H2020-FoF-02-2016. 723853 COROMA

Data Management Plan

COROMA

Cognitively enhanced robot for flexible manufacturing of metal and composite parts

D1.4
Data Management Plan

Deliverable name	D1.4 – Data Management Plan
Due date	31/03/2017
Delivery date	31/03/2017
Authors	Alexander Fabisch, DFKI-RIC Tom Runge, DFKI-RIC
Responsible of the deliverable	Alexander Fabisch, DFKI RIC alexander.fabisch@dfki.de
Version	Version 1
Dissemination level	Public

Table of Contents

LIST OF ACRONYMS

Acronym	Title
IDK	Ideko S.Coop.
UOS	University of Sheffield
UNA	University of Nantes
KTH	Royal Institute of Technology
DFKI	German Research Center for Artificial Intelligence
STA	Stäubli Faverges SCA
ITR	IT+Robotics s.r.l.
CIT	Convergent Information Technologies GmbH
BAS	BA Systemes SAS
SRC	Shadow Robot Company Ltd.
SOR	Soraluce S.Coop.
EUT	Europe Technologies SAS
ENSA	Equipos Nucleares S.A.
BEN	Beneteau S.A.
ACI	Aciturri Metallic Parts S.L.U.
DIN	German Institute for Standardization

1. DATA SUMMARY

1.1 OBJECTIVE

This document constitutes the first version of the Data Management Plan (D1.4) of COROMA project, elaborated under Task 1.4 Data management plan for Open Research Data Pilot of Work Package 1 Requirements definition.

The objective of this deliverable is to give an overview of the data that will be collected during the runtime of the COROMA project. This document will define processes how data will be stored, published, and distributed amongst the consortium partners. COROMA is a large project with many partners which makes a good data management strategy to share and archive documents inevitable.

Project partners will collect data in order to ensure that all conducted experiments are reproducible and to share knowledge between partners within the project and with other researchers.

A key objective of the COROMA project is to develop the robot into an autonomous system. Autonomous systems rely more and more on experiences and collected data therefore data can be used to learn and improve various cognitive components of the robot as well as for validation of approaches and reproduction of results.

1.2 TYPES AND FORMATS

Several types of data will be produced during the project runtime. These are for example:

- Experimental data (e.g. sensor data)
- Equipment data and environment data (e.g. CAD files)
- Configuration files for software and hardware
- Illustrations and videos
- Documents, e.g. dissemination and communication material (publications, posters, presentations, etc.), deliverables
- Software

It is important that data is provided in standard formats that are easily accessible with standard software which should ideally be free of charge to simplify sharing the data among partners or to the public. Examples of preferable formats are

- Documents (e.g. publications, presentations): PDF
- Illustrations: PDF, SVG
- Images: high quality, e.g. TIFF, BMP, JPEG 2000
- CAD: ISO-10303-21 (STEP file)
- Dummy part data / process related data: SciLab, GNU Octave, CSV

Process related data such as force measurements, distance measurements, etc. must be provided in SI.

1.3 OVERVIEW

This Data Management Plan will briefly describe the general categories of data that will be generated during the project and in some cases it will already define specific details about the data that will be produced in this section. However, everything is subject to change. This part of the data management plan will be updated as soon as we know more details and it will form the basis of deliverable 8.5 “Data Compilation Open Research Report”.

Partner	Data	Details in Section	Type	Shared with
ITR	Models of parts	1.3.1	CAD models	Consortium
ITR	Semantic dataset	1.3.2	Sensor measurements	Public
SRC	CORO-hand	1.3.3	CAD / simulation models and metadata	Consortium
SRC	Experimental Data (SRC)	1.3.4	Sensor measurements	Specific partners
IDK	Experimental Data (IDK)	1.3.5	Sensor measurements	Public
BEN	Demonstration data (BEN)	1.3.6	Sensor measurements and metadata	Public
ACI	Demonstration data (ACI)	1.3.7	Sensor measurements and metadata	Consortium and public (after approval)
ENSA	Demonstration data (ENSA)	1.3.8	Sensor measurements and metadata	Consortium or public (after approval)
UNA	Experimental data (UNA)	1.3.9	Sensor measurements and metadata	Specific partners
All	Publications	1.3.10	Documents	Public
All	Internal documents	1.3.11	Documents	Consortium
All	Deliverables	1.3.12	Documents	Public or consortium
DFKI	Machine learning datasets	1.3.13	Datasets	Unknown

1.3.1 PART MODELS

Responsible Partner: ITR

Description: CAD models of products

Purpose: Evaluation of visual scanning methods or procedures

Type / Format: CAD, STL (Stereo Lithography Interface format)

Metadata: -

Required Software: STL files can be opened with e.g. FreeCAD or Blender

Data Collection: CAD models will be provided by partners that are involved in the demonstration

Linked to Publications: -

License / Access: Shared with the consortium

1.3.2 SEMANTIC DATASET

Responsible Partner: ITR

Description: Dataset for Semantic Segmentation

Purpose: Training set for new algorithms, can be used for benchmarking

Type / Format: Dataset for machine learning, ROS Bags (ROS logfiles), PCD (point cloud library document), XML

Metadata: -

Required Software: ROS Bags can be opened with the robot operating system (rosbag, rqt_bag), PCD can be opened with pcd_viewer

Data Collection: Dataset will be acquired using 3D sensors in experimental setups

Linked to Publications: -

License / Access: Shared publicly

1.3.3 CORO-HAND (SRC)

Responsible Partner: SRC

Description: (1) CAD and simulation data for mechanical parts, schematics and CAD for electronics, (2) videos, images

Purpose: (1) used to prototype, design and verify the design of the CORO-hand, (2) dissemination and communication activities

Type / Format: CAD data: STEP (ISO 10303), simulation models: URDF[1]

Metadata: -

Required Software: URDF files can be used and visualized with the robot operating system (ROS), ROS supports multiple operating systems but Ubuntu is preferred and has several tools to visualize data (e.g. rviz), STEP files can be opened by any 3D CAD package (e.g. Solidworks, AutoCAD, Blender)

Data Collection: SRC designs and provides the CORO-hand models

Linked to Publications: -

License / Access: Design files shall not be made publicly available. Any CAD and simulation models of the CORO-hand that are made available to the consortium shall be regarded as confidential and may be simplified models which have been de-featured, e.g. external geometries only.

1.3.4 EXPERIMENTAL DATA (SRC)

Responsible Partner: SRC

Description: (1) Sensor data (joint angles, joint torques, motor temperature/voltage/current, tactile sensors if they are used), (2) control data (controller status/commands, driver status/commands)

Purpose: Data is used to monitor the health of the CORO-hand, debugging and performance characterization

Type / Format: ROS Bags (ROS logfiles)

Metadata: Data shall be associated to which objects are being grasped

Required Software: ROS Bags can be used and visualized with the robot operating system (ROS), ROS supports multiple operating systems but Ubuntu is preferred and has several tools to visualize data (e.g. rviz)

Data Collection: (1,2) shall be recorded either live from using the CORO-hand or during simulation

[1] http://wiki.ros.org/urdf

Linked to Publications: -

License / Access: Sensor data coupled with simulation models might be useful to consortium members, but shall remain private unless specifically requested.

1.3.5 EXPERIMENTAL DATA (IDK)

Responsible Partner: IDK

Description: Robot dynamics dataset

Purpose: Monitoring of the dynamics of the robot during the machining process it is involved (as a fixture)

Type / Format: Accelerometer signal, time signals and/or FFT signals, force signal, time signal, Matlab MAT (*.mat) file[2]

Metadata:

- Estimated acceleration measurement frequency: 5000 Hz
- Acceleration unit: m/s²
- Force unit: Newton

Required Software: Matlab files can be opened with Matlab, SciLab, GNU Octave, etc.

Data Collection:

- Ingesys IC3 real time control system
- Industrial or laboratory accelerometer
- Force signal: force sensing plate, robot tool tip 1 or 6 axis force sensor

Linked to Publications: Linked to two peer reviewed scientific papers, one about using the robot as a mobile fixturing system, one about robotic drilling

License / Access: Shared publicly

1.3.6 DEMONSTRATION DATA (BEN)

Responsible Partner: BEN

Description: Technical data recorded during the demonstration at the facilities of BEN and corresponding metadata

Purpose: Technical documentation and analysis, reproduction of results in a paper

Type / Format: Video, images (JPEG 2000), logfiles

Metadata: -

Required Software: -

Data Collection: Data will be logged and recorded during the demonstration

Linked to Publications: Linked to a publication about the demonstration

License / Access: Shared publicly

[2] https://www.mathworks.com/help/pdf_doc/matlab/matfile_format.pdf

1.3.7 DEMONSTRATION DATA (ACI)

Responsible Partner: ACI

Description: Technical data recorded during the demonstration at the facilities of ACI and corresponding metadata

Purpose: Technical documentation and analysis, reproduction of results in a paper

Type / Format: Video, images (JPEG 2000), logfiles

Metadata: -

Required Software: -

Data Collection: Data will be logged and recorded during the demonstration

Linked to Publications: -

License / Access: Data related to the process itself, images, or videos without produced parts can be shared with the consortium and publicly after approval. Images and videos of the produced parts are property of the customers and cannot be shared.

1.3.8 DEMONSTRATION DATA (ENSA)

Responsible Partner: ENSA

Description: Technical data recorded during the demonstration at the facilities of ENSA and corresponding metadata

Purpose: Technical documentation and analysis, reproduction of results in a paper

Type / Format: Video, images (JPEG 2000), logfiles

Metadata: -

Required Software: -

Data Collection: Data will be logged and recorded during the demonstration

Linked to Publications: -

License / Access: Videos or pictures must be recorded by ENSA personnel. Logged and recorded data must be approved by ENSA and its customers before released to public. Customer information must not be published at all.

1.3.9 EXPERIMENTAL DATA (UNA)

Responsible Partner: UNA

Description: Experimental data (process monitoring, environment monitoring), equipment data, and configuration files

Purpose: Benchmarking, evaluation of a method or procedure, reproduction of results in papers

Type / Format: Equipment data (part file, path file equipment model file, environment model file), configuration file (robot position file, process conditions), videos, images (high quality, e.g. PNG, TIFF, BMP, JPEG 2000), CAD (STEP file, CATPART, IGS, STL, IFC, …), process related data (SciLab, GNU Octave, Matlab, Excel, PsiConsole (AGV)), force measurements (CSV)

Metadata: -

Required Software:

- Required licenses: CATIA V5, Matlab, Autodesk / Powermill, MS Word
- Dependencies: Matlab robotic tool box, Windows 7

Data Collection: -

Linked to Publications: Possibly

License / Access: Data will be shared according to WP distribution and limited to collaboration

1.3.10 PUBLICATIONS

Responsible Partner: UNA, IDK, BEN, …

Description: Scientific or industrial publications

Purpose: Dissemination of project results, making results reproducible

Type / Format: Publication, MS Word, LaTeX, PDF

Metadata: -

Required Software: MS Word, PDF viewer

Data Collection: -

Linked to Publications: -

License / Access: Scientific publications must be open access

1.3.11 INTERNAL DOCUMENTS

Responsible Partner: all

Description: Internal documents, confidential documents or temporary documents

Purpose: Documents that are required to communicate among the partners to achieve the project goals

Type / Format: Documents, PDF, DOCX, XLSX, PPTX

Metadata: -

Required Software: MS Office, Adobe Acrobat Reader

Data Collection: -

Linked to Publications: -

License / Access: Shared within the consortium

1.3.12 DELIVERABLES

Responsible Partner: all

Description: Deliverables

Purpose: Documentation and planning of the project COROMA

Type / Format: PDF

Metadata: -

Required Software: PDF viewer

Data Collection: -

Linked to Publications: -

License / Access: Public or confidential, as defined in the GA (annex 1, part A, pp 6-10)

1.3.13 MACHINE LEARNING DATASETS

Responsible Partner: DFKI

Description: Datasets that will be preprocessed and used for machine learning

Purpose: Benchmarking, reproduction of results in papers

Type / Format: -

Metadata: -

Required Software: -

Data Collection: Data will be collected by partners in the project COROMA

Linked to Publications: Probably

License / Access: Depends on the owner of the original data

2. FAIR DATA

In COROMA, three different kinds of data will be aggregated and shared within the project:

(1) Public data and documents
(2) Restricted data
(3) Internal documents

Task leader DFKI will rely on the platform Zenodo[3] for data that has to be made publicly available (1) and DFKI will maintain an overview of publicly shared data and documents at the COROMA website[4]. Note that as COROMA is an H2020 project, all publications must be open access. For data that will be shared with a selected audience (e.g. the project consortium), DFKI *recommends* to use Zenodo as a platform which supports detailed access control (2). It enables interested researchers to request access to restricted data directly from the owner of the data and it allows the owner of the data to define and revoke access rights per request. For internal documents (3) such as minutes, confidential deliverables, internal presentations, etc. the internal project website[5] will be used to share these documents among the project consortium.

2.1 MAKING DATA FINDABLE

Publicly shared data produced in the project COROMA must be findable. Zenodo provides search functionality which will guarantee that COROMA datasets can be found. The search functionality will for example allow filtering by keywords and name of the data repository. Each responsible partner must provide appropriate keywords during the data publication process. These keywords should assist people that search for the data at Zenodo. Each upload receives a digital object identifier (DOI) which will make it easily findable even though the URL to the dataset might have changed. In addition, we will provide an overview of publicly available datasets and documents at our project website.

All partners are encouraged to upload restricted data to Zenodo. Zenodo allows restricted access for confidential data. Note that the data is still findable although it is not visible if the access is restricted.

The internal COROMA website provides a section to store documents that have been produced within the project. It will be used mostly to share confidential data between partners of the project. Documents will be findable through the hierarchical directory structure (e.g. ordered by categories

[3] https://zenodo.org/communities/coroma-project

[4] Datasets will be shared at http://www.coroma-project.eu/data-and-results/

[5] http://www.coroma-project.eu

like deliverables, meetings, general information, work packages, etc.). The project website will be available at least as long as the project COROMA is running.

2.2 MAKING DATA OPENLY ACCESSIBLE

Due to the industrial character of the COROMA project and the participation of industrial partners, not all data generated within COROMA can be made public according to the Consortium Agreement. However, important scientific data that can be used to reproduce and validate results must be made openly accessible and publications must be published open access. The data will be published in forms that are specified at the beginning of Section 2. The data publication process is described in Section 2.5.

Only free and standard software *should* be required to use the data. Specific formats are described in Section 1 of this document. Some proprietary tools produce data in formats that can only be used by these tools. These tools must either be standard tools that are used by experts that are interested in the data or there must be a way to extract relevant information with free software. If neither the first of the second case is applied, there is no reason to release the data. All partners must be aware of these issues and there must be a good reason to violate these guidelines (e.g. unreasonable effort to convert a proprietary format).

Access to restricted data will be controlled by the responsible partners. Zenodo allows searching for restricted datasets and it allows requesting access to these datasets over the platform. The responsible partner has to decide on each individual case whether the access will be granted.

2.3 MAKING DATA INTEROPERABLE

Data must be interoperable. Each dataset that is released will have sufficient documentation included that describes the process of reading and using the data. Sufficient means that professionals will easily be able to make use of the data. For example, in the case of a publication in PDF format, no further documentation would be required. The documentation should be plain text or PDF that describes the required tools and procedures to make use of the data. It might be required to provide a short example script in a free programming language that loads and visualizes the data or metadata like information about the columns of a CSV (e.g. measured quantity, units). The format won't be specified any further because the produced datasets and the systems used in the project are so diverse that further constraints could make the cost of releasing datasets unreasonably high.

A good example of an interoperable public dataset is the MNIST dataset[6] which has been used in the machine learning community as a benchmark dataset for more than 15 years. Although the format of the data is very unusual, the website contains a short and sufficient description of the format which makes it transferable in almost any programming language. The website explains the purpose of the dataset, how it is related to other datasets, and it includes a comparison of various methods that have been evaluated with the dataset.

2.4 INCREASE DATA RE-USE

Data and documents have to be published with a corresponding license. The responsible partner must make several decisions that affect the choice of an appropriate license, for example:

- Is commercial use permitted?
- Is modification permitted?
- Is distribution permitted?

[6] http://yann.lecun.com/exdb/mnist/

- Is private use granted?
- Is redistribution mandatory?

Possible licenses for published data and documents are

- Creative Commons Attribution 4.0, CC BY 4.0, https://creativecommons.org/licenses/by/4.0/: allows copying and redistribution, allows adaption for any purpose, cannot be revoked by the author, users must give appropriate credit to the author
- Creative Commons Attribution-ShareAlike 4.0, CC BY-SA 4.0, https://creativecommons.org/licenses/by-sa/4.0/: similar to CC BY 4.0 with the additional obligation to distribute adaptions under the same license
- Creative Commons Attribution-NonCommercial 4.0, CC BY-NC 4.0, https://creativecommons.org/licenses/by-nc/4.0/: similar to CC BY 4.0 with the restriction that commercial use is not possible
- Creative Commons Attribution-NoDerivatives 4.0, CC BY-ND 4.0, https://creativecommons.org/licenses/by-nd/4.0/: similar to CC BY 4.0 with the restriction adaptions must not be shared
- Any other Creative Commons license: https://creativecommons.org/licenses/

Other types of licenses are available for software. The choice of license depends on the specific requirements of the copyright owner. In case software should be released as open source partners can generally choose between licenses that allow commercial use (e.g. BSD) and those that do not (e.g. GNU General Public License, GPL). There are various platforms that help to select an appropriate license.[7]

Public data will be made available as soon as it has been produced and documented, the corresponding publication has been published, or at the latest at the end of M36.

DFKI is responsible for checking whether published data is findable, interoperable, and reusable. The publication process is described in Section 2.5. By relying mostly on standard formats and requiring a documentation for each published dataset, COROMA consortium will make the data reusable as long as possible. No further guarantees are given.

2.5 DATA PUBLICATION PROCESS

We will briefly describe the process to publish public and restricted data and documents at the platform Zenodo. The responsible partner will prepare the data for publication. The preparation includes:

- Definition of appropriate keywords that make the dataset findable for potential users of the dataset or document
- Definition of access rights (open access, embargoed access and embargo date, restricted access)
- Selection of license (see Section 2.4 for details)
- Description of dataset, must be sufficient for experts in the field to make use of the data and for non-experts to understand for which purpose the data can be used
- Optional: implementation of an example that loads the data and demonstrates how it can be used, e.g. by visualization

[7] For example, https://choosealicense.com/ or https://opensource.org/licenses

After the preparation, the responsible partner uploads the data and generated metadata to the COROMA community at the Zenodo platform.[8] DFKI is responsible for checking whether all of the criteria mentioned above have been fulfilled and will either contact the responsible partner to complete the process or directly accept the dataset for the community. After the dataset has been accepted for the Zenodo community, DFKI will contact IDK to add the dataset to the COROMA website.

3. ALLOCATION OF RESSOURCES

Collected data will be prepared for publication in T8.3. Datasets that require documentation and quality assurance are prepared and released by DFKI, IDK, UOS, UNA, and ITR. Other documents that will be used for communication or dissemination will be released in T8.1 and T8.2 respectively and usually do not require much further preparation for release apart from the work that is required to generate these documents (e.g. articles, posters).

Each partner is responsible for his datasets as described in Section 1. Each responsible partner must guarantee that the data is in a format that can be published, documented so that an expert is able to use the data, and can be read with free or standard software. The release process is supervised by DFKI. Details are described in Section 2.5. Each responsible partner must communicate about the release of public data or dissemination documents with DFKI. The release of communication documents must be communicated to IDK.

There are no costs for long-term preservation of public data. The platform Zenodo is free and we cannot give any guarantees beyond the lifetime of Zenodo and its successors.

4. DATA SECURITY

Publicly shared data is stored at the platform Zenodo. Zenodo guarantees to retain the data for the lifetime of the platform which is at least 20 years. Zenodo further guarantees that data is backed up nightly and stored in multiple online replicas. We rely on the platform's security in terms of data recovery, secure storage of restricted data and transfer of sensitive data.

5. SUMMARY

COROMA project partners have completed the data management plan. After analysing the planned data generating activities and the requirements of each partner, a plan has been produced to give an overview of these datasets and give guidelines for data management. In particular DFKI will focus on the process to make data available to the public and how the datasets have to be prepared. This data management plan is intended to be a living document that will be extended during the project and will finally result in the deliverable 8.5 "Data Compilation Open Research Report".

[8] The upload link is https://zenodo.org/deposit/new?c=coroma-project