Ref. Ares(2024)6896643 - 30/09/2024

Deliverable Number 5.2

RP1 update of the Data Management Plan

Related WPs: WP5

Related tasks: T5.1

	Affiliation / Position	**Name**	
Author	UNIGE / Deputy CIO	Pierre-Yves Burgi	
Verifier	Prof. Dr.	Renaldas Raišutis	
Issued:	*2024.09.30*	*Update:*	

KEYWORDS:

Data Management, Data Management Plan

TABLE OF CONTENTS

1. INTRODUCTION 4
2. PLAN OVERVIEW 4
2.1 Data Collection 5
2.1.1. What data will you collect or create? 5
2.2 How will the data be collected or created? 7
2.2.1 Source code of control software: 7
2.2.2 Input data produced for experimentation and tests of control software: 7
2.2.3 Output data produced for experimentation and tests of control software: 7
2.3 Documentation and Metadata 7
2.3.1 What documentation and metadata will accompany the data? 7
2.4 Ethics and Legal Compliance 8
2.4.1 How will you manage any ethical issues? 8
2.4.2 How will you manage copyright and Intellectual Property Rights (IPR) issues? 8
2.5 Storage and Backup 8
2.5.1 How will the data be stored and backed up during the research? 8
2.5.2 How will you manage access and security? 8
2.6 Selection and Preservation 9
2.6.1 Which data are of long-term value and should be retained, shared, and/or preserved? 9
2.6.2 What is the long-term preservation plan for the dataset? 9
2.7 Data Sharing 9
2.7.1 How will you share the data? 9
2.7.2 Are any restrictions on data sharing required? 9
2.8 Responsibilities and Resources 9
2.8.1 Who will be responsible for data management? 9
2.8.2 What resources will you require to deliver your plan? 9
3. PLANNED RESEARCH OUTPUTS 10
3.1 Physical object - "New archiving service based on DNA" 10

1. INTRODUCTION

The purpose of this deliverable report is to present the updated Data Management Plan (DMP) for the project, incorporating all changes and developments up to Month 12 (M12). This updated DMP includes the addition of new datasets and reflects the evolving data management practices implemented during the first year of the project.

The construction of the DMP was carried out using the online tool DMPonline, which is based on the open-source DMPRoadmap codebase jointly developed by the Digital Curation Centre (DCC) and the University of California Curation Center (UC3). The DCC and UC3 collaborate closely with research funders and universities to produce a tool that generates active DMPs, catering to the entire lifecycle of a project—from bid preparation through to completion.

[https://dmponline.dcc.ac.uk/about_us]

2. PLAN OVERVIEW

A Data Management Plan created using DMPonline
Title: DNAMIC
Creator: Pierre-Yves Burgi
Principal Investigator: Pierre-Yves Burgi
Data Manager: Pierre-Yves Burgi
Project Administrator: Ignas Galminas
Contributor: Renaldas Raisutis, Jerome Charmet
Affiliation: Other
Funder: European Commission
Template: DCC Template
ORCID iD: 0000-0002-4956-9279

Project abstract:

We live in a society that produces and requires more and more data every year. It has become apparent that our current data storage strategies are not adequate to meet this increasing demand. The quest for robust, high density, sustainable, and economically viable data storage solutions has highlighted the enormous potential and suitability of DNA data storage. However, to date, none of the solutions have considered a comprehensive vision of anchoring this technology in standardised archival frameworks that will support long-term storage. Besides, most DNA data storage applications are limited to specialist laboratories due to complex (and expensive) protocols. Our project entitled DNAMIC (DNA Microfactory for Autonomous Archiving) proposes an autonomous solution based around a low energy consumption microfactory that will be developed for end-to-end DNA data archiving (from encoding to decoding via synthesis, storage, quality control and sequencing among others). The microfactory is interoperable and future proof thanks to

technology block that can be easily modified or replaced. Our solution will be compliant with the Open Archival Information System (OAIS) reference model (ISO 14721). To enable the implementation of disaster recovery strategies, critical for long term storage, we propose a novel DNA encoding scheme.

ID: 146083
Start date: 02-10-2023
End date: 30-09-2026

Last modified: 19-09-2024

Grant number / URL: 101115389

2.1 Data Collection

2.1.1. What data will you collect or create?

This project will work with and generate several main types of data :

1. Source code for the control software
    - Origin : produced for the project, could include source code produced in previous or other projects
    - Format : programming language files
    - Volume estimated : 1 MB – 50 MB
2. Flow sequences used for the experimentation and tests
    - Origin : produced for the project
    - Format : JSON files
    - Volume estimated : <1 MB
3. Input data produced for experimentation and tests : status of machines, status of robots, position of robots, velocity of robots, machine operations duration.
    - Origin : produced for the project
    - Format : During experimentation saved in PostgreSQL database. Exported as JSON files after experimentation.
    - Volume estimated : 10 MB – 10 GB
4. Output data produced for experimentation and tests, as time series : decisions of software, events for all machines, events for all robots
    - Origin : produced for the project

- Format : During experimentation saved in InfluxDB database. Exported as JSON files after experimentation.
- Volume estimated : 100 MB – 1 GB

5. In use cases, we will simulate DICOM datasets to test the archiving system.
   - Origin : produced for the project
   - Format : DICOM.
   - Real-world patrimonial data (e.g. some industry digital register, or technological archive data, specific data format related to archive). Any privacy concerns will be addressed by de-identifying sensitive information.
   - Simulated DICOM images with appropriate variations of shape and pattern.
   - Digital health images (DICOM) , e.g. veterinary images or anonymized health images with removed sensitive or patient information.
   - Non-destructive testing (NDT) (DICOM) images from Industry 4.0 sectors (energy, transport, aviation, materials) with removed sensitive or confidential information.
   - Volume estimated : 10 MB – 50 MB
6. Sequencing data
   - Origin : produced for the project, during sequencing experiments
   - Format : FastQ, Fast5, BAM, SAM, FASTA.
   - Volume estimated : 10 GB – 100 TB (depending on the input size)

7. Report data used for measuring efficiency of sequencing: plots, charts, histograms, tables describing output results.
   - Origin: produced for the project, during sequencing experiments
   - Format: HTML, CSS, JavaScript
   - Volume estimated: 10-200MB

8. Dependencies required during creation of report: virtual environments, intermediate results, AI models.
   - Origin: produced for the project, before and during experiments
   - Format: Programming languages files
   - Volume estimated: 10GB – 1TB

Other data we might potentially collect are technical data for qualification of the processes.

2.2 How will the data be collected or created?

2.2.1 Source code of control software:

- Apply HE-Arc Engineering coding and naming conventions.
- Use Git as versioning system with Gitflow workflow.
- Save on GitLab provided by HE-Arc Engineering : https://labinfo.ing.he-arc.ch/gitlab

2.2.2 Input data produced for experimentation and tests of control software:

- One folder for each experimentation with naming YYMMDD-Experimentation-Number with a readme file describing the content.
- Saved during experimentation on GitLab provided by HE-Arc Engineering : https://labinfo.ing.he-arc.ch/gitlab
- Saved after experimentation in project folder provided by HE-Arc Engineering (network server shared folder).

2.2.3 Output data produced for experimentation and tests of control software:

- Data saved during experimentation in InfluxDB database server provided by HE-Arc Engineering.
- Apply Micro Lean Lab data acquisition procedure : specify the list of tasks to be done by collaborators to ensure data are correctly and completely saved in database.
- Exported after experimentation in JSON files and saved in project folder provided by HE-Arc Engineering (network server shared folder). One folder for each experimentation with naming YYMMDD-Experimentation-Number with a readme file describing the content.

2.3 Documentation and Metadata

2.3.1 What documentation and metadata will accompany the data?

The project deals with the archiving of data following the OAIS standard. Consequently, metadata will be an important part of the data description following standard format such as datacite, METS, PREMIS, etc.

Source code of control software is provided with the following metadata :

- Comments according to coding convention using Doxygen for C-like programming languages.
- Comments and annotation according to ReStructuredText for Python programming language files.
- A readme file for each source code project, with the standard information such as name, description, installation, usage, dependencies, authors and license

Input and output data produced for experimentation and tests are provided with the following metadata :

- Spreadsheet describing the configuration of the experimentation, with list of devices (robots…), settings of devices, initial positions and status, pictures taken during experimentation and human oriented comments (free text)
- Readme file describing the JSON file format : structure, type of value, description of values, unit of values

2.4 Ethics and Legal Compliance

2.4.1 How will you manage any ethical issues?

Generated data will be free of any ethical issues.

DNA synthesis we will provided by TWIST Bioscience, which applies strict rules that comply with bio-engineering, see https://www.twistbioscience.com/legal/ethical-business-practices

2.4.2 How will you manage copyright and Intellectual Property Rights (IPR) issues?

IPR is managed by a patent expert at HES-SO ARC, who coordinate the IPR issues among all partners.

Source code of the control software will be published with an open source license.

Input and output data produced for experimentation and tests of control software will be released as open data under Creative Commons CC0 license.

2.5 Storage and Backup

2.5.1 How will the data be stored and backed up during the research?

The University of Geneva has developed a long-term preservation system for archiving research data: OLOS.swiss. Consequently, all pertinent documentation and research data generated during the project will be archived on OLOS.swiss. The control software data is stored on institutional storage facilities (network file server, GitLab server, InfluxDB server). All these servers are configured by the IT department to be automatically backed up on a daily basis.

2.5.2 How will you manage access and security?

OLOS.swiss offers all functionalities to guarantee access levels, embargoes, and security, if needed.

The control software data is stored on institutional storage facilities (network file server, GitLab server, InfluxDB server). Access and security is managed by the IT department.

2.6 Selection and Preservation

2.6.1 Which data are of long-term value and should be retained, shared, and/or preserved?

All documentation needed to realize the targeted system will be retain on the long term on OLOS.swiss

Generated data to test the system will not be retained on the long term, but only backup for the duration of the project.

2.6.2 What is the long-term preservation plan for the dataset?

We expect to archive for a period of 10 years. Given we will keep only documentation, we should not need more than 1 TB. For up to 50 GB it costs about 50 euros per year (https://olos.swiss/pricing). Price will be evaluated later in the project when we will have a better idea about the volume needed.

2.7 Data Sharing

2.7.1 How will you share the data?

OLOS.swiss has all the features for data sharing as it is a system which is compliant with the FAIR principles.

Source code of control software with open source license will be published on github.com.

2.7.2 Are any restrictions on data sharing required?

Some documents that will remain private and shared only within the project partners until appropriate IP protection is sought. Then, everything should in principle become open access.

2.8 Responsibilities and Resources

2.8.1 Who will be responsible for data management?

Pierre-Yves Burgi of the University of Geneva will be in charge of data management, and will regularly revise this DMP.

2.8.2 What resources will you require to deliver your plan?

Pierre-Yves Burgi will assume the role of data expert. No additional resources will be needed as this project is not intended to generate data but rather is intended to archive data based on new principles using DNA as the medium storage. As indicate above, price for long-term archiving should remain modest and easily absorbed in the general budget. This situation will however be re-evaluated in the course of the project.

3. PLANNED RESEARCH OUTPUTS

3.1 Physical object - "New archiving service based on DNA"

The main output of this project is a proof-of-concept of an end-to-end system for archiving information within DNA synthetic molecules.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
New archiving service based on DNA	Physical object	2026-10-01	Open	None specified	Non	Creative Commons Attribution (Non) Commercial (No) Derivatives 4.0 International	None specified	No	No