible to protect individual privacy. The project will ensure that all research activities are ethically justified and aligned with the interests of the participants, with a focus on transparency in the study's goals, methods, and data management. **Ethics deliverables** will be reviewed regularly to ensure compliance with ethical standards.

Which security standards are relevant for your data? Are you bound by a confidentiality agreement?
Data security will be governed by the highest standards applicable to medical and research data. The **GDPR** guidelines, alongside institutional data protection policies, will be strictly followed to ensure the protection of personal data. Data will be stored securely in **cloud-based systems** (such as **Microsoft Azure**) or **local secure facilities**, with access granted only to authorised personnel. Confidentiality agreements will be signed by all project team members, ensuring that any sensitive information is protected, and that participants' privacy rights are upheld. Secure access to data will be granted only through encrypted channels and with strict authentication processes. Pseudonymised data will be used whenever possible to minimize any potential risk to participant privacy.

Do you have the necessary authorizations to obtain, process, store and further process the data?
Yes, all necessary **ethics approvals** have be obtained prior to the commencement of the study. These approvals will be granted by relevant national and institutional ethics committees, in line with applicable laws and regulations, including the **GDPR**.

Have the people whose data will be used been informed or given their consent?
Written informed consent will be obtained from all participants. Participants will be clearly informed about the study's objectives, how their data will be used, the measures taken to protect their privacy, and their rights regarding the withdrawal of consent.

What methods and precautions do you plan to use to protect personal data and other sensitive data?
To protect personal and sensitive data, the following measures will be implemented:

- **Data encryption**: All personally identifiable information (PII) will be encrypted and stored separately from research data, in accordance with data protection protocols.
- **Secure data storage**: PII will be stored on encrypted cloud servers (e.g., **Microsoft Azure**) or in **secure local facilities**.
- **Access control**: Only **approved personnel** (e.g., investigators, data managers) will have access to PII through a **dedicated administration portal**, requiring username and password authentication.
- **Data anonymisation and pseudonymisation**: PII will be pseudonymised wherever possible, ensuring that no direct identification of participants can occur without specific access to the key.
- **Access restrictions**: Outside of secure data storage systems, access to paper records and other personal data will be strictly controlled, with only essential personnel allowed to access sensitive information.
- **Daily backups**: Backup instances of the data will be created daily to ensure data security and recovery in case of system failure.

In all cases, data will be handled in line with the study's protocols and with full compliance with **GDPR** and other relevant data protection regulations.

How do you resolve copyright and intellectual property issues?
The intellectual property (IP) rights to data collected during the BINGO project remain with the PI. If data is transferred to other project research team members, the IP rights will not be altered. All project team members will have access to the data collected during the project period, but personal data will be pseudonymised prior to sharing to ensure compliance with data protection laws. The project team

members will agree upon the terms of **data sharing** and **publishing rights** to ensure fair use and dissemination of project results.

Who is the owner of the data?
The data ownership rests with the PI of the project. This will be clearly outlined in the data management agreements and project documentation.

Which licenses apply to the data?
The raw data collected during the project will be used exclusively within the context of the BINGO project and will not be shared externally without prior consent, except where permitted under the terms of the participants' consent. Evaluation data will be pseudonymised for future publications, ensuring that privacy and confidentiality are maintained.

What restrictions apply to the further use of external data?
As the BINGO project will primarily generate its own data, restrictions on external data usage are not expected to apply. Any external data used within the project will be subject to the permissions and restrictions of the data providers, ensuring compliance with data-sharing agreements and legal constraints.