CONVERGE

view-to-communicate and communicate-to-view

CONVERGE - Telecommunications and Computer Vision Convergence Tools for Research Infrastructures

# D5.1: Initial Data Management Plan

| | | | |
|---|---|---|---|
| **Work package** | WP5 | **Task** | Task 5.3 |
| **Deliverable number** | D5.1 | **Due date** | 31/07/2023 |
| **Editor** | Liisa Marjamaa-Mankinen (CSC) and Satu Tissari (CSC) | | |
| **Internal reviewers** | Sérgio Silva (Adapttech), Maria-Cristina Marinescu, Nadia Tonello and Cristian Xum (BSC CNS) | | |
| **Version** | 1.0 | | |
| **List of contributors** | INESC TEC<br>UOULU<br>BSC CNS<br>EURECOM<br>SORBONNE<br>INRIA<br>CSC<br>ALLBESMART LDA<br>Adapttech<br>Finwe Oy<br>FinCloud Ltd.<br>Queen’s University Belfast<br>INTERDIGITAL | | |
| **Dissemination Level** | PU<br>***PU:*** *Public, fully open, e.g., web (Deliverables flagged as public will be automatically published in CORDIS project’s page)*<br>***SEN:*** *Sensitive, limited under the conditions of the Grant Agreement* | | |

Co-funded by the European Union

The CONVERGE project has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101094831, including top-up funding by UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee.

Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union, SNS JU or UKRI. The European Union, SNS JU or UKRI cannot be held responsible for them.

# CHANGE REGISTER

| Version | Date | Editor | Organization | Changes |
|---|---|---|---|---|
| A | 30-05-2023 | Liisa Marjamaa-Mankinen, Satu Tissari | CSC | Initial draft for coordinator review |
| B | 06-06-2023 | Liisa Marjamaa-Mankinen, Satu Tissari | CSC | Moved the content from Horizon Europe template into the deliverable template of the project. Added some detailed information from the first tables from the longer version. |
| C | 28-06-2023 | Liisa Marjamaa-Mankinen, Satu Tissari | CSC | Modified based on comments from the coordination organization, INESC TEC (Luís Pessoa and Ana Sequeira) as well as reviewers from ADAPTTECH (Sérgio Silva) and BSC (Maria-Cristina Marinescu, Nadia Tonello and Cristian Xum). |
| D | 21-07-2023 | Ana Sequeira, Luís Pessoa | INESC TEC | Formatting and language improvements. |

# EXECUTIVE SUMMARY

The main objective of the CONVERGE project (Telecommunications and Computer Vision Convergence Tools for Research Infrastructures) is the development of an innovative toolset aligned with the motto “view-to-communicate and communicate-to-view”. This toolset is a world-first and consists of vision-aided large intelligent surfaces, vision-aided fixed and mobile base stations, a vision-radio simulator and 3D environment modeler, and machine learning algorithms for multimodal data including radio signals, video streams, Radio Frequency (RF) sensing, and network traffic traces. This toolset will be deployed into seven research infrastructures (RI) mostly aligned with the European Strategy Forum on Research Infrastructures (ESFRI) Scientific Large-scale Infrastructure for Computing/Communication Experimental Studies (SLICES-RI) and improve their competitiveness. It is emphasized that the focus of the project is on the development of technical tools.

In this project, the creation of an initial Data Management Plan (DMP) is agreed to occur between M1-M6. The DMP is planned to have an initial and a final version considering that the data-related questions often are updated as the project evolves. Thus, the CONVERGE DMP will be developed and updated throughout the project. The final version will be released by M33.

The DMP offers a description of the decisions, taken by each partner and the consortium as a whole, regarding the management of data. In particular, it will comprise the aspects related for example to the data creation, data handling, data storage, data categories, formats and sizes, as well as the associated metadata. The data management process is performed to ensure that all the data is findable, accessible, and interoperable for reuse as well as published for the open reuse by other researchers after the project.

This deliverable presents the initial step in the creation of a plan to be used for managing data and other research related outputs within the project. This first version of the deliverable covers the timeframe M1-M4 of the project. At this phase of the project, this initial document includes conclusions based on the information gathered from the partners including the current practices. Taking in account that the project must follow the findable, accessible, interoperable, and re-usable (FAIR) principles, the questions analysed throughout the document were based on the Horizon Europe DMP template combined with some additional questions used for gathering information from the partners. To better align this document with the SLICES-RI, references to the classifications made by SLICES-SC (SC stands for Starting Community) were added to the current version.

The need for the training of partner’s participants has been identified regarding FAIR principles and alignment with SLICES-RI. This need for training and related required resources will be clarified in the next phase of the project. Thus, we will be able to adopt more detailed views in different issues possibly through the involvement of the partners in workshops. One aspect to be taken into account in the next phase of the project, to help defining the full research life cycle of CONVERGE, is the solution of SLICES-SC for the support of the full research life cycle presented in the document of SLICES-SC D2.4 Definition of common standards and practices for research reproducibility [FDI22].

At the time writing this deliverable, the work on WP 1 is ongoing, targeting the collation of the relevant toolset technologies from a continuously evolving state of the art, definition of use cases including target groups, and identification of requirements.

# TABLE OF CONTENTS

# LIST OF TABLES

# ABBREVIATIONS

3D – 3 Dimensional

DMP – Data Management Plan

EM – Electromagnetic

EOSC – European Open Science Cloud

ESFRI – European Strategy Forum on Research Infrastructures

FAIR – Findable, accessible, interoperable, and re-usable

LiDAR – Light Detection and Ranging

LIS – Large Intelligent Surface

MIMO – Multiple Input Multiple Output

ML – Machine Learning

RF – Radio Frequency

RI – Research infrastructure

RIS – Reconfigurable Intelligent Surface

SLICES-RI – Scientific Large-scale Infrastructure for Computing/Communication Experimental Studies

SLICES-SC – Scientific Large-scale Infrastructure for Computing/Communication Experimental Studies – Starting Community

UE – User equipment

# 1. INTRODUCTION

Telecommunications and computer vision have evolved as separate scientific areas. This is envisioned to change with the advent of wireless communications with radios characterised by line-of-sight ranges which could benefit from visual data to predict the wireless channel dynamics. Computer vision applications will also become more robust if helped by radio-based imaging. This new joint research field relies on wireless communications, computer vision, sensing and machine learning, and it has a high innovation potential because of the large domain of innovative applications it enables and the relevant know-how available in Europe. However, the full potential of this new area can only be evaluated if adequate RIs and tools are available [POEU23].

The main objective of the CONVERGE project is the development of an innovative toolset aligned with the motto "view-to-communicate and communicate-to-view". This toolset is a world-first and consists of vision-aided large intelligent surfaces, vision-aided fixed and mobile base stations, a vision-radio simulator and 3D environment modeler, and machine learning algorithms for multimodal data including radio signals, video streams, RF sensing, and network traffic traces. This toolset will be deployed into 7 RIs mostly aligned with the ESFRI SLICES-RI and improve their competitiveness.

CONVERGE will also provide the scientific community with open datasets of experimental and simulated data obtained with the toolset in the RIs, meet scientific and industrial requirements by addressing relevant 6G verticals, enhance the competitiveness of the involved companies, extend the European influence to world-wide recognised RIs, enable the creation of new RIs, contribute to the development of new environment-friendly tools, and help European Union to address its societal challenges.

Throughout the project, an innovative toolset will be developed, including the tools listed in Table 1.

Table 1. Overview of the planned toolset to be developed within the project.

| Tool | Description |
|---|---|
| **Tool 1: Vision-aided large intelligent surface** | Vision-aided large intelligent surface (LIS) aimed at allowing experimentation, in a controlled room environment, of massive Multiple Input Multiple Output (MIMO) wireless communications, high precision 3D positioning and environment sensing, including human sensing and microwave holography, through the combination of a smart programmable meta-surface antenna with a camera array. Two LIS versions will be developed, addressing both sub-6 GHz and mm-Wave frequency bands, enabling research based on multi-frequency data. |
| **Tool 2: Vision-aided base station** | Vision-aided base station aimed at enabling communications and experimentation with mobile terminals mostly related to beamforming, multi-user access, and opportunistic scheduling by taking advantage of environment mapping made by video-cameras and of the LIS. Both fixed (but relocatable) and mobile versions of the base station will be developed, the latter adding the possibility of controlled mobility (different positions along the time or predefined trajectories), while also taking advantage of video cameras and LiDAR and enabling its cooperation with the fixed base station and the LIS. Vision-aided mobile user equipment (terminals) will complement the tool, aimed at exchanging network traffic with the base stations and obtaining multiple video perspectives of the environment. |
| **Tool 3: Vision-radio simulator and 3D environment modeller** | Vision-radio simulator and 3D environment modeller aimed at creating a digital 3D representation of the environment to enable simulating observations at any location and producing geometric models suitable for radio signal propagation simulation based on information from |

| | |
|---|---|
| | material dielectric properties and antenna geometries. This tool will consist of a ray-tracing software employing data from visible light and radio signal interactions with the environment. The simulator will help to understand the level of accuracy needed from geometric and electromagnetic models to achieve and surpass the performance of statistical channel models in radio communications. |
| **Tool 4: ML algorithms** | Machine learning algorithms aimed at facilitating the processing of a variety of data made available by the above tools including videos from cameras and RF sensing signals from the large intelligent surface, as well as timestamped and space-referenced communications traces including received powers, signal-to-noise ratios, antenna radiation patterns, objects positioning, and network traffic performance indicators such as bitrates or delays. |

# 2. DATA SUMMARY

The discussions surrounding the DMP at this initial stage of the CONVERGE project, were mainly meant to raise awareness for the importance of planning ahead issues regarding the data generation, reuse, and storage. During the information collection phase, it was possible for the project partners to identify and reflect upon their initial plans regarding the purpose of data generation and data reuse, details about data itself and their needs for storing data during and after the project.

Currently, the partners have identified their roles and activities concerning data generation and reuse aligning them with the objectives of the project, na,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,mely with the planned tools (refer to Table 2).

Table 2. Responses to the question: What is the purpose of the data generation or re-use and its relation to the objectives of the project?

| Tool | Purpose |
|---|---|
| **Tool 1: Vision-aided large intelligent surface** | Full-wave simulation studies for the design of LIS type apertures will be implemented to study RF mapping and wireless coverage enhancement. Measurements are also planned to verify the practicality of the studied scenarios. |
| | Performance measurements and numerical studies will be carried out to generate new data and study LIS architectures to demonstrate the RF performance of the LIS in far-field conditions. |
| | The assessment of the practicality of using LIS/ Reconfigurable Intelligent Surface (RIS) for gathering sensory information, in particular to enhance the system's knowledge of the 3D environment where the equipment is located. |
| | Antenna unit cell simulation and experimental results (magnitude and phase of reflection coefficient) will be collected in the scope of the LIS development. |
| | Existing 3D virtual models of LIS structures will be re-used to speed up the development of new 3D environment model designs. |
| | Image/Video data will be generated by Multi Camera Networks System in order to allow for the deployment of computer vision algorithms. |
| **Tool 2: Vision-aided base station** | Data generated by LiDARs (3D cloud points) will be used to detect Line-of-Sight radio link blockage to trigger the 5G handover procedure and decrease the handover latency. |
| | Radio signals, network traffic traces and video streams will be collected to assess the practicality of using vision-based tools to enhance the efficiency of the radio network (e.g., influence beamforming and power-control) |
| | RF sensing data will be collected to enhance current vision-based machine learning (ML) pose estimation detection algorithms |
| | Existing 3D virtual models of radio components and user equipment (UE) will be re-used to speed up the development of new 3D environment model designs. |
| **Tool 3: Vision-radio simulator and 3D environment modeller** | Data generated by the developed physical model of the controlled environment will be used to carry out several learning tasks. |
| | Data generated by cameras and/or LiDARs will be used to create a model of the environment suitable to fit the simulator. Simulations of raytracing of visual light and radio will be generated at the simulator itself. |
| | Based on the 3D environment model generated by cameras, the simulator will produce electromagnetic (EM) coverage and therefore data communication performance information, suitable to develop indoor/outdoor propagation models which can be used offline or in real-time simulation/emulation. |

| | |
|---|---|
| **Tool 4: ML algorithms** | Datasets collected from real environments using Tools 1 and 2 (cameras, LiDARs, radios) or generated at the vision-radio simulator (Tool 3) will be used to train sensing algorithms based on ML approaches, enhancing the accuracy of training ML models, and exploring novel semantic sensing approaches. New data will be generated that can be outputted in several forms: arrays of features extracted from raw data, results from decision support algorithms, benchmarking metrics, generated data resulting from data augmentation processes, etc. |
| | Data collected by radio equipment (i.e., RIS/LIS and radio-unit antenna arrays) will be used to assess the practicality of using ML approaches for enhancing sensory information and to act on the radio transmission chain to improve the efficiency of the radio network. |
| | Data will be captured to support the development of techniques for human pose estimation or user identification based on radio signals where the optical camera data will be used as ground truth. |

The CONVERGE partners have provided justifications for generating new data and will also make use of relevant existing public datasets as when appropriate (refer to Table 3).

Table 3. Utilisation of existing data (to minimize unnecessary data generation).

| Tool | Re-use of existing data | Reasons why existing data will not be used |
|---|---|---|
| **Tool 1: Vision-aided large intelligent surface** | Existing Public Datasets if deemed appropriate for benchmarking and comparison | |
| | | We do not have existing data suited for this tool. Full-wave simulations, performance measurements and numerical studies will be carried out to generate new data and study LIS architectures. |
| **Tool 2: Vision-aided base station** | Existing Public Datasets if deemed appropriate for benchmarking and comparison | |
| | | Existing 3D cloud points data available in public repositories do not fit the scope of this tool (see Deliverable 1.1). |
| **Tool 3: Vision-radio simulator and 3D environment modeller** | Existing Public Datasets if deemed appropriate for benchmarking and comparison | |
| | | There are existing datasets of point clouds and triangle meshes of environments relevant for CONVERGE, but there is the need to collect new datasets specifically aligned with the targeted project use cases. A new physical model will be developed to study a controlled environment. |
| **Tool 4: ML algorithms** | Existing Public Datasets if deemed appropriate for benchmarking and comparison | |
| | | Some existing vision-radio datasets are available but do not |

| | | |
|---|---|---|
| | | support the project use cases appropriately. They might be used for pre-training or self-supervised learning, but there is the need to collect new data using Tools 1-3 in order to collect datasets to be used specifically for ML training (e.g. ground truth vision data) and ML testing (e.g. RF data) |

The CONVERGE partners have given a preliminary description of data types, data formats and data volumes. To ensure the alignment with SLICES-RI, the categories of SLICES-SC for data types and formats were utilized [ZIE21] (see the “Data category” column in Table 4).

Table 4. Responses to the question: What types and formats of data will the project generate or re-use? What is the expected size of the data you intend to generate or re-use?

| Tool | Data category | Data formats | Expected data size |
|---|---|---|---|
| **Tool 1: Vision-aided large intelligent surface** | • Observational<br>• Experimental<br>• Simulation<br>• Derived | • Open file formats<br>• Unity/other selected 360 or other video and image formats<br>• Radar formats<br>• Configuration script formats<br>• Page layout formats | • Up to 100’s of GBs per partner |
| **Tool 2: Vision-aided base station** | • Observational<br>• Experimental<br>• Simulation<br>• Derived | • Open file formats<br>• Unity/other selected 360 or other video and image formats<br>• LiDAR formats<br>• Complex Q15 format<br>• Configuration script formats<br>• Page layout formats | • Up to 100s of GBs per partner<br>• One partner to TBs |
| **Tool 3: Vision-radio simulator and 3D environment modeller** | • Observational<br>• Experimental<br>• Simulation<br>• Derived | • Open file formats<br>• Unity or other selected 360 video, image and 3D storage formats<br>• Radar formats<br>• Packet capture formats<br>• Configuration script formats<br>• Page layout formats | • Up to 100s of GBs per partner |
| **Tool 4: ML algorithms** | • Experimental<br>• Simulation<br>• Derived | • Open file formats<br>• Real or synthetic video/images formats and radio signal formats<br>• Enhanced sensory information in video/image formats | • Several GBs per partner |

Currently, the CONVERGE partners do not anticipate the necessity of engaging with external infrastructures and systems beyond the scope of the project. As the work on WP1 is still ongoing, the partners have yet to determine the specific target groups for which the data will be valuable.

In terms of data storage, the CONVERGE partners have primarily recognized the requirement for shared storage among partners throughout the project. Additionally, there is a need for permanent storage to support future utilization of the developed toolset beyond the project’s duration. The partners have expressed the necessity for utilizing SLICES-RI resources as well as special computing resources using NVIDIA GPU and DGX100. Some partners have also identified a potential future need for accessing third-party computing resources.

# 3. FINDABLE, ACCESSIBLE, INTEROPERABLE, AND RE-USABLE (FAIR) DATA

The discussions surrounding the DMP at this initial stage of the CONVERGE project stimulated the individual reflexion and global discussion of issues related to the concepts of findable, accessible, interoperable, and re-usable (FAIR) data. During the next phases of the project the detailed conditions on Open Science described in the Grant Agreement will be taken into account.

Through this process, the need for the training of partner participants in FAIR (Findable, Accessible, Interoperable, and Reusable) principles and aligning with the SLICES-RI has been identified. Further discussions in the next phase of the project will address this training need and the required resources. As CONVERGE progresses to a more mature stage, we can take more sustained actions including assessing deployment of the solutions proposed by SLICES-SC for supporting the full research life cycle, as presented in the SLICES-SC document D2.4 Definition of common standards and practices for research reproducibility [ZIE21].

## 3.1. Making data findable, including provisions for metadata

In the context of data collection, the objective of making data, and the respective metadata, findable must be assured with the assignment of a globally unique and persistent data identifier. To accomplish the goal, there is a need to agree on the common metadata standards to be used for complementing collected/generated data with rich metadata. Both data and metadata will be registered or indexed in a searchable resource.

Currently the CONVERGE partners have various processes for the creation of metadata. On the next phase of the project, agreed metadata standards will be implemented by all partners involved in data collection, generation, sharing, and publishing.

Some of the partners have employed the methodology of assigning a DOI (persistent identifier) for their publications. In the upcoming phase of the project, we will collectively decide on a common practice for publishing data and utilizing persistent identifiers. If necessary, the partners can also agree to use alternative identifiers during the project to ensure the reliability of all data processing.

## 3.2. Making data accessible

A common repository for the administrative data of CONVERGE project has been created using the Nextcloud platform, which is planned to be further developed throughout the project. Within the tasks of WP1, including the preparation of D1.1, the partners have been contributing and collaborating for the creation of this common repository.

As agreed in the Grant Agreement the principle “as open as possible as closed as necessary” will be followed. The Grant Agreement guides one to take in the account the possible restrictions to open the data. For making data accessible data should be accompanied by discovery metadata, such as access level information on who can access the data and under what conditions it can be reused, as well as information about licence to explain how it can be legally reused. In the Grant Agreement it is recommended that specific Common Creative licences need to be used regarding both data and metadata.

The partners have not yet acknowledged varying levels of access to their data. When deciding on repositories, metadata standards, etc., it will be considered what level of access and license information is required.

## 3.3. Making data interoperable

Data must be made interoperable in order to enable data sharing and reuse within and between disciplines. This calls for proper handling and preparation so that it may be transferred between services and systems as well as combined with other data without losing or disorganizing its informational integrity.

The partners have not yet mentioned techniques, semantic artefacts (for example metadata schemas, vocabularies, ontologies), or interoperability best practices that have the support of the community.

Some partners have expertise in standardisation which will take place on WP4.

## 3.4. Increase data re-use

To promote increased data re-use, it is crucial to document the data in a manner that enables others to find, access, and utilize it effectively. Documentation involves creating descriptive information that elucidates the context, methods used for data capture and processing, data structure, chosen filing system, and other relevant details.

Certain partners have implemented practices for data documentation to facilitate data re-use, such as utilizing ReadMe files and GitHub. Furthermore, some partners have expressed their willingness to make certain data publicly available under Creative Commons licenses.

Quality assurance is an integral part of the project and will be addressed in subsequent stages.

# 4. OTHER RESEARCH OUTPUTS

In addition to data management, it is important for partners to also consider and plan for the management of other research outputs that may be generated or re-used during the course of the project. While these outputs are not yet fully defined at this initial stage of the project, it is conceivable that they could encompass software, workflows, protocols, models, and other related items that may emerge. At present, partners have not identified any other research outputs.

# 5. ALLOCATION OF RESOURCES

The allocation of resources will be defined in a later stage of the project.

# 6. DATA SECURITY

The selection of data repositories for the project will be made from a pool of trusted data repositories. This ensures that data management is conducted in a reliable and secure manner.

Furthermore, WP6, which focuses on ethics, will play a crucial role in managing potential sensitive personal data within the project. This work package will provide guidance and protocols to ensure the proper handling and protection of sensitive information.

# 7. ETHICS

It has been acknowledged that certain personal and sensitive data may be included among the various types of data collected throughout the project. It is important to differentiate whether the sensitive personal data is collected solely for internal use or for other purposes. The data collected during the development of the tools to achieve the project's main objectives are intended for internal use within the project and not for publication. Any potential changes to this plan, such as using the collected data in any project result or output, would present complex legal considerations and require a case-by case analysis.

The use cases to be studied and implemented in CONVERGE are currently under discussion among all the partners and will be described in greater detail within the scope of WP1.

Ethical and legal issues are deliberated among the partners in WP6, particularly within the project's ethical board. The work conducted in WP6 aims to help the management of the data, from its creation to its storage for future use, in compliance with all legal requirements This includes principles such as the necessity to involve human participants in experiments and the need to obtain their consent~~.~~

The CONVERGE project incorporates Artificial intelligence (AI) methodologies, and their implications will be evaluated and assessed within WP6.

# 8. OTHER ISSUES

From the discussions among the partners regarding the DMP, at this moment there are no other procedures envisaged for the data management.

# 9. CONCLUSION

The work to be developed by the partners of the CONVERGE project has been launched. This marks a journey towards developing the world-first innovative toolset from a continuously evolving state of the art.

This Data Management Plan presents the initial stage (M1-M4) in the creation of a plan to manage data and other related research outputs within the project. This document includes insights and initial targets based on the information gathered from the partners, including their current practices.

At the time writing this deliverable, the work on WP1 is progressing, focusing on the collation of the relevant toolset technologies from a continuously evolving state of the art, the definition of use cases including target groups, and the identification of specific requirements.

This DMP will be updated as the project advances to the next phases, taking in account the entire research data management life-cycle, as well as FAIR principles and the guidance of EOSC and SLICES-SC in providing re-usable datasets and other research outputs for the scientific community. This will support the deployment of the toolset into the seven targeted RIs to improve their competitiveness in the future.

# REFERENCES

[FDI22] S. Fdida, H. Rahich, S. Gallenmüller and N. Makris, “SLICES SC D2.4 - Definition of common standards and practices for research reproducibility”, September 2022, [Online]. Available: https://slices-sc.eu/wp-content/uploads/2023/01/SLICES-SC-D2.4.pdf [Accessed June 2023]

[POEU23] “The Community Research and Development Information Service (CORDIS) project ID 101094831,” Publications Office of the European Union, [Online]. Available: https://cordis.europa.eu/project/id/101094831. [Accessed February 2023].

[ZIE21] S. Ziegler, C. Crettaz, A. Q. Rodriguez, E. Mespoulhes and N. Makris, “SLICES-SC D3.1,” August 2021, [Online]. Available: https://slices-sc.eu/wp-content/uploads/2023/01/SLICES-SC-D3.1.pdf. [Accessed February 2023].