Data Management Plan

Deliverable 8.2

Deliverable Title	D8.2 Data Management Plan
Deliverable Lead:	University of Bremen (UOB)
Related Work Package:	WP8: Project management
Related Task(s):	T8.3: Innovation, data and knowledge management
Author(s):	Prof. Michael Beetz
Dissemination Level:	Public
Due Submission Date:	07/31/2021
Actual Submission:	07/28/2021
Project Number	101017089
Instrument:	Research and innovation action
Start Date of Project:	01.01.2021
Duration:	51 months
Abstract	This deliverable describes the initial Data Management Plan of the TraceBot Project. The key method to make data visible and accessible is the cloud-based openEASE web-service, that is developed and hosted by the University of Bremen. Key experimental results are targeted for data release on openEASE. This will allow internal partners as well as third parties to access data in a transparent and structured way.

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089), 2021-2025

Versioning and Contribution History

Version	Date	Modified by	Modification reason
v.01	14.07.21	Prof. Michael Beetz(UOB)	Initial version
v.02	26.07.21	Prof. Michael Beetz(UOB)	Internal revision

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

Table of Contents

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

Horizon 2020
TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

1 Executive Summary / Introduction

The TraceBot project focuses on the development of new robotic approaches in the domain of lab automation. To tackle this challenge a large variety of data will be generated, processed and stored over the course of the project.

This includes very specific data in the form of geometric models up to data about semantic execution traces of robot executions. Our goal is to share the data from the project as transparent as possible.

The key method to fulfill this goal is the cloud-based openEASE web-service, that is developed and hosted by the University of Bremen. Key experimental results are targeted for data release on openEASE. This will allow internal partners as well as third parties to access data in a transparent and structured way.

The structure of this deliverable directly follows the suggested outline of the template data management plan from the website of the Horizon 2020 program [1].

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

2 Data Summary

2.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

The main purpose of data collection is to evaluate the performance of the developed systems and demonstrators that will be developed in the TraceBot project. It will also provide means to analyze structured execution traces of traceable laboratory actions and ultimately provide legible data.

Data is stored in the form of NEEMs (narrative enabled episodic memory). More information can be found in the NEEM-Handbook[2].

2.1.2 What types and formats of data will the project generate/collect?

We aim to use the service of the openEASE platform to openly publish the data and make it available to others. After the original data is stored in an openEASE server, the data can be visualized and collected according to the standards defined by openEASE. The structured execution traces will also include environmental and object models, sensor data(e.g. vision, tactile) as well as robot configuration streams.

openEASE is a web-based knowledge service targeting the easy and structured access of robot activity data. It provides an interface to query semantically annotated data of robot manipulation actions. The data representation also contains references to signals, beliefs and sensory streams occurred during the manipulation actions to ultimately provide a comprehensive tool to analyze robotic systems.

It is planned to support the collection of relevant raw data used in the submodules of demonstrated experiments of all TraceBot partners, by uploading them to openEASE as well and therefore making them available. Experiments could for example be full-stack robot manipulation demos or specific use-case demonstrations which are part of the complete TraceBot scenario.

2.1.3 Will you re-use any existing data and how?

No.

2.1.4 What is the origin of the data?

The data is produced by:

- Robot sensors
- Robot control software
- Observations by the developers
- Manual creation of data by the developers

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

2.1.5 What is the expected size of the data?

Several MiBs to several GiBs, depending on the duration of the experiment, the bandwidth of the sensor data, and the level of detail of the stored data.

2.1.6 To whom might it be useful ('data utility')?

- Scientific evaluators, seeking to confirm the validity of the results.
- Researchers interested in measuring the performance of their algorithms against known data sets using real-world data.
- Other TraceBot partners.

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

3 FAIR Data

3.1 Making data findable, including provisions for metadata

3.1.1 Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

We will use the openEASE mechanisms to generate an DOI for each experiment, which will group several types of data streams. Each collection of NEEMs (Narrative Enabled Episodic Memory) will be stored as a git repository on a Gitlab which is part of the NEEM-Hub at the University of Bremen. DVC (Data Version Control) is used for versioning. Each repository is linked to a metafile in JSON format and will contain at least name, a description a list of keywords and unique ID.

OpenEASE acts as a front end for the NEEM-Hub (https://neemgit.informatik.uni-bremen.de/neems) and will utilize the metafiles to make the NEEMs searchable. The Gitlab can also be used to search for repositories. It allows the lookup of data sets by searching for them by name and author.

3.1.2 What naming conventions do you follow?

Each openEASE experiment related to TraceBot will be named *TraceBot-<experimentname>-<version>.*

3.1.3 Will search keywords be provided that optimize possibilities for re-use?

Yes, the researchers will be able to choose keywords freely, which get attached to the metadata describing the experiment.

3.1.4 Do you provide clear version numbers?

Yes, a DVC system is used.

3.1.5 What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Gitlab already provides the following information:

- Name of the uploader
- Copyright holder
- License
- Version
- Date of publication

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

The metafiles of the experiments contain the following entries:

- Name for the experiment
- List of keywords
- Description

We recommend, that the description contains at least the following additional information:

- Date in which the data was recorded
- DOIs / URLs of related scientific publications

3.2 Making data openly accessible

3.2.1 Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

The data used to produce scientific publications by the academic partners is the main target for publication. If the data produced contains confidential business information originating from the industrial partners, the data will be modified before publication to protect their business interests unless agreed otherwise by the affected industrial partner.

We will encourage all the researchers producing project related publications to publish also their data according to this DMP, but they have the option to decline after stating their reasons.

3.2.2 How will the data be made accessible (e.g. by deposition in a repository)?

By uploading the data as an experiment in the openEASE web platform.

3.2.3 What methods or software tools are needed to access the data?

We offer two options to interact with the data via openEASE.

For visualizing the supported types of data, only an internet browser is necessary. This is a very low barrier of entry made possible by the openEASE service. Semantic information can be queried through standardized languages. To have full access to the data, it will be possible to download the datasets and work on them locally using openEASE and ROS tools.

It is also possible to get the data by using DVC, git and hadoop to pull the data directly from the NEEM-Hub and work locally with KnowRob or their own tools.

3.2.4 Is documentation about the software needed to access the data included?

Yes, this is available in the openEASE website and in the NEEM handbook.

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

3.2.5 Is it possible to include the relevant software (e.g. in open source code)?

The openEASE website currently does not support direct linking of relevant open source software.

3.2.6 Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

The data will be deposited on the openEASE server that is hosted at the University of Bremen.

3.2.7 Have you explored appropriate arrangements with the identified repository?

Yes, the openEASE project is managed by the Artificial Intelligence department of the University of Bremen, which is a partner in the TraceBot project and has agreed to host the data generated in the project.

3.2.8 If there are restrictions on use, how will access be provided?

There are currently no restrictions: the data hosted there is available freely without cost.

3.2.9 Is there a need for a data access committee?

No.

3.2.10 Are there well described conditions for access (i.e. a machine readable license)?

Yes, a machine readable LICENSE.md file will be stored in the git repository of the openEASE NEEM-Hub Gitlab at the University of Bremen.

3.2.11 How will the identity of the person accessing the data be ascertained?

Users who want to upload data need to create an account. Accessing the data is possible without one.

The creation of users in the system is administered by the TraceBot project members of the University of Bremen(UOB). Every partner of the TraceBot project can contact UOB to get an account for data uploading.

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

3.3 Making data interoperable

3.3.1 Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

The data produced in the project is interoperable using the tools provided by the openEASE platform. It offers a low barrier access using a web server and a modern browser-based client. It also offers its' own standard for data storage and complete data access, which is based on JSON and MongoDB.

Additionally, you can download the data directly from the NEEM-Hub.

3.3.2 What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

We will follow the openEASE method that focusses on ensuring compliance with the data structures of the Robot Operating System (ROS) to ensure interoperability within the robotics community.

3.3.3 Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

We are supporting inter-disciplinary interoperability since any kind of data can be uploaded to the NEEM-Hub and also accessed by anyone.

3.3.4 In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Yes, the SOMA ontology (https://ease-crc.github.io/soma/) used by openEASE uses DOLCE+DnS Ultralite (DUL) as upper-level ontology.

3.4 Increase data re-use (through clarifying licences)

3.4.1 How will the data be licensed to permit the widest re-use possible?

The researchers will have the option to choose their preferred data license, but we strongly suggest to use one of the Creative Commons(CC) licenses, for example CC-BY-SA, which grants

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

the right to re-distribute the data and make derivative works only if they give the authors credit, and that those derivative works are also released using the same license.

If the project produces code to be released to the public, a free-software license is recommended, such as the GNU General Public License (GPLv3+ or LGPLv3+). If the researchers desire a more permissive license, the Apache 2.0 License is recommended. The choice of license is left to the authors/copyright holders.

3.4.2 When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

The data will be available for the selected scientific publications of the project once the respective papers have been accepted for publication. The data should be available no later than 6 months after the submission of the final version of the manuscript.

3.4.3 Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

Yes, the data is reusable by third parties, even after the end of the project.

3.4.4 How long is it intended that the data remains re-usable?

Long term reusability of the data is a difficult issue given the speed of change of the robotics software eco-system. For this we will rely on the assurances of the openEASE service, and it is expected to be at least ten years from now (until 2031). This ensures an enduring data availability for following projects and research endeavors. To fulfill this goal, openEASE implements multiple measures to ensure the availability and security of the data. The data is backed up by multiple servers in different buildings to ensure that hardware defects will not cause a loss of data. To prevent unauthorized access and manipulation of the system, the management interfaces of the servers are secured. The physical access to the system is also strongly limited.

3.4.5 Are data quality assurance processes described?

No.

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

4 Allocation of resources

4.1.1 What are the costs for making data FAIR in your project?

There is the time budget needed to collect, organize, and prepare the data for publication, which is covered by the authors. The costs of the data hosting and openEASE services are covered by the University of Bremen, currently the only responsible for the openEASE service.

4.1.2 How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

N/A.

4.1.3 Who will be responsible for data management in your project?

The University of Bremen as the host of the openEASE/NEEM-Hub server will act as the responsible for the data management within TraceBot.

4.1.4 Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Not discussed.

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

5 Ethical aspects

5.1.1 Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

No.

5.1.2 Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Currently we don't expect data sets that include personal information.

If this changes over the course of the project, we will revise the DMP accordingly. In general, personal data would be treated confidentially and will only be processed for the purposes of the TraceBot project management and related dissemination and communication activities. GPDR-related issues would also be considered and outlined in the DMP if personal data would be featured in the data sets in the future.

6 Other issues

6.1.1 Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

No.

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

7 Conclusion

TraceBot is developing novel ways to enable robots to verify their own actions to ultimately perform more robustly. A key to reach that goal is the availability of rich semantic data which the robot can analyse to reason about the correct executions of actions, by building on a large database of episodic memories and other relevant data.

In TraceBot, we will use the openEASE platform as a foundation for such kind of data, which not only enables robots to analyse execution traces, but also allows researchers worldwide to interact with this kind of data in a structured way. The consortium will continuously analyse the data requirements to support verifiable robot manipulation scenarios and provide this data via openEASE. In general, data will be published transparently which makes re-use by other parties easy and therefore directly supports the transfer of the research results of TraceBot into the industry, society and academia.

8 References

[1] EUROPEAN COMMISSION Directorate-General for Research & Innovation - Guidelines on FAIR Data Management in Horizon 2020; URL: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] Michael Beetz, Daniel Beßler, Sebastian Koralewski, Mihai Pomarlan, Abhijit Vyas, Alina Hawkin, Kaviya Dhanabalachandran, Sascha Jongebloed - NEEM Handbook; URL: https://ease-crc.github.io/soma/owl/current/NEEM-Handbook.pdf

Horizon 2020

TraceBOT receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)