Autonomous Robotic Inspection and Maintenance on Ship Hulls and Storage Tanks

Deliverable report – D10.6 (iii) Data Management Plan

Context	
Deliverable title	Data Management Plan
Lead beneficiary	CNRS
Author(s)	Laura MONNIER
Work Package	WP10
Deliverable due date	March 2024 (M51)
Document status	
Version No.	1
Type	ORDP: Open Research Data Pilot
Dissemination level	**Public**
Last modified	26 March 2024
Status	RELEASED
Date approved	26 March 2024
Approved by Coordinator	Prof. Cédric Pradalier (CNRS) Signature:
Declaration	Any work or result described therein is genuinely a result of the BUGWRIGHT2 project. Any other source will be properly referenced where and when relevant.

This project has received funding form the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 871260

TABLE OF CONTENTS

History of changes 1

Referenced documents 1

Executive summary 2

I. Data summary 2

1. Purpose of technical data collection/generation and its relation to project’s objectives 2

2. Types and formats of data generated/collected in the project 3

3. Re-use of existing data 6

4. Origin of the data 6

5. Expected size of data 6

6. Data utility 6

II. FAIR data 6

1. Making data findable, including provisions for metadata 6

2. Making data openly accessible 7

3. Making data interoperable 8

4. Increase data re-use (through clarifying licences) 9

III. Open access 10

IV. Allocation of resources 10

V. Data security 11

VI. Ethical aspects 11

HISTORY OF CHANGES

Date	Written by	Description of change	Approver	Version No.
05/02/24	Laura Monnier	Collection of data from partners	Cédric Pradalier	1
06/03/24	Laura Monnier	Validation	Cédric Pradalier	2

REFERENCED DOCUMENTS

1. BUGWRIGHT2 Grant Agreement (GA) Number 871260
2. Deliverable D11.3 Development Repositories and Shared Servers
3. D10.6 (i) Data Management Plan
4. D10.6 (ii) Data Management Plan

These documents are stored on the file sharing site hosted by CNRS.

Executive summary

This document describes the Data Management Plan of the project BUGWRIGHT2. In particular this deliverable describes the types and size of the data collected and gathered during the execution of the project; the procedures to be followed to ensure a FAIR (Findable, Accessible, Interoperable, Re-usable) access to data and the possible technical and legal/ethical issues. This document is more a methodology focus, it is used to produce materials and the method used is as important as the production of these materials.

This deliverable describes the open-access approach to be followed for publishing scientific results. The plan will also comply with by the Consortium Agreement provisions in order to protect IP generated by the project.

This is the final version of the Deliverable 10.6, a complete one, delivered at M51. It is taking into account changes, regular updates, especially new data generated and software used for the project. This deliverable is to be considered a live-document.

I. Data summary

1. Purpose of technical data collection/generation and its relation to project's objectives

The main purpose of data collection and generation in the BUGWRIGHT2 project is to document the state of an infrastructure. Meaning the creation of a database of images, videos, schemas, codes and other data aimed at providing meaningful input and/or test data for benchmarking tools.

More precisely, within BUGWRIGHT2, the consortium expects to generate significant knowledge in multiple areas:

- Robotic acoustic tomography
- Autonomous navigation on and around ship hulls and large storage tanks
- Mapping and inspection of large industrial structures
- Multi-robot coordinated mission
- 3D immersive interface and data exploitation for decision support.

In this regard, proper monitoring, protection and management of commercially exploitable IPR (Intellectual Property Rights) are crucial.

The creation of such database meets the objective of BUGWRIGHT2, which is to “bridge the gap between the current and desired capabilities of ship inspection and service robots by developing and demonstrating an adaptable autonomous robotic solution for servicing ship outer hulls” (ref. Grant Agreement n°871260, Annex 1 part B, Ch1.1 “Objectives”).

2. Types and formats of data generated/collected in the project

The data generated and collected in the framework of BUGWRIGHT2 are expected to be mainly images, video streaming and other data such as point clouds, thickness measurements, data sensors and data for the localization of robotic platforms within the confined space they operate in publication only with respect to privacy and security limitations).

The format of data would follow industry standards as much as possible (e.g. .jpeg or .png for images or .mpg for videos, .ppt, .doc, .vsd for documentation or self-report measures, .csv or .xls for data tables or way-point list for mission execution, .hpp .cpp .py for code, .ply or .pcl for point clouds, or .pdf for general reading documents), with preference for the standards for which open source software is available (like the Point Cloud Library). When custom data formats are chosen, suitable documentation will be provided.

Access to navigation data will be offered in rosbag format, a format widely used by the robotic community, in particular in the context of the development of localisation algorithms. Concerning the robots, the data collected will be mainly based on the lidars, cameras, sensors which will monitor the general environment and will be saved in rosbag, the native file of the middleware operating system (ROS).

The robotic platforms developed in the project (such as Altiscan by Roboplanet, Blueye Pro by Blueye, Deeptrekker and DJI Matrice 300 RKT used by Glafcos Marine) use standard format data like .csv, .pdf and .mp4.

ROS (Robot Operating System) provides libraries and tools to help software developers create robot applications. It provides hardware abstraction, device drivers, libraries, visualizers, message-passing, package management, and more. ROS is licensed under an open source, BSD licence. The great flexibility of ROS allows it to be deployed on very different robots (mobile robot, industrial arm, multicopter) and which evolve in various environments (land, air, marine and submarine).

ROS
Robot Operating System

Rosbag is a command line tool used to record and playback ROS message data. Rosbag uses a file format called bags (.bag). Rosbag may be used to store data on the robots including, but not limited to images, videos, code and other raw-sensor data (IMU, GNSS, air pressure, etc.), mission paths planned and executed, and point clouds, before downloading them to a secure storage (the media data used are .jpeg, .png, .mp4). Furthermore, the data can also be extracted for processing and analysis.

UNI-KLU uses ROS data recordings, Robot Operating System data streams including, but not limited to images and other raw-sensor data (IMU, GNSS, air pressure, DVL, water depth, magnetic field, UWB distance, etc.), mission paths planned and executed, localization information in position and attitude of external navigation aids (such as e.g., visual markers, or UWB anchors) and point clouds. Furthermore, UNI-KLU uses Mission way-points for mission execution (.csv), and External navigation aids or markers calibration information such as position and attitude of the markers with respect to a common reference frame that can be used for boot-strapping online localization algorithms (.csv).

LSL also uses ROB for the pose of drones, position and orientation of drones with respect to a common coordinate system defined by ArUco markers, with the format geometry_msgs/TransformStamped.msg. For the inspection of images, the ROS format is sensor_msgs/CompressedImage.msg containing a jpeg. The telemetry data to know the status of the drones during a mission (various ROS messages). For the mission status, they use an array indicating area inspected, corrosion detected and closely inspected, and remaining area to inspect, the format being arrays of integers. And for the mission commands, the various commands to prepare and carry out the mission, including points to define the inspection area (with respect to a common coordinate system defined by ArUco markers), Various ROS messages including an array of ROS geometry_msgs/Point messages are used.

DUNE (DUNE Uniforms Navigation Environment) is a runtime environment for unmanned systems on-board software. It is used to write generic embedded software at the heart of the system, e.g. code or control, navigation, communication, sensor and actuator access, etc. It provides an operating-system and architecture independent platform abstraction layer, written in C++, enhancing portability among different CPU architectures and operating systems. It provides a small footprint control software that is suitable to be used to control autonomous vehicles with low power requirements. This framework provides hardware abstraction, device drivers, libraries, several communication transports, and is based on IMC (Inter-Module Communications) message protocol. The software is open-source and free to use for non-commercial purposes. DUNE is licensed as dual licence (to provide commercial use) based on modified European Union Public Licence V. 1.1 and is mainly used by UPORTO.

IMC (Inter-Module Communications) is a message protocol definition designed by UPORTO LSTS to allow interaction and control of unmanned vehicles. Ever since the very beginning IMC has been open-sourced to allow wide usage. This message protocol is designed to be transport agnostic, with several code bindings available. Interconnection between ROS and the UPORTO LSTS toolchain will use IMC through a broker. The serialized IMC data is stored for logging and analysis of data and also can be used for replying data. IMC uses serialised binary IMC, CSV, ROSbag and MP4.

A few partners use Gazebo, a robot simulation. Gazebo is a 3D dynamic simulator with the ability to accurately and efficiently simulate populations of robots in complex indoor and outdoor environments. While similar to game engines, Gazebo offers physics simulation at a much higher degree of fidelity, a suite of sensors, and interfaces for both users and programs. Gazebo is typically use for testing robotic algorithms, designing robots and performing regression testing with realistic scenarios. Gazebo is free and has a huge community worldwide. Furthermore, it is possible to launch Gazebo in a ROS launch file and pass in relevant parameters. Partner INSA is developing a plugging called “SimuDronesGR” (available for partners) which handles .xacro and .dae type files/data.

RWTH, as a specialist in visualization and user interfaces, uses the Unreal Engine in the project. The Unreal Engine is a game development engine, which uses the C++ and C# programming languages together with a visual scripting language to code applications. The usage of a game engine is required to allow for a quick development of the needed virtual reality interfaces and 3D visualizations without reimplementing basic concepts. To build the application the Unreal Editor is needed, which is free for non-commercial use and open source to developers under a special license. The applications written with Unreal are able to bundle a lot of media formats which are encoded in ".uasset" and ".umap" files. The project is stored in a ".uproject" file. The source code is stored in ".h" (C++ headers), ".cpp" (C++ source), ".cs" (C# build system files), ".usf" (shader files) and in case of the visual scripting language in ".uasset" or ".umap" files.

In addition to the source code, some 3D models are needed by RWTH. The Unreal Engine can import a variety of standard 3D formats like e.g. ".fbx" and ".gltf" and stores them in ".h", ".cpp", ".cs", ".uproject", ".uasset", ".umap", ".usf" files internally. The 3D models are then used to visualize objects in the application like the different robot types or the ship model.

And RWTH uses a Web-App project Data, it assests and code related to the development of the user-interface for mission planning and monitoring files are available in ".html", ".js", ".json", ".tsx", ".css", ".schema.d.ts", ".ts".

MATHLAB is an interactive calculation software. It makes it possible to carry out numerical simulations based on numerical analysis algorithms. It can therefore be used for the approximate solution of differential equations, partial differential equations or linear systems, etc. For BUGWRIGHT2 MATHLAB is mainly used to solve acoustic equations to obtain maps. It is also used to simply draw curves. It is one of the most used calculation software in the world. And the format used is .m.

COMSOL is a numerical simulation software based on the finite element method. Here this software is used to build the most faithful digital models of the studied structure (boat hull). In addition, it allows to the generation and propagation of acoustic waves to be simulated in a way that is very close to reality. Thus, it is possible to perform deep numerical analyses of physical problems. This software is also a leader in its field. The format used is .mph.

Glafcos Marine, as a service provider, uses technical drawings software such as 2D Designed Models (format .dwg), 3D CAD (format .3dm) and 3D Printed Models (format .stl).

Roboplanet, as an industrial, uses its own software such as Datascan, a field software for the ALTISCAN. It generates .csv raw ultrasonic measures, positions, logs. Datareport, an office software for the .csv ALTISCAN data. It generates .jpg reports.

And Solidworks 2022, a 3D mechanical design software. It generates 3D files type .stl. SolidWorks is professional 3D CAD software for engineering applications. Its functionalities encompass computer-aided design (CAD) and computer-aided engineering (CAE). SolidWorks is extremely rich in tools and features.

The dataset that will be provided by DANAOS of the past records of thickness measurements will be in JSON files.

3. Re-use of existing data

No re-use of any existing data is foreseen at the present stage.

4. Origin of the data

The data collected or generated in BUGWRIGHT2 will be acquired mainly during field tests and simulations in the partner lab but also during the Integration weeks where demonstrations will be organised. Other data may be collected or generated during the development of robotic platforms or during the development/assessment as well as during the development of analysis tools for the collected data.

For Trier University, further self-report measures will be collected through paper-pencil, online surveys, or interviews. Another form of data collection are literature researches.

The project will produce various sets of raw data resulting from the various robots used during this project i.e. the robot navigation: ROV (Remotely Operated Vehicle), MAV (Micro Aerial Vehicle), AUV (Autonomous Underwater Vehicle) or Crawlers, on and around the large structures being inspected.

5. Expected size of data

The size of data will depend on the extent and type of tests carried out during the project. These datasets will comprise a large volume of visual and acoustic data. We estimate that the data gathered by the partners and the materials and its activities should represent **10 to 20 TB** during the project.

6. Data utility

The data collected or generated in BUGWRIGHT2 can be useful for the partners. The data could be used as a test set for the comparison of results obtained by new platforms against benchmarks or results obtained as outcomes of the BUGWRIGHT2 project.

II. FAIR data

1. Making data findable, including provisions for metadata

The images, videos and other data (see sections 1.1 and 1.2) will be discoverable, identifiable and locatable by means of expressive and mnemonic resource names. Data will be categorized according to their type and purpose according to the partners.

Naming conventions will be adopted by each partner to facilitate identification and discovery.

Descriptions of data with meaningful keywords and other metadata will be created as deemed useful, aimed at facilitating data discovery, contextualisation, selection and access. Metadata will be machine- and human-understandable.

Where necessary, version numbers and/or timestamps will be indicated.

All partners will document their data in a different way, either logging relevant data, or using dedicated software (e.g. Gazebo), libraries and IP management systems (e.g. ROS, DUNE). Others prefer to document it after designing, simulating and measuring the components using also MS office (Visio, Excel, Word etc.) or MATLAB. Almost half of the partners do not use metadata standards, as it is not relevant to them. The rest, like end-users, may use internationally recognised Standards and Codes published by relevant organisations, national industry organisations or standardisation institutions (e.g. ISO, IEC or IACS).

2. Making data openly accessible

In line of principle, it is the intention of the consortium to make most of the collected data publicly available at the end of the project, so that they can be re-used by the project partners and by third-parties. Some of the data will be restricted and kept confidential. Critical data will be selected by the consortium members and will be shared with the scientific and innovation communities. The academic partners will be mostly involved in the process. This curated data will be described using the meta-data standard recommended by the chosen repository.

Exceptions to this general principle will be made based on the basis of:

- Any underwater or boat imagery or state of the boat is specific, so it will need to be anonymized;
- Regarding data acquired during on-board tests carried out on real ships, explicit consent from the Ship Owner will be necessary prior to use and publication of the data themselves.

The openly accessible data will be made available on the project website. The data will also be uploaded on a standard repository (Nextcloud) and made accessible to the consortium. The access to the database will be through a login and password (see deliverable D11.3 Development Repositories and Shared Servers).

As already detailed in section 1.2, the main software used to access the data is ROS. The technical data generated is either raw data or processed data provided in the most common storage formats. ROS is an open-source, meta-operating system for robots. ROS is a distributed framework of processes that enables executables to be individually designed and loosely coupled at runtime. These processes can be grouped into *Packages*, which can be easily shared and distributed. ROS also supports a federated system of code *Repositories* that enable collaboration to be distributed as well. A general document about the ROS system is available on their website (http://wiki.ros.org/).

Should custom software be needed for accessing or using the data, the software will be made available or relevant documentation on the data format will be provided.

The data and associated metadata, documentation and code will be deposited on the project servers. Whereas data and metadata are deposited on Nextcloud, a GitLab repository is used for code styles and documentation as it is a free, open-sourced and development collaborative platform.

For Trier University, interview data and self-reported data will be made openly available in an anonymised format that summarises analytic results (e.g., results report, original article, and factsheets) but will not be open access in their raw form.

Access to data and repositories is limited to registered users, for security purposes. The personal data of the registered users (name, last name and email) are accessible only to the system administrator. There will be no need to identify an external person accessing the data as the data will be public. The identity of users and other sensible data will be managed according to laws and regulations in force (e.g. GDPR) and according to recommended practices and standards for data security.

The possibility to also upload the material on a public repository for research data sharing will be explored. However, at the current stage this solution seems non-practicable given the generated data is either raw data or very large.

As there are no ethical problems concerning data access, a committee is not necessary. Should ethical problems arise during the project, a committee will be created.

As the various software used for access are free and open-sourced, their licenses are available online and free of charge.

3. Making data interoperable

Since the developed data will be stored in the most common formats, it is reasonable to expect that data could be re-used with a good level of interoperability. The use of the ROS system to explicit the data types and that both the language-independent tools and the main client libraries are released under the terms of the BSD license, and as such are open source software and free for both commercial and research use will facilitate the interoperability.
Human-related data will be stored in line with standards of good scientific work and ethical guidelines defined by European and national research associations (European Federation of Psychologist Association; Deutsche Gesellschaft für Psychologie) and made available in an anonymized form (e.g., results report, article, PowerPoint presentation).

As data will be generated by the BUGWRIGHT2 partners, different methodologies come into operation. Half of the partners will generate data via research. Others will do different types of measurements and simulations or design flow. However, for some partners the exact methodology is unknown yet but it will be important to use methodologies that are commonly used and known. In case the data gets collected and not generated in BUGWRIGHT2, data originally will come from literature research, internal databases, company internal instrumentation, and through design, historical datasets of end-users, simulations and measurements or survey methods.

The data in the databases will be exposed in a text format following well-known and established standards (e.g., CSV, JSON or XML).

Data collected by the social media tools (LinkedIn and Facebook) and the website will only be used to further improve the quality of the produced output i.e. dissemination. Data related to social media posts and the outputs of the tools analysis will not be made available to anyone and will be used exclusively by the consortium.

In principle the data that will be collected are expected to be available for five years after the project's end.

4. Increase data re-use (through clarifying licences)

The data will be publicly released under **Creative Common Attribution-NonCommercial-ShareAlike** license. The licensing model will be agreed among the partners and will be clearly presented to users before giving access to the data.

Summarizing from the Creative Common website (https://creativecommons.org/licenses/by-nc-sa/4.0/) this license allows to freely:

- Share – Copy and redistribute the material in any medium or format
- Adapt — remix, transform, and build upon the material

Under the following conditions:

- Attribution — One must give appropriate credit, provide a link to the license, and indicate if changes were made. One may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- NonCommercial — One may not use the material for commercial purposes.
- ShareAlike — When remixing, transforming, or building upon the material, one must distribute his/her contributions under the same license as the original.

The consortium will ensure the public access to the generated database starting from the beginning of the third year of the project. There will be no restrictions to the use of data during and after the project that could impair or restrict the access to or the use of data beyond the original intent.

It is foreseen that the data should remain available also after the end of the project, at least four years after the end of the project. The partners intend to preserve data beyond the project duration. Details on long-term preservation of data, including costs and potential value, what data will be kept and for how long, will be defined by the end of the project.

III. Open access

All scientific outcomes will be provided in open access mode and through dataset papers. The project partners will provide free online access to their publications by archiving them in online repositories (Green open access). They also plan to publish their articles in open access mode (Gold open access) when appropriate. A specific budget has been planned to cover the fees that will be charged for articles which will be published in non-Open-Access Journals to provide immediate open access to the readership.

Every scientific outcome generated in the project will be self-archived in three locations to ensure maximum visibility:

- on the partner’s institutional repository,
- or on OpenAIRE Orphan Repository (https://www.openaire.eu)
- on Zenodo (https://zenodo.org/)

Through such repositories, metadata and digital object identifiers could be assigned to BUGWRIGHT2 datasets in order for them to be located via search after the end of the project. The datasets in Zenodo will be preserved in line with the European Commission Data Deposit Policy. The data will be maintained indefinitely (minimum 5 years) ensuring no costs for archiving. It should be noted that any unforeseen costs related to open access to research data in Horizon 2020 are eligible for reimbursement for the duration of the project.

References and links to the publications will also be available on the project website.

Open Access to project data will also be provided. Partners will make sure that open source software developed in the project are shared by the end of the project. They will share them on open platforms such as OpenSLAM (https://openslam-org.github.io/) when appropriate.

It is also worth mentioning that the costs for making data FAIR highly depends on the yet to be specified details of the amount data that will be collected and their processing effort. In any case, all costs related to FAIR data management that will occur during project implementation will be covered by the project’s budget. However, any other cost that may relate to long term data preservation will be discussed among consortium members, but as stated before the services of free of charge research data repositories will be pursued.

IV. Allocation of resources

The costs for making the data FAIR in BUGWRIGHT2 are essentially related to the Gold access mode, the purchase of a server if needed and a suitable license. The costs will be covered mainly by the planned budget of the Coordinator for the creation and maintenance of the website (WP10) and for management activities (WP11).

The resources for long-term preservation, including costs and potential value, what data will be kept and for how long, will be defined by the end of the project.

V. Data security

Scientific data is stored in a server which is physically located at CNRS-GTL (Metz, France) and protected by a firewall (see deliverable D11.3 Development Repositories and Shared Servers).

University Trier (Germany) will store the human-related data collected at the University server, which is subject to the same privacy rules as France. The following safety measures (hardware and software) have been secured: backup copies, antivirus software, and password-protected access.

On the security side, a secure channel for sensitive information exchanged on the Internet has been to enable HTTPS, also known as SSL (secure socket layers) so that any information going to and from the server is automatically encrypted. The BUGWRIGHT2 project does not involve any sensitive data raising security issues or any ‘EU-classified’ information. In addition, whenever possible, the other partners of the consortium will keep copies of the data sets to ensure some redundancy against possible failures.

Key data will be backed up at CNRS’s premises and stored on servers. If needed, the CNRS will purchase a server for data storage purpose.

VI. Ethical aspects

No ethical aspects concerning data sharing is expected. If any should raise (e.g. images capturing unexpected people passing by), proper actions will be taken, e.g. data removal. At the current stage is foreseen that the database will not contain any personal information.