BUGWRIGHT2

Autonomous Robotic Inspection and Maintenance on Ship Hulls and Storage Tanks

Deliverable report – D10.6 Data Management Plan

Context		
Deliverable title	Data Management Plan	
Lead beneficiary	CNRS	
Author(s)	Laura MONNIER	
Work Package	WP10	
Deliverable due date	June 2020	
Document status		
Version No.	1.1	
Type	ORDP: Open Research Data Pilot	
Dissemination level	**Public**	
Last modified	07 August 2020	
Status	RELEASED	
Date approved	07 August 2020	
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

Date	Written by	Description of change	Approver	Version No.
26/06/2020	Laura Monnier	Initial version	Cédric Pradalier	1.0
07/08/2020	Laura Monnier	Revision + validation	Cédric Pradalier	1.1

REFERENCED DOCUMENTS

1. BugWright2 Grant Agreement (GA) Number 871260
2. Deliverable D11.3 Development Repositories and Shared Servers

This document will be 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. Finally, this deliverable is to be considered a live-document and is subject to regular updates, to cover possible changes and new data generated, that is at M24 and M48.

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.

The format of data would follow industry standards as much as possible (e.g. .jpeg or .png for images or .mpg for videos, .csv or .xls for data tables, .hpp .cpp .py for code, .ply or .pcl for point clouds), with preference for the standards for which open source software is available. 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 localization 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).

Rosbag is a command line tool used to record and playback ROS message data. Rosbag uses a file format called bags. Rosbag may be used to store data on the robots (images, videos, code, point clouds) before downloading them to a secure storage. Furthermore, the data can also be extracted for processing and analysis.

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).

In principle, the collected/generated data can be divided in the following classes:

Entity	Name	Type	Format

A more detailed overview of the data will be collected as the project progress.

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.

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). Others prefer to document it after designing, simulating and measuring the components using also MS office 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.

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.

To ensure that human-related data can be re-used by as many researchers as possible, data will be stored and made available 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).

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, simulations and measurements.

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/)

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.

IV. Allocation of resources

The costs for making the data FAIR in the Bugwright2 project 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.