CORSMAL - Data Management

CORSMAL is generating different data types: sensor data, annotations, performance scores, source code (ROS environment, C/C++ and python) and documents. CORSMAL complies with RCUK, ANR and SNSF standards on research Data Management, Open Research Data, and Open Access publishing. The partners support the October 2017 European Open Science Cloud Declaration Action List and will ensure that other researchers will be able to find, re-use, deposit and share data produced during the project. Shared datasets and documents will have a Digital Object Identifier. Reusability of data is at the core of CORSMAL through the open GitHub Repository and the Challenges, which encourage researchers to share their data, tasks and scenarios and help validate the reproducibility of the results.

Software. CORSMAL uses a GitHub repository (https://github.com/CORSMAL) for public release of software produced and revised according to the standards for reproducibility, modularity, re-usability, and sustainability. Project partners also share a GitHub repository with private privileges to develop, keep synchronised and available any piece of software to easily reproduce components of the system in their own labs, as well as other shared confidential documents. Dedicated ad-hoc GitHub private repositories are created for the development of intermediate algorithms during the project. Libraries, environments, programming languages, simulators include code written in C++, Python, OpenCV, PyTorch, ROS, Gazebo, RViz.

Digital Object Identifier (DOI). A DOI helps cite any piece of data (dataset, software, paper, poster, slides, etc.) and it permanently refers to a location. In addition to DOIs for publications, a DOI will be defined for datasets and the main CORSMAL documents. A Digital Object Identifier (DOI) will help to permanently refer to an Internet location associated with the data produced, especially datasets and source codes, for long-term sustainability and version control. The DOI will also help researchers in the community to cite any piece of data generated out of CORSMAL without having a publication associated with it. For small-size data (up to 50 GB), such as source code, slides, and other documents, CORSMAL will be publicly released on open, FAIR, repositories. Domain-specific data repositories will be preferred, and generic data repositories will be considered as alternatives in case of absence of any domain-specific, open/FAIR, option. As a generic data repository, CORSMAL will prefer Zenodo (https://zenodo.org), an OpenScience platform supported by CERN and EU OpenAire and with a GitHub integration already available. Zenodo will also provide DOIs linked to the data for long-term sustainability and accessibility. For large volumes of data, e.g. multi-modal datasets, and/or sensitive data, which require ethical approval and anonymization, each partner will interact with the appropriate teams and the ITS service in their respective institutions/organizations. QMUL will refer to the QMUL Repository and Research Information Team to discuss and prepare the data and to obtain the DOI.

Datasets. Large-sized data (50 GB) will be stored, managed and maintained internally to each partner's organisation with the responsibility of each partner team. The distribution of each of these data will be through the CORSMAL website. Large-sized data could be difficult to access and download by external users and hence recent solutions should be considered, discussed and properly planned. Examples that CORSMAL is currently

looking at are the Cloud Storage Buckets that enable users to download data locally with the ability to resume a download when interrupted (sharding), or directly consume the data on the cloud without copying the data to a local machine. Such an option should be discussed within each partner institution and the IT service team to know if available and how to better implement it before relying on any external service, such as Google or Amazon AWS. CORSMAL will guarantee the long-term sustainability of its data with last generation backup systems within each institution, e.g. IT service, and via the service provided by GitHub. The IT service of QMUL will provide regular, over-night backups of the server where the CORSMAL website is hosted. QMUL is partnering with Microsoft for the cloud hosting, sharing and backup of institutional data, for example using OneDrive. QMUL in collaboration with GitHub, also provides the GitHub Enterprise service and therefore all data can be hosted, backed up, and tracked the versions within QMUL. The private CORSMAL repository, shared among partners, is part of the QMUL account, thus providing backup, version control, and sustainability over time, even in the case people involved in the project are no longer available (e.g. due to a change of job, retirement, end of project, or any other reasons). CORSMAL publicly releases datasets, including data and their annotations under the public Creative Commons CC0 License or Creative Commons Attribution 4.0 International License. Software will be publicly released under the MIT license or GNU general public licenses (GPL) will be primarily considered. The above licenses allow to copy and redistribute the material in any medium or format (share), to remix, transform, and build upon the material (adapt and re-usability). The license Creative Commons Attribution 4.0 requires in addition to give appropriate credit (attribution) to the authors of the dataset. This license is the standard one provided by Zenodo when uploading and releasing a dataset. While CORSMAL aims to employ software and data that are open source, partners will always verify the copyright and the owner the licence, negotiating the usability of the data and/or a new licence with the supplier in case copyleft license and open source rights are not satisfied. Moreover, CORSMAL will cite and provide appropriate references to the data used in any publication associated with the project in accordance with each institutional policy.

Examples of CORSMAL data

1. CORSMAL Containers Dataset -> https://doi.org/10.17636/101CORSMAL1
2. CORSMAL Containers Manipulation Dataset -> https://doi.org/10.17636/corsmal2
3. Pre-trained Models -> https://zenodo.org/record/4518951#.YC9-z-qnw5k