European Cluster EMF and Health

Deliverable

Shared Data Management Plans between cluster partners

Document Summary Information

Start Date	01/09/2022	**Cluster URL**	www.emf-health-cluster.eu
Deliverable	Data Management Plans between cluster partners		
Working Group	WG2		
Contractual due date	31/08/2023	**Submission date**	31/08/2023 (initial), 15/04/2024 (revised)
Type	Report	**Dissemination Level**	Confidential
Lead Project	NextGEM	**Deliverable Editor**	Nikolaos Petroulakis (FORTH)
Contributors	NextGEM: Nikolaos Petroulakis, Maryse Ledent, Eleni Zarogianni, Panos Chatziadam, Andrea Gavrielides SEAWave: Theodoros Samaras, Caterina Merla, Ioannis-Anestis Markakis ETAIN: Anke Huss, Silvia Ariccio GOLIAT: Mònica Guxens		

This cluster is comprised from four projects that have received funding from the European Union's Horizon Europe research and innovation programme: ETAIN (101057216), GOLIAT (101057262) NextGEM (101057527), SEAWave (101057622). Funded by the European Union. Views and opinions expressed are, however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Table of Contents

Executive Summary

The current deliverable “Shared Data Management Plans between cluster partners” aims to share data management plans among partners and to agree on a common management of data produced within CLUE-H. That includes a short presentation of each project’s data management plans, to achieve the convergence and harmonization of data management plans. In addition, the presentation of a unified data storage and framework is also presented. Thus, the common approaches as well as common repositories and databases are also provided. Finally, the organization of outputs of individual projects and CLUE-H as a whole, is also described.

1 Introduction

The purpose of this section is to introduce the deliverable providing the outline of a shared data management plans (DMPs) with regards to the project's respective outputs and work performed.

Based on the objectives and work carried out in this deliverable, the document starts with the Executive Summary followed by the introduction of the document in Section 1.

Section 2 provides convergence/harmonization of data management plans.

Section 3 aims to establish unified data storage common approaches as well as common repositories and databases and data exchange for the development of a unified framework within CLUE-H.

Section 4 provides the projects' generated outputs organized in CLUE-H.

Finally, Section 5 concludes the deliverable.

2 CLUE-H projects data management plans

2.1 Summary of CLUE-H projects of data management plans

2.1.1 NextGEM data management plan

The purpose of the NextGEM DMP deliverable (D1.1, submitted on 26/11/2022) is to outline how data will be handled throughout the project's lifecycle improving their Findability, Accessibility, Interoperability and Reusability (FAIR) properties. NextGEM DMP documents the methods and procedures for collecting, organizing, documenting, storing, sharing and preserving data generated during the project. To coordinate and monitor the execution of its DMP, NextGEM has appointed a DMP task force, charged with applying rules meant to ensure an ethically permissible balance between data protection and accessibility. The DMP Task Force includes the partners in charge of the DMP and the data quality process, as well as partners from the various research output areas within NextGEM. It will also be in charge of defining how research outputs are processed and preserved, and how associated metadata (categories, topics, characteristics of the research outputs...) are set. Following the requirement "as open as possible, as closed as necessary", data and other research outputs will be open and reusable after the end of the project, as long as it is kept by the selected open repository.

2.1.2 GOLIAT data management plan

The GOLIAT DMP deliverable (D10.2, submitted on 30/11/2022) describes the data management life cycle within the GOLIAT project, including: i) the handling of research data during and after the end of the project; ii) what data will be collected, processed, and/or generated; iii) which methodology and standards will be applied; iv) which data will be open access; v) how data will be curated and preserved, including after the end of the project. The Data Management Plan ensures that the data generated in GOLIAT is appropriately managed to ensure its usability, access, and preservation.

GOLIAT will follow the FAIR principles of data sharing to maximize the findability, accessibility, interoperability, and reusability of data. Results will be published in open access peer-reviewed journals through which potential users will become aware of the data generated by GOLIAT. After publication of the results, data will be made available to external researchers to maximize their impact by encouraging secondary analyses to address other research questions.

2.1.3 SEAWave data management plan

SEAWave DMP deliverable (D11.2, submitted on 30/11/2022) describes the data management life cycle for the data to be collected, processed and/or generated by the project. As part of making research data findable, accessible, interoperable and reusable (FAIR), the DMP included precise information on:

- the handling of research data during and after the end of the project,
- what collected, processed and/or generated,
- methodology and standards applied,
- policy for data shared/made open access,
- policy for data curation and preservation (also during and after the end of the project).

The above points were all addressed in detail in the submitted DMP and this document can be found online at the project website where it is accessible to the public.

2.1.4 ETAIN data management plan

ETAIN has described in its DMP deliverable (D8.2, submitted on 22/12/2022) the data management plan of the ETAIN project, which covers handling of research data during and after the end of the project; type, formats, size and provenance of research data; data utility, data usage outside of the project, other research output, as well as FAIR principles, and includes data management until the end of the project. In summary, the Data Management Plan ensures that the data generated in ETAIN is appropriately managed to ensure its usability, access, and preservation.

In line with the other CLUE-H projects, ETAIN follows FAIR principles to maximize findability, accessibility, interoperability, and reusability of data. Publications in open access peer-reviewed journals will mean that potential users will become aware of the results obtained in ETAIN. Where possible, after publication of results, data will be made available to external researchers to maximize their impact by encouraging secondary analyses to address other research questions.

2.2 Convergence and harmonization of data management plans

As projects have different outcomes in terms of experimental activities, exposure assessment procedures, output data and tasks, a total harmonization of data format and type is practically impossible to achieve. This option would result in a complex and time-consuming effort with a limited practical interest.

Each CLUE-H project establishes its own means for managing, sharing and exchanging data, as described in its own consortium specific DMP. However, one of the scope of this WG2 output, is to harmonize metadata files in order to have a common and homogenized description of the different type or format of data, if possible. These metadata files will be made largely findable and accessible as described in Section 3. Integration into a metadata catalogue is being considered. Then these data will be accessible on each single repository in accordance with what reported on the single DMP of each project. The total exchange of data would have resulted in an extremely complex procedure and in some cases, this would have been even not possible for ethical issues of privacy and confidentiality of data related, for example, to human trials.

3 Establishment of a unified data storage/unified framework on data documentation

3.1 CLUE-H projects data types and structure

3.1.1 NextGEM data types and structure

Prior to the development of the initial version of the DMP, NextGEM completed a dedicated template aimed at defining the different outputs that might be generated through the different work packages and tasks. The information related to the outputs included the identification name in NextGEM, re-use of existing data, origin, title, category, purpose and relation to the objectives of the project, responsible person, data format, expected size of the file, status, location of the running document / raw versus processed dataset, back-up frequency of the running document /, location of the final document / final dataset, duration (preservation), ethics issues and privacy level/accessibility. As regards metadata aimed at providing descriptive, administrative and structural information on the outputs generated, the outputs generated by NextGEM will be structured according to these metadata, extracted from the open repository where outputs are uploaded. As regards the additional metadata, the aim is to provide further details on the different outputs, to make them more findable such as generic metadata identical for all domains and domain-specific metadata for searches and for preview of the outputs obtained. NextGEM has also developed templates pertaining to the project's presentation (in .pptx format) and how metadata trees will be updated (in .xlsx and .csv formats) where partners will be able to update their data requirements and provide inputs concerning types and ever evolving metadata associated with the DMP.

3.1.2 GOLIAT data types and structure

Different types of data generation with the purpose of fulfilling the specific objectives of each task of GOLIAT will be performed, in accordance with the "data minimization" purpose. GOLIAT comprises 25 sub-projects with different characteristics and objectives, each one corresponding to a different task of the project. Data generated within GOLIAT will have very different natures and formats. We will work with types of data that range from text to video, and from diverse sources like digital platforms and online questionnaires, cohorts, focus groups or surveys, in-situ exposure measurements, simulations, and experimental data from in silico to human. For each of the sub-projects, the different types and formats of data generated, its expected size, whether we will re-use data and which one, the origin of the data, to whom the data might be useful, and whether the data will be made openly available.

3.1.3 SEAWave data types and structure

The data in SEAWave are collected in text, numerical and image formats, following the associated DMP instruction, and they are stored in files whose extension are defined by the equipment/software used for the generation and the collection. Some examples include but are not limited to .xlsx (Excel spreadsheet), .ppt (PowerPoint), .mat (Matlab), .txt (text), .s2p (touchstone), .avi (video). Since either individuals or groups of researchers from the consortium partners will generate data, the identification of the partner that created the specific file or data set is also described in project's DMP.

3.1.4 ETAIN data types and structure

Depending on the specific objectives within ETAINs work packages and tasks data types and structure varies widely and have been described in detail in the Data Management Plan deliverable. Data types range for example from maps to algorithms to quantify RF-EMF dose, to quantification of RF-EMF exposure in insects, to videos, text, numerical data and so on. ETAIN will work together with the other CLUE-H partners on a harmonised meta-data collection to make individual project data FAIR.

3.1.5 Definition of a unified data type and structure in CLUE-H

It is important to set the different kind of outputs and defined common metadata. The working group will develop a common format to collect metadata from the individual projects, in order to enhance FAIR principles across the projects. The metadata will be updated regularly and published on the

CLUE-H website. Metadata will include a short description of the type of data, format, size, accessibility, ownership, and contact persons. As a matter of complying with the FAIR process, metadata will be assigned to each research outputs. European metadata standards will be used to make data interoperable with other data sets of similar type. Proprietary data formats will be avoided as much as possible. Research outputs will be made available in platform and software independent data formats.

The unified data type and structure in CLUE-H is established based on the work undertaken in this WG2 for the harmonization of a common metadata template based on the format keeping the common fields between the Zenodo (NextGEM, SEAWave), Dataverse (GOLIAT) and Yoda (ETAIN) that projects use for storing their outcomes. In this envisioned way, the created metadata will be able to be shared and stored in the different project and open databases, under a shared metadata catalogue within the NextGEM Innovation and Knowledge Hub (NIKH) platform, as described in the following subsection.

3.2 CLUE-H project data storage

3.2.1 NextGEM data storage

All data generated within the NextGEM shall remain at the corresponding's partner local premises, following security policies and procedures defined by the relevant partner. Additionally, research institutes of the NextGEM consortium retain the option to make any data and research outputs they wish publicly available either to open repositories, such as the Zenodo, the NIKH and the NextGEM's Nextcloud-based file repository. Simultaneously, the NIKH serves as a centralized repository, offering a standardized approach for storing metadata and managing project outcomes, EMF measurements, research data, and risk assessment information. NIKH achieves this by hosting a rich metadata catalogue, enabling efficient organization, searchability, and discovery of the stored data. This metadata catalogue provides an essential tool for stakeholders to locate relevant information, enhancing the overall usability and accessibility of the NIKH. Predominantly, the NIKH platform will host metadata records, detailing the structure, generic metadata information and also domain-specific (EMF-specific) information. In this way, the NIKH platform will become a central hub for the storage and access of EMF-related data, in that an authorized user of the NIKH platform will be able to browse through the metadata catalogue, in search of any data, documents, and/or research outputs, etc. someone is interested in and once the desired data or research output is identified, a separate procedure for secure data exchange is launched. Finally, the Nextcloud platform provides a feature-rich environment for sharing, synchronizing, and collaborating on project files, facilitating efficient teamwork.

3.2.2 GOLIAT data storage

Data collected from each partner will be digitized and stored on their institution's server which is subject to regular back-up that is controlled by their IT personnel. The IT department performs operations by type: mission critical (user data, virtual machines, scientific results, etc.) and static (scientific data sets, intermediate files, etc.). Content will be checked regularly to preserve its integrity, security, and durability. These procedures are designed, set and applied in order to fully comply with personal data as ruled by Regulation 2016/679 (General Data Protection Regulation (GDPR)) and other current national legislation and institutional regulations. Research team members will have an appropriate access level according to their role in the project. They will obtain the correct IT credentials for accessing the server. It will not be possible to download data from this server; instead, all analyses will be carried out remotely using analytical software installed on it. In order to satisfy data protection and ethical concerns, any information that might permit identification of survey respondents will be removed or obfuscated. GOLIAT will use Dataverse (http://dataverse.org), an open-source platform to share and archive data developed at Harvard's Institute for Quantitative Social Science. This repository platform allows a process for data access that ensures transparency and accountability for data providers, data requesters, and the data release decision making process. For the data that will be made open, data will be deposited at the CORA.RDR data repository within the Dataverse web application (https://dataverse.csuc.cat). For the data that will be restricted but usable by third parties under a proper agreement, each partner will store the data they will collect on their institution's server and will follow the data security procedure of their IT

department which includes regular back-ups and content checked regularly to preserve its integrity, security, and durability. Data will be stored for at least 10 years.

3.2.3 SEAWave data storage

SEAWave repository is based on a Zenodo account which guarantees a secure and solid space for data storage and management. The Zenodo system is managed by CERN assuring full access to data and periodic backup of the stored data performed regularly. It represents one of the most used systems for data sharing and storage and it is fully accepted by the EU.

SEAWAVE created a Zenodo repository available to all partners and in agreement with the policy of different availability of data access. Therefore, the generated data and files will be fully open (gold open access), subjected to restrictions individuated by the SEAWave consortium, or made available with an embargo period established again based on the agreement among the partners of the SEAWave consortium.

3.2.4 ETAIN data storage

ETAIN uses YODA, a research data management service (https://www.uu.nl/en/research/yoda). Yoda is Utrecht University's institutional research data repository. It is registered as such with re3data.org. Yoda complies with Utrecht University's Information Security policy for data classified as public, internal use, sensitive or critical. The YODA repository enables researchers from Utrecht University and their partners to securely deposit, share, publish and preserve large amounts of research data during all stages of a research project. YODA enables to establish a data management system on which ETAIN data and knowledge are FAIR (findable, reusable, accessible, interoperable), while applying ethical and privacy considerations where applicable. Regarding data storage of ETAIN data in CLUE-H, ETAIN will actively contribute to collecting meta-data, and to keep those data up to date until the end of the CLUE-H cluster.

3.2.5 Development of a unified data storage in CLUE-H

Currently, no unified data storage of individual project's results are planned. Data storage in CLUE-H pertains to CLUE-H activities (e.g. policy briefs, meta-data, inventories e.g. of measurement devices and approaches, etc). The corresponding information will be provided on the CLUE-H website, where it will be regularly updated (until the end of the CLUE-H lifetime. CLUE-H file repository, powered by Nextcloud and hosted in FORTH's premises, are securely hosted in a cutting-edge datacenter. The implementation prioritizes reliability, security, and performance. The datacenter's infrastructure ensures stable operation, stringent security measures protect valuable assets, high-speed connectivity and network redundancy optimize performance, and the Nextcloud platform, along with the NIKH's metadata catalogue, fosters efficient collaboration, seamless data exchange, and effective organization of project outcomes and research data.

3.3 CLUE-H project data exchange

3.3.1 NextGEM data exchange

In order to ensure the secure and sovereign data exchange, dedicated software components will be utilized within the NextGEM project. Specifically, Data Space services will be deployed to enable a decentralized data sharing approach, whereby data physically remain at their source and only transferred to another interested party once specialized, trustworthy exchange requests are issued. The Data Space Connectors constitute the core component of a data space architecture and allow the data exchange between involved participants. The parties involved in a Data Space scenario usually comprise of:

- A Data Provider, who is in possession of some data and would like to make them available and exchange them in a secure manner;
- A Data Consumer, who would like to become in possession of the data.

Each participant is represented by a Data Space Connector, which allows the registration of offered data resources, along with the metadata that describe them. Each participant can attach usage restrictions to their data and monitor data transactions through continuous monitoring. The Connector executes the complete data exchange process and ensures bidirectional communication

and enforcement of usage policies upon data, which ultimately leads to the bilateral agreement between involved parties. In NIKH, Data Connectors are located either at the user's local premises for NextGEM members (e.g., scientific research institutes), offering data resources or within the Data Spaces Services of the NIKH Service Layer for NextGEM members, non-certified users and Mobile App users (e.g. citizens) requesting access to specific data or intending to upload data to the NIKH Platform central storage.

3.3.2 GOLIAT data exchange

GOLIAT will produce three types of data:

- Data that will be made openly available and usable by third parties which will be available for download at the CORA.RDR data repository within the Dataverse web application;
- Data that will be restricted due to privacy concerns but usable by third parties under a proper agreement. Details of a contact person from the partner that owns the data will be provided and data access will be given by the partner conditional on relevant local ethical, legal, and social principles and signing a data transfer agreement. The fulfillment of these requirements will be solely evaluated and approved by the partner that owns the data;
- Data contains personal identifiable variables that will always be restricted.

An embargo will be applied for making the data of each sub-project open or allowing access to it until the scientific manuscripts planned in GOLIAT with this data are published, or 3 years after the end of the project, whichever comes first.

3.3.3 SEAWave data exchange

Data exchange within the SEAWave consortium will be managed using the Zenodo repository whereas data transfers between SEAWave partners working in the same WP will take place using each partner's institutional repositories, where a virtual area for data sharing will be created and used following the established procedures. In all cases the final data and procedures will be transfers within the SEAWave Zenodo repository following the file naming which is described in the project's DMP. Metadata will be shared in a common repository of the CLUE-H cluster or a new repository created for the Cluster on Zenodo. In any case, the type of repository for the CLUE-H project will be discussed based on accessibility, security and reliability of the platform.

3.3.4 ETAIN data exchange

ETAIN collects data that will be openly accessible (e.g. spatial maps of exposure) and that can be viewed and accessed freely. Other data are restricted but accessible, for example licensed data on insect models or the network planning tool. In these cases, data can become accessible via data transfer agreements if the respective conditions are met. Finally, some data underlie strict GDPR protection and will always be restricted.

3.3.5 Development of a unified data exchange in CLUE-H

Due to the confidentiality of some data and/or the requirement of ethical approval for others, related to human trials combined with the difficulty in harmonizing file formats and types, WG2 group decided to not share full data in CLUE-H but to share only metadata. Then metadata containing file and data set information will be available in the specific repository of each project under the established conditions of each single consortium i.e. full open access, with embargo or restricted access. The exchange of different processed data and metadata within the CLUE-H cluster is foreseen to be established through the development of a shared metadata catalogue based on the integration of open repositories and the NIKH platform.

3.4 Design a unified data framework in CLUE-H

The unified data framework in CLUE-H will be developed concerning metadata of the individual projects. For this aim, a harmonised form will be used and the collected results will be made accessible in a searchable way on the CLUE-H website in a timely manner. WG2 group discussed and established the creation and organization of a common metadata template usable for the four projects. This task will be performed under the Common European Research Information Format (CERIF) standard. Different types of data have to be taken into account for the creation of the

metadata table as .xlsx (excel spreadsheet), .ppt (power point presentation), .mat or .m (Matlab related file extension), .txt (text file), .s2p (touchstone), .avi (video files) and others. Metadata will have specific fields to be filled depending on the specific type of file. The different fields required for each type of file are reported in the next Table. Once the metafile has been created the data set identifier has to be generated using the following organization:

Project"_"Date"_"Time"_"Name"_"Type"_"Extension"_"Place"_"Creators"_"Target user"_"Other".

"Project" is the name of the specific project at which the metadata belongs so for example SEAWave. Date is the date in format "YYMMDD", "Time" is the time of metadata creation in the format "HHMMSS". "Name" is a short name of the data eventually with the specification of WP in which these data were generated. "Type" describes the type of data as publication, measured data, simulation data, protocol description or others. Finally, extension describes the data file extension.

Table 1. Metadata fields for the different data types

Data type/ Metadata type	Extension	Directory	Creator name	Editor name	File name	File size	Bits per sample	Height	Width	Resolution	Variables	Description
Image file	x	x	x	x	x	x	x	x	x	x		
Video file	x	x	x	x	x	x	x			x		
mat (Matlab file)	x	x	x	x	x	x					x	
m script (Matlab file)	x	x	x	x	x	x						x
Excel file	x	x	x	x	x	x						x
Powerpoint file	x	x	x	x	x	x						x
Text file	x	x	x	x	x	x					x	x
Touchstone file	x	x	x	x	x	x						x
Other												

4 Organisation of project outputs

4.1 Summary of CLUE-H projects outputs

4.1.1 NextGEM outputs

NextGEM will generate reviews, technical (exposure, measurements and simulations/modelling) and biological data from *in vitro*, *in vivo*, *ex vivo* and human studies Other documents will also be generated, e.g. standard operating procedures (SOP) related to experiments and exposure measurements, study protocols, reports, deliverables, etc. In order to promote the NextGEM outputs to the relevant stakeholders and engage with audiences the project has developed its own unique visual identity and a website that is continuously updated and contains all information relevant to the project. In addition to the external communication, the project has developed internal communication and data management tools to enable the consortium to provide updates to the data management plan and its associated activities for the team responsible with upkeeping the data management plan of the project to do so more effectively.

4.1.2 GOLIAT outputs

Research generated outputs in GOLIAT include: Protocols, Codebooks, General descriptive tables of aggregated data, Scripts/codes of analyses, Add-on sensor fixed on mobile devices for conversion of monitored mobile phone power values into field levels, Next-generation multiband RF-EMF nodes for temporal monitoring in 5G telecom networks, Tri-axial 5G RF-EMF prototypes FR1sub 6 GHz and FR2 mm-waves 26 GHz band, Expert RF-EMF dose model, Public RF-EMF dose mode, Exposure setups working at 0.7, 3.5, and 26 GHz, Cell and animal samples, Matrices of expert and citizen models on misconceptions and gaps of understanding and Scientific manuscripts and public reports.

4.1.3 SEAWave outputs

Within SEAWave, the purpose of data collection/generation is to gather evidence of the developed quantification of the 5G exposure, the creation of new tools and instruments for this purpose, and the generation of new scientific data for assessing, in a solid fashion, the possible health risk for 5G-FR2 exposures. Additionally, data from different environments (ex. urban areas and smart industries) will be collected in order to permit the assessment of exposure in real 5G scenarios to establish appropriate standards and validate the results coming out from the biological investigations in vitro, in vivo and on human volunteers.

Therefore, the collected data will include the experimental and theoretical procedures for characterization of the exposure along with the protocols for biological testing. Finally, simulation data and codes, for example for the automatic control of measurement instruments, or the dosimetric assessment will be also collected.

4.1.4 ETAIN outputs

ETAIN produced outputs will include: an RF-EMF exposure app and an exposure portal, citizen engagement for app design, uptake, citizen co-designed communication materials, citizen events, networking, exposure calibration functions for the app, as well as exposure maps, future 5 and 6G exposure predictions, exposure maps, calibrated population dose distributions, algorithm for dose calculations, protocols, codebooks and measurements, channel sounding measurement campaigns as well as front-end exposure reduction techniques, exposure-aware network planner tool, 3D-insect models, numerical quantification of insect exposure, quantification of effects in insects including fruit flies, honey and solitary bees and insect biodiversity, planetary health impact assessment framework, dissemination & communication Plan, visual identity, website, social media, press releases, publications, project management structure and processes, contracts, reports, exploitation plan, DMP, management of ethics, project coordination and communication.

4.2 Organization of CLUE-H outputs

CLUE-H web portal has been developed to serve as the public presence of the cluster. The web portal utilises technology features and is a part of the communication and dissemination activities undertaken for this cluster. It can be accessed at https://www.emf-health-cluster.eu/ and www.clue-h.eu. The web portal page includes the on-going CLUE-H generated outputs and results, such as Public Deliverables, Publications, Press Releases, Leaflets & Brochures, Photo & Video Gallery. Also, important information for the industry as well as the general public will be published on this page, as it becomes available. Such information includes: Policy Briefs, Standards and Guidelines and they will available through the website of the cluster.

5 Conclusion

This deliverable provides the data management process between the CLUE-H projects. That includes the summary of the data management plans between projects, aiming to share data management plans between the cluster partners. In addition, the aim for the establishment of a unified framework is based on the collection of the individual data types and structure, data storage and data exchange from the different projects. Finally, the different types of generated outputs are provided to assist the organizations of the variety of the produced data within CLUE-H.