CloudSkin

Funded by the European Union

HORIZON EUROPE FRAMEWORK PROGRAMME

CloudSkin

(grant agreement No 101092646)

Adaptive virtualization for AI-enabled Cloud-edge Continuum

D2.6 Data Management Plan, 3rd version

Due date of deliverable: 31-12-2025
Actual submission date: 29-12-2025

Start date of project: 01-01-2023 Duration: 36 months

Summary of the document

Document Type	Report
Dissemination level	Public
State	v1.0
Number of pages	14
WP/Task related to this document	WP2 / T2.3
WP/Task responsible	URV
Leader	Marc Sanchez-Artigas (URV)
Technical Manager	Marc Sanchez-Artigas (URV)
Quality Manager	Julen Bohoyo (URV)
Author(s)	Vanesa Ruana (URV)
Partner(s) Contributing	URV, ALT, NCT, IBM, KIO
Document ID	CloudSkin_D2.6_Public.pdf
Abstract	This deliverable presents the third version of the project Data Management Plan (DMP). It is submitted on Month 36 as a Final review of the CLOUDSKIN Data Management Plan.
Keywords	Data Management Plan, Open Access, Open Research Data, FAIR data, ORDP.

History of changes

Version	Date	Author	Summary of changes
0.1	03-11-2025	Vanesa Ruana	First draft.
0.2	05-12-2025	Several partners	Add the datasets.
1.0	29-12-2025	Vanesa Ruana	Final version.

Table of Contents

List of Abbreviations and Acronyms

API	Application Programming Interface
AWS	Amazon Web Services
CC	Creative Commons
Cholec80	Endoscopic video dataset containing 80 videos of cholecystectomy surgeries from University of Strasbourg
CSV	Comma-separated values
DMP	Data Management Plan
DOI	Digital Object Identifier
DSAD	Dresden Surgical Anatomy Dataset
HeiChole	Surgical Dataset for surgical workflow and skill analysis from University Hospital Heidelberg und NCT
SME	small medium enterprise

1 Executive summary

CloudSkin is committed to maintaining robust data management practices. To support the development of a complete data management lifecycle essential for verifying scientific outputs, the third and final version of the Data Management Plan (DMP) has been released as Deliverable D2.6. This updated version details the approaches for ensuring that research data is findable, accessible, interoperable, and reusable. It also summarizes the datasets currently identified for use over the course of the project.

2 Data Summary

The CloudSkin project aims to promote open access to, and the reuse of, research data generated within Horizon Europe initiatives. To this end, CloudSkin is committed to:

- Developing and continuously updating a Data Management Plan (DMP).
- Depositing the project's data in an appropriate research data repository.
- Ensuring that third parties can freely access, analyse, exploit, reproduce, and disseminate the data.
- Providing the necessary contextual information and identifying (or supplying) the tools required to use the raw data for validating the project's research outcomes.

Specifically, this applies to:

- All data (and accompanying metadata) required to validate the results presented in scientific publications.
- Other curated and/or raw data (and associated metadata) that is specified within this Data Management Plan.

The overarching objective of the CloudSkin project is to develop a cognitive cloud–edge continuum platform capable of fully leveraging heterogeneous resources across cloud and edge environments. The platform seeks to identify the optimal distribution of workloads (the "sweet spot") between these layers and to dynamically adapt to application behaviour changes using AI. To support automated deployment, service mobility, and enhanced security, CloudSkin will leverage an innovative container-like execution abstraction based on WebAssembly, adapting and extending existing approaches to enable seamless, secure, and portable execution of both new and legacy applications across the cloud–edge continuum.

Table 1 provides an overview of the existing datasets that will be processed to validate the results of the CloudSkin project.

Table 1: Used datasets

UD1	
Name:	Dresden Surgical Anatomy Dataset (DSAD) [1]
Origin:	University Hospital Dresden and National Center for Tumor Diseases Dresden
Access:	Dresden Surgical Anatomty Dataset
Volume:	11 different organs, 33 patients (20.5G)
Variety:	Organ Segmentation
Frequency of update:	NA
UD2	
Name:	HeiChole [2]
Origin:	University Hospital Heidelberg and National Center for Tumor Diseases Dresden
Access:	HeiChole Dataset
Volume:	33 patients, ∼200GB
Variety:	Surgical Workflow Analysis

Frequency of update:	NA
UD3	
Name:	Cholec80 [3]
Origin:	University of Strasbourg
Access:	Cholec80
Volume:	80 patients, ∼80GB
Variety:	Surgical Workflow Analysis
Frequency of update:	NA
UD4	
Name:	Synthetic video data [3]
Origin:	National Center for Tumor Diseases Dresden
Access:	Synthetic Video Data
Volume:	∼30GB
Variety:	Surgical Workflow Analysis and Segmentation
Frequency of update:	NA
UD5.1	
Name:	Agricultural data analysis - Campo de Cartagena (WDC)
Origin:	WIDHOC Smart Solutions
Access:	Agricultural technical laboratory
Volume:	13 sensors from nov 2021 to June 2023. 15,6 MB
Variety:	Aggregated data from agricultural sensors
Frequency of update:	30 mins
UD5.2	
Name:	Agricultural data analysis - Campo de Cartagena (WDC)
Origin:	WIDHOC Smart Solutions
Access:	Extracted from agricultural sensors in various areas measuring volumetric humidity and flow meter information
Volume:	172 sensors from july 2023 to april 2024. 1,3 MB
Variety:	Aggregated data from agricultural sensors
Frequency of update:	5 mins
UD5.3	
Name:	Agricultural data analysis - Campo de Cartagena (WDC)
Origin:	WIDHOC Smart Solutions
Access:	Extracted from agricultural sensors in various areas measuring volumetric humidity and flow meter information
Volume:	176 sensors from july 2023 to june 2024. 1,2 MB
Variety:	Aggregated data from agricultural sensors
Frequency of update:	5 mins

UD5.4	
Name:	Agricultural data analysis - Campo de Cartagena (WDC)
Origin:	WIDHOC Smart Solutions
Access:	Extracted from agricultural sensors in various areas measuring volumetric humidity and flow meter information
Volume:	172 sensors from sept 2023 to june 2024. 1,5 MB
Variety:	Aggregated data from agricultural sensors
Frequency of update:	5 mins
UD5.5	
Name:	Agricultural data analysis - Campo de Cartagena (WDC)
Origin:	WIDHOC Smart Solutions
Access:	Extracted from agricultural sensors in various areas measuring volumetric humidity and flow meter information
Volume:	176 sensors from sept 2023 to june 2024. 1,5 MB
Variety:	Aggregated data from agricultural sensors
Frequency of update:	5 mins
UD5.6	
Name:	Agricultural data analysis - Campo de Cartagena (WDC)
Origin:	WIDHOC Smart Solutions
Access:	Extracted from agricultural sensors in various areas measuring applied irrigation
Volume:	1 sensor from sept 2023 to mach 2024. 13 kB
Variety:	Aggregated data from agricultural sensors
Frequency of update:	daily
UD5.7	
Name:	Agricultural data analysis - Campo de Cartagena (WDC)
Origin:	WIDHOC Smart Solutions
Access:	Extracted from agricultural sensors in various areas measuring applied irrigation
Volume:	1 sensor from sept 2023 to mach 2024. 14 kB
Variety:	Aggregated data from agricultural sensors
Frequency of update:	daily
UD6	
Name:	Ion images of annotated molecules ready to off-sample
Origin:	METASPACE
Access:	Extracted from the METASPACE pipeline and saved privately in AWS S3
Volume:	161,137 ion images from 11 spatial metabolomics datasets
Frequency of update:	NA

UD7	
Name:	Spatial Spectrometry results from the METASPACE data space
Origin:	METASPACE
Access:	Public dataset from Justus-Liebig-Universität Gießen: `https://metaspace2020.eu/dataset/2019-08-19_11h28m42s`. Replicated and concatenated 1, 3, 7 and 35 times.
Variety:	Metabolomics data in imzML format
Volume:	7GB (1x), 21GB (3x), 49GB (7x) and 250GB (35x)
Frequency of update:	NA
UD8	
Name:	Enviromental Data from BeOpen Project
Origin:	Cartagena Council
Access:	Agricultural technical laboratory
Volume:	Sensor information from range (08/11/2012 - 07/06/2023)
Frequency of update:	NA
UD9	
Name:	Illumination and temperature Data from BeOpen Project
Origin:	Cartagena Council
Access:	Agricultural technical laboratory
Volume:	Sensor information from range (01/01/2023 - 30/09/2023)
Frequency of update:	NA
UD10	
Name:	Initial infrastructure performance data for Agriculture dataspace
Origin:	KIO Network
Access:	Agricultural technical laboratory
Volume:	Platform performance information from M1-M18
Frequency of update:	NA

In addition to these datasets and benchmarks, the CloudSkin project is anticipated to produce other data to validate the findings outlined in scientific publications, including test data, APIs, and source code utilized for analysis purposes. All such data will be released as open data, promoting its reuse. Table 2 offers an overview of the datasets already generated during the validation process of the CloudSkin project results.

Table 2: Generated dataset

GD1	
Name:	The Dresden Dataset for 4D Reconstruction of Non-Rigid Abdominal Surgical Scenes
Description:	The **D4D** dataset offers paired endoscopic stereo video and structured-light 3D geometry from porcine cadaver procedures for evaluating non-rigid surgical scene reconstruction. It includes diverse deformation sequences, precise camera registration, and rich geometric and semantic annotations.
Access:	Publicly accessible at DOI below, or via `https://reubendocea.github.io/d4d/`
Volume:	448GB
Variety:	Surgical data with multimodal annotations
DOI:	`doi.org/10.25532/OPARA-1033`
GD2	
Name:	Castelloli circuit Video Data (24 cameras) (currently not public)
Description:	Videos from the 24 cameras of the Castelloli circuit that capture 3 vehicles from the project going around the circuit (1h30min) for subsequent analysis (24 video files).
Access:	Data will be made accessible to project partners up request
Volume:	Currently $\sim$ 7GB
Variety:	
DOI:	NA

The data produced by CloudSkin will serve not only the current and forthcoming cohorts of researchers in big data and cloud technologies but also prove invaluable to big data practitioners and companies (from SMEs to multinationals). These entities, keenly interested in novel programming models for data analytics, stand to benefit significantly from the insights gleaned from CloudSkin's data.

3 FAIR data

In general terms, research data should be **FAIR**, that is **findable**, **accessible**, **interoperable** and **reusable** [4]. These principles precede implementation choices and do not necessarily suggest any specific technology, standard, or implementation/solution.

We follow the Horizon Europe DMP template [5], that is inspired by FAIR as a general concept. In the following sections, we try to answer the template questions in an appropriate level of detail. As the implementation of the project progresses, we have updated this document with information on a finer level of granularity.

3.1 Making data findable, including provisions for metadata

Used data. In order to ensure that the data used in the project is easily findable, we have made an effort to include standard identification mechanisms in all our publications, source code and tutorials. Although not all datasets used in the project provide these identification mechanisms, we have taken special care to provide the necessary instructions, metadata and tools for locating and processing those datasets.

Produced data. CloudSkin deposits generated data in an open online research data repository. We selected Zenodo as our data repository of choice. Zenodo is an OpenAIRE and CERN collaboration that allows researchers to deposit both publications and data, providing tools to link related items

through persistent identifiers and data citations. Zenodo automatically assigns a Digital Object Identifier (DOI) to each item to make them easily and uniquely citable. Moreover, Zenodo is set up to facilitate the finding, accessing, re-using and interoperating of data sets, which are the basic principles that ORD projects must comply with.

To this end, we created a CloudSkin community in Zenodo[1] to gather all the open data contributions of the project, as Figure 1 shows.

The repository allows to assign specific keywords to each dataset as well as a minimum of the DataCite's Metadata Schema [6] recommended terms.

Figure 1: CloudSkin community in Zenodo.

Whenever possible (according to publisher copyright policies regarding open access), research publications are uploaded to this repository to ensure the maximum dissemination of the results of the project. Publications are linked to its associated research data.

Source code. To make the source code open to the general public, we created a code repository in GitHub for CloudSkin[2] as Figure 2 illustrates.

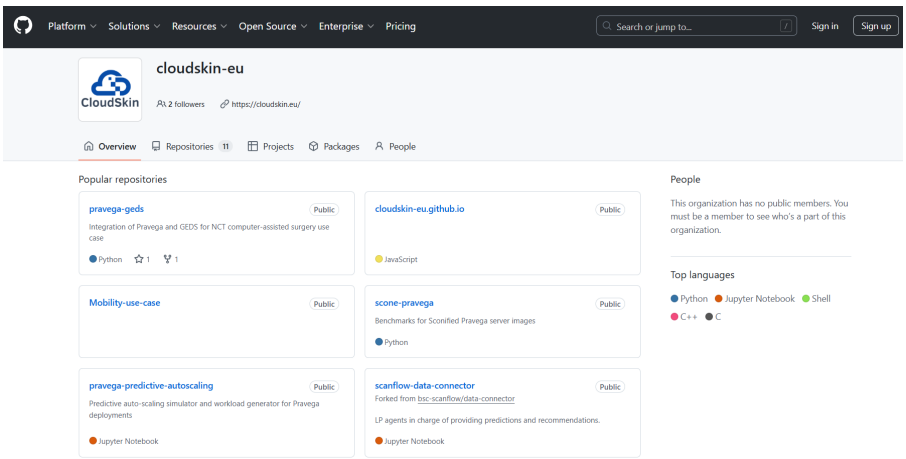

Figure 2: Public repository in Github for CloudSkin.

GitHub is currently one of the most popular code management systems due to the advanced fea-

[1] https://zenodo.org/communities/cloudskin-eu/
[2] http://github.com/cloudskin-eu

tures and easy management that it provides to developers. This has various potential benefits to the management and dissemination of CloudSkin source code: for instance, GitHub is well-known across developer communities, which facilitates the access to the source code of CloudSkin. Moreover, GitHub offers a plenty of options to fork/branch/merge versions of a software project that enables third-parties to easily extend the source code developed in CloudSkin (even for internal use). Additionally, we also made source code citable and uniquely identifiable by automatically archiving software releases in Zenodo [7].

As of the last release of this document, the CloudSkin Github profile contains 13 individual repositories hosting software results: 11 public and 2 private repositories.

Finally, the CloudSkin web page[3] lists all project results as Figure 3 illustrates. And its footer page provides links to their respective repositories in Zenodo or GitHub.

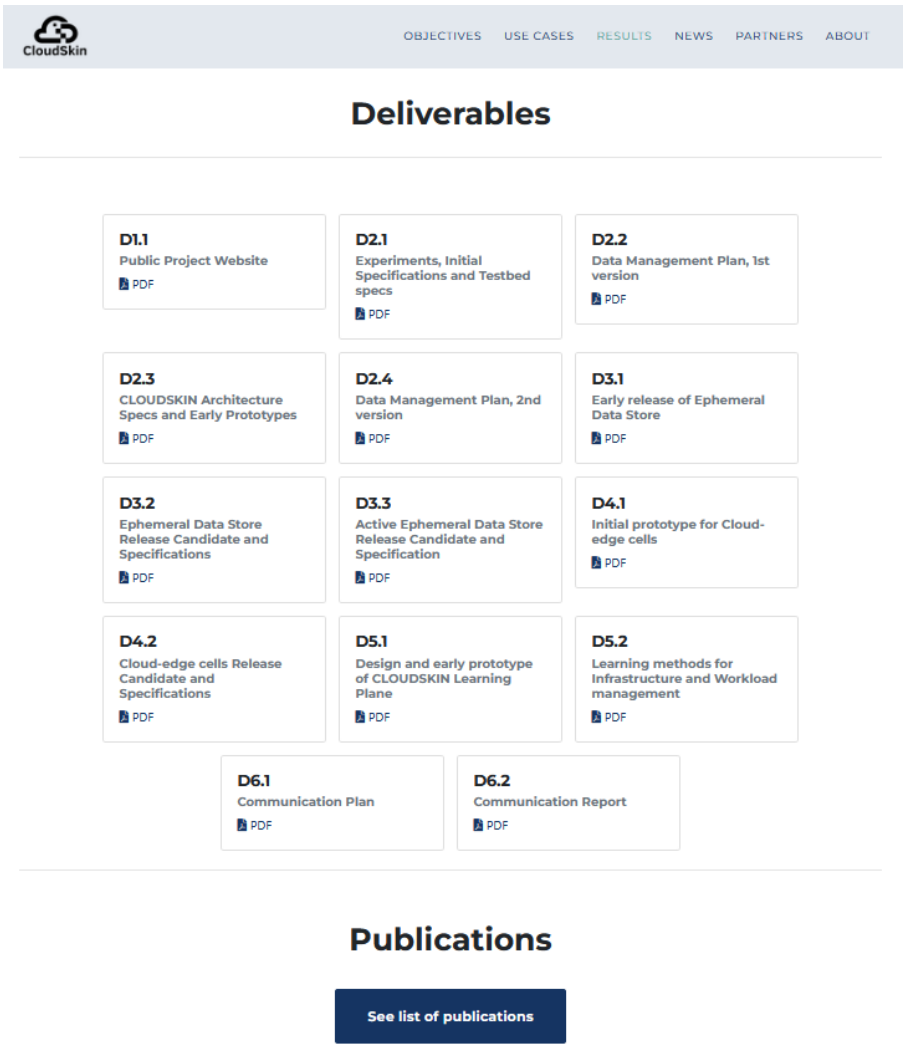

Figure 3: Project results on the CloudSkin website.

3.2 Making data accessible

It has been our intention that all data produced during the CloudSkin project is openly accessible as the default. Pre-existing datasets used in the experiments are mostly public and openly available (see Table 1).

Potential users will find out about the data through publications and the CloudSkin website. Data

[3] `http://cloudskin.eu`

will be made available on publication of the associated paper and will be made accessible through the Zenodo repository.

3.3 Making data interoperable

Interoperability of data produced within the CloudSkin project is promoted through best practices. Data formats should adhere to widely used standards and should be compliant with available software applications. Where possible, standard codes will be followed (e.g.: ISO 639 for language codes, ISO 3166 for country codes, NUTS for region codes, ...).

3.4 Increase data re-use

Data is made accessible, and therefore available for re-use, within one month of the publication of the related peer-reviewed scientific article. Data is shared under the Creative Commons Attribution 4.0 International Public License (CC BY 4.0) [8]. This license guarantees the widest possible re-use and redistribution while only requiring that appropriate credit is given.

As CloudSkin delegates the archiving of data to Zenodo, their policies regarding data maintenance apply. The data is stored in CERN Data Center. CERN has a commitment to maintain this data centre over the next 20 years. In the highly unlikely event that Zenodo will have to close operations, CERN guarantees that they will migrate all content to other suitable repositories, and since all uploads have DOIs, all citations and links to Zenodo resources (such as CloudSkin data) will not be affected.

The shared data will remain re-usable after the end of the project by anyone interested in it, with no access or time restrictions.

As the project progresses and data is identified and collected, further information on making data re-usable will be outlined in subsequent versions of the DMP. In specific, information about data quality assurance processes.

3.5 Management principles

The protocol below summarizes the management principles behind making generated research data FAIR:

PROTOCOL: Storing generated research data in CloudSkin project and making it FAIR

Beneficiaries follow these procedures for each dataset collected or generated during the CloudSkin project:

- Store and make findable the dataset in the CloudSkin community of the Zenodo repository.
- Ensure that publications and research data behind them are cross-referencing each other through standard identification mechanisms.
- Ensure that each dataset provides metadata, particularly regarding access rights, licenses, and funding information.
- Each Work Package Leader is responsible for storing relevant research data to the repository.
- Data will be made accessible within one month of the publication of the related peer-reviewed scientific article.

Beneficiaries will follow these procedures for source code generated during the CloudSkin project:

- Store the source code under the CloudSkin organization in GitHub repository.
- Provide a comprehensive README file with instructions to run the code.
- Store each release of the source code to Zenodo repository and cross-reference related datasets and publications.
- Each Work Package Leader is responsible for storing relevant source code to the repository.

3.6 Open Access Guideline.

Under Horizon Europe, Open Science has become a mandatory requirement. This entails that both publications and research data must comply with Open Access requirements. As publishing in open access may raise certain questions or uncertainties, a guideline document was developed to support project partners in following the correct procedures. It is a comprehensive guide to support the proper publication of research data. This guide has been presented and distributed among the consortium partners to summarize and clarify the most relevant concepts, thereby assisting them in meeting all Open Access requirements effectively. Figure 4 shows some slides of this guideline.

4 Allocation of resources

Regarding Open Access to research data, archiving at Zenodo is free of charge. Storing source code at the GitHub repository is also free of charge. Therefore, no costs are currently foreseen regarding the long term preservation of data.

URV provides its infrastructure to host the project web site (http://cloudskin.eu), and commits to keep the web site active after the end of the project.

The project coordinator has the ultimate responsibility for the data management in the project.

5 Data security

As CloudSkin delegates the archiving of data to Zenodo, their policies regarding data security apply:

- **Replicas**: All data files are stored in CERN Data Centres, primarily Geneva, with replicas in Budapest. Data files are kept in multiple replicas in a distributed file system, which is backed up to tape on a nightly basis.

Figure 4: Screenshots of the Open Access guideline.

- **Retention period**: Items will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least.
- **File preservation**: Data files and metadata are backed up nightly and replicated into multiple copies in the online system.
- **Fixity and authenticity**: All data files are stored along with a MD5 checksum of the file content. Files are regularly checked against their checksums to assure that file content remains constant.
- **Succession plans**: In case of closure of the repository, best efforts will be made by CERN to integrate all content into suitable alternative institutional and/or subject based repositories.

6 Ethics

There is no sensitive ethical issue of collecting, storing, processing and archiving data raised by the research of the CloudSkin project. Any potential ethical issue raised during the life of the project may be reported to the CloudSkin project board, which would, if necessary, raise immediate awareness of internal consortium members' executives, in order to take appropriate actions to resolve this issue.

Concerning potential ethical conflicts all issues will be resolved through the procedures depicted in relative legal documents (e.g., Consortium Agreement) and Commission guidelines.

7 Data Management Plan review process and timetable

As a *living* document, the Data Management Plan has been updated periodically. Particularly, the DMP has been updated whenever significant changes arise, such as:

1. New data
2. Changes in consortium policies (e.g. new innovation potential, decision to file for a patent)
3. Changes in consortium composition and other external factors (e.g. new member joining or current member leaving)

An up-to-date version has been available in time with each periodic review of the project. Table 3 summarizes the scheduled updates of the Data Management Plan.

Table 3: Timetable for Data Management Plan updates

Deliverable title	**Del. No.**	**Month**	**Date**
Data Management Plan, 1st version	D2.2	M6	June 2023
Data Management Plan, 2nd version	D2.4	M18	June 2024
Data Management Plan, 3rd version	D2.6	M36	December 2025

8 Conclusions

This document is the third version of the CloudSkin Data Management Plan. Its objective is to provide a clearly defined management strategy of data for the consortium. The CloudSkin DMP explains the life-cycle for the project data, whether it is collected, processed, or generated by the consortium. This DMP describes how the research data has been, and will be made findable, accessible, interoperable and reusable.

With this last version of the DMP, it is expected to present all the additional datasets that have been generated during the project, alongside plans for further management of test data and generated source code. In addition, issues like data quality assurance or data/metadata vocabularies for interoperability have been tackled in order to provide a refined version of the Data Management Plan.

References

[1] M. Carstens, F. M. Rinner, S. Bodenstedt, A. C. Jenke, J. Weitz, M. Distler, S. Speidel, and F. R. Kolbinger, "The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science," Scientific Data, vol. 10, no. 1, pp. 1–8, 2023.

[2] M. Wagner, B.-P. Müller-Stich, A. Kisilenko, D. Tran, P. Heger, L. Mündermann, D. M. Lubotsky, B. Müller, T. Davitashvili, M. Capek, et al., "Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark," Medical Image Analysis, vol. 86, p. 102770, 2023.

[3] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," IEEE transactions on medical imaging, vol. 36, no. 1, pp. 86–97, 2016.

[4] M. Wilkinson and et al, "The FAIR Guiding Principles for scientific data management and stewardship," Nature Scientific Data, no. 160018, 2016.

[5] European Commission, "Data management plan template," 2021.

[6] DataCite Metadata Working Group, "DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.2. DataCite e.V.." https://doi.org/10.5438/bmjt-bx77, 2019.

[7] GitHub, "Making your code citable." https://guides.github.com/activities/citable-code/, 2016.

[8] Creative Commons, "Creative Commons Attribution 4.0 International Public License." https://creativecommons.org/licenses/by/4.0/legalcode, 2013.