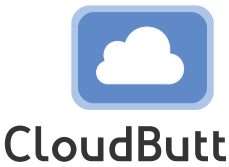

HORIZON 2020 FRAMEWORK PROGRAMME

CloudButton

(grant agreement No 825184)

Serverless Data Analytics Platform

D2.4 Data Management Plan, 2nd version

Due date of deliverable: 30-06-2020
Actual submission date: 31-07-2020

Start date of project: 01-01-2019 Duration: 36 months

Summary of the document

Document Type	ORDP: Open Research Data Pilot
Dissemination level	Public
State	v1.0
Number of pages	14
WP/Task related to this document	WP2 / T2.1
WP/Task responsible	URV
Leader	Gerard París (URV)
Technical Manager	Ana Juan (ATOS)
Quality Manager	Josep Sampé (URV)
Author(s)	Gerard París (URV)
Partner(s) Contributing	URV
Document ID	CloudButton_D2.4_Public.pdf
Abstract	Second version of the data management plan. The different experiments, workloads, benchmarks, and results will be delivered as Open Research Data for the community. This deliverable will evolve during the lifetime of the project in order to present the status of the project's reflections on data management.
Keywords	Data Management Plan, Open Access, Open Research Data, FAIR data, ORDP.

History of changes

Version	Date	Author	Summary of changes
0.1	05-06-2020	Gerard París	First draft
0.2	30-06-2020	Gerard París	Updated generated datasets.
0.3	07-07-2020	Gerard París	Integrated comments from internal reviewers.
0.4	07-07-2020	Gerard París	Updates used datasets with specific METASPACE datasets that are made available for the project.
1.0	31-07-2020	Gerard París	Final version.

Table of Contents

List of Abbreviations and Acronyms

AEMET	Agencia Estatal de Meteorología (*State Meteorological Agency (Spain)*)
API	Application Programming Interface
CC	Creative Commons
CERN	European Organization for Nuclear Research
CNIG	Centro Nacional de Información Geográfica (*National Center for Geographic Information Systems (Spain)*)
CSV	Comma-separated values
DMP	Data Management Plan
DOI	Digital Object Identifier
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ESA	European Space Agency
FAANG	Functional Anotation of ANimal Genomes
FAIR	Findable, Accessible, Interoperable and Reusable
GIS	Geographic information system
ICGC	International Cancer Genome Consortium
IGN	Instituto Geográfico Nacional (*National Geographic Institute (Spain)*)
ISO	International Organization for Standardization
LiDAR	Light detection and ranging
NUTS	Nomenclature des unités territoriales statistiques (*Nomenclature of Territorial Units for Statistics*)
ORDP	Open Research Data Pilot
SIAM	Sistema de Información Agraria de Murcia (*Murcia Agricultural Information System*)
SIGPAC	Sistema de Información Geográfica de parcelas agrícolas (*Spanish Land-parcel identification system*)
SME	Small and medium-sized enterprises
TCGA	The Cancer Genome Atlas
TIFF	Tagged Image File Format
URV	Universitat Rovira i Virgili

1 Executive summary

CloudButton is committed to good data management. In an effort to provide a management life-cycle of the data needed to validate results in scientific publication, a second version of the Data Management Plan (DMP) has been provided as deliverable D2.4. This DMP describes how the research data will be made findable, accessible, interoperable and reusable. This second version also presents a summary of the existing datasets that are currently known to be used over the course of the project. As the project progresses and data is identified and collected, further information about generated data will be provided.

2 Data Summary

The Open Research Data Pilot aims to enable open access and reuse of the research data generated by Horizon 2020 projects. CloudButton, as an action participating in the Open Research Data Pilot, has the commitment to:

- Develop (and keep up-to-date) a Data Management Plan (DMP).
- Deposit the project's data in a research data repository.
- Ensure third parties can freely access, mine, exploit, reproduce and disseminate our data.
- Provide related information and identify (or provide) the tools needed to use the raw data to validate our research.

In particular, the Open Research Data Pilot applies to:

- The data (and associated metadata) needed 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 main goal of the CloudButton project is to create a Serverless Data Analytics Platform. CloudButton aims to "democratize big data" by overly simplifying the overall life cycle and programming model thanks to serverless technologies. To demonstrate the impact of the project, we target two settings with large data volumes: bioinformatics (genomics, metabolomics) and geospatial data (LiDAR, satellital). This ambitious objective requires of a consistent evaluation of its productivity and performance. Therefore, in order to validate the results of the project, we are processing several existing datasets from the bioinformatics and geospatial domain. Additionally, CloudButton will make use of several relevant, industry-validated benchmarks to supply a complete and comprehensive evaluation of the project.

In addition to public datasets corresponding to general benchmarks (E-commerce Transaction Data, Wikipedia Entries, ...), the consortium will have access to extremely large and complex European and international domain-specific datasets hosted at EMBL-EBI, the Spanish National Geographic Institute (IGN) and the European Space Agency (ESA) amongst others. These datasets cover different data types and formats: structured/unstructured data, text, satellite images, LiDAR point clouds (LAZ format), mass spectrometry imaging data (imzML format), biological sequences (FASTQ format), ...

Table 1 presents a summary of the existing datasets that will be processed to validate the results of the CloudButton project.

Table 1: Used datasets

UD1	
Name:	Administrative areas [1]
Origin:	Spain's National Geographic Institute (CNIG-IGN)
Access:	Public
Volume:	31 MB for all Spain regions
Variety:	Shapefile format.
Frequency of update:	Fixed

UD2	
Name:	Sentinel-2 [2]
Origin:	European Comission
Access:	Public
Volume:	300 GB
Variety:	Raster graphics images (TIFF format)
Frequency of update:	Data must be downloaded with a frequency of between 15 days and a month.
UD3	
Name:	SIGPAC [3]
Origin:	Spanish Ministry of Agriculture, Fisheries and Food
Access:	Public
Volume:	200 GB
Variety:	Shapefile format
Frequency of update:	Fixed
UD4	
Name:	LiDAR [4]
Origin:	Spain's National Geographic Institute (CNIG-IGN)
Access:	Public
Volume:	8 TB
Variety:	LAS/LAZ files. LAS is the industry standard binary format for storing air LiDAR data. LAZ is a compressed data format often used to transfer large amounts of LiDAR data.
Frequency of update:	Fixed
UD5	
Name:	SIAM Meteorologic Information [5]
Origin:	Servicio de Información Agraria de Murcia (SIAM)
Access:	Public
Volume:	< 10 KB for each meteorological station and day
Variety:	CSV files
Frequency of update:	Data is updated every day. The experiments will use data for a range of dates.
UD6	
Name:	AEMET Meteorologic Information [6]
Origin:	Spanish Meteorological Agency (AEMET)
Access:	Public
Volume:	< 10 KB for each meteorological station and day
Variety:	CSV files
Frequency of update:	Data is updated every day. The experiments will use data for a range of dates.
UD7	

Name:	Irrigation communities [7]
Origin:	Irrigation communities of Murcia Region
Access:	Subject to the owner permission
Volume:	65 KB
Variety:	This information can be offered in different formats depending on each community. Most common format will be Shapefile.
Frequency of update:	Fixed
UD8	
Name:	Natura 2000 [8]
Origin:	European Environment Agency
Access:	Public
Volume:	80 MB for Spain
Variety:	OGC Geopackage
Frequency of update:	Fixed
UD9	
Name:	Functional Annotation of ANimal Genomes (FAANG) [9]
Origin:	FAANG consortium
Access:	Public
Volume:	5 TB
Variety:	FASTQ files. Dataset accessions: PRJEB26787, PRJEB19268, PRJEB24166, PRJEB28219, PRJEB19199, PRJEB25677, PRJEB23119, PRJEB27337, PRJEB28653, PRJEB23196, PRJEB19386, PRJEB27455, PRJEB25226, PRJEB24920. These datasets explore genomic and functional information on livestock species (Equus caballus, Sus scrofa, Bos taurus, Ovis aries, Bos indicus, Capra hircus, Gallus gallus, Bubalus bubalis).
Frequency of update:	Fixed
UD10	
Name:	Virus-host interaction [10]
Origin:	European Nucleotide Archive (ENA)
Access:	Public
Volume:	1 TB
Variety:	FASTQ/BAM files. Dataset accessions: ERP104372, ERP004390, SRP042295, SRP051574, SRP069043, SRP012102, SRP075180, SRP076509, SRP055968, SRP082191,
Frequency of update:	Fixed
UD11	
Name:	Cancers of the immune system [11]
Origin:	The Cancer Genome Atlas (TCGA) / International Cancer Genome Consortium (ICGC)
Access:	Private. Access can be obtained by principal investigators upon nominal request.

Volume:	3 TB
Variety:	FASTQ/BAM files
Frequency of update:	Fixed
UD12	
Name:	METASPACE public raw data [12]
Origin:	The METASPACE platform
Access:	Public/Private
Volume:	> 100 TB of raw data
Variety:	Datasets are provided in the imzML format, the main open format in the field of imaging mass spectrometry. EMBL provides a Python library to parse the data (`https://github.com/alexandrovteam/pyimzML`).
Frequency of update:	The METASPACE platform is growing 2x/year. In CloudButton, we will use representative pre-selected datasets that can be shared within the consortium (either public data or private data from EMBL team). The following datasets tagged with id UD12.X are made available specially for CloudButton.
UD12.1	
Name:	Brain02_Bregma1-42_02
Origin:	The METASPACE platform. Author: Régis Lavigne, University of Rennes 1
Access:	Public
Volume:	0.05 GB
Variety:	imzML format
Frequency of update:	Fixed
UD12.1	
Name:	Brain02_Bregma1-42_02
Origin:	The METASPACE platform. Author: Régis Lavigne, University of Rennes 1
Access:	Public
Volume:	0.05 GB
Variety:	imzML format
Frequency of update:	Fixed
UD12.2	
Name:	AZ_Rat_Brains
Origin:	The METASPACE platform. Author: Nicole Strittmatter, AstraZeneca
Access:	Public
Volume:	0.7 GB
Variety:	imzML format
Frequency of update:	Fixed
UD12.3	
Name:	CT26_xenograft
Origin:	The METASPACE platform. Author: Nicole Strittmatter, AstraZeneca
Access:	Public

Volume:	1.8 GB
Variety:	imzML format
Frequency of update:	Fixed
UD12.4	
Name:	Mouse brain test434x902 Captured with AP-SMALDI5 and Q Exactive HF Orbitrap
Origin:	The METASPACE platform. Author: Dhaka Bhandari, Justus-Liebig-University Giessen
Access:	Public
Volume:	3.9 GB
Variety:	imzML format
Frequency of update:	Fixed
UD12.5	
Name:	X089-Mousebrain_842x603 Captured with AP-SMALDI5 and Q Exactive HF Orbitrap
Origin:	The METASPACE platform. Author: Dhaka Bhandari, Justus-Liebig-University Giessen
Access:	Public
Volume:	7.0 GB
Variety:	imzML format
Frequency of update:	Fixed
UD12.6	
Name:	Microbial interaction slide
Origin:	The METASPACE platform. Author: Don Nguyen, European Molecular Biology Laboratory
Access:	Public
Volume:	56.7 GB
Variety:	imzML format
Frequency of update:	Fixed

Aside from these datasets and benchmarks, the CloudButton project will likely generate other data to validate the results presented in scientific publications (test data, APIs, source code used to perform analysis, documented Jupyter notebooks, captured performance results of benchmarking CloudButton toolkit, etc.). All this data will be made available as open data and its re-use will be encouraged. The expected size of this kind of data is relatively small, of the order of MB. As the project progresses and data is identified and collected, further information on data details will be provided. Table 2 presents a summary of the already generated datasets in the process of validating the results of the CloudButton project.

Table 2: Generated datasets

GD1	
Name:	CloudButton Serverless Benchmark results (May 2020)
Description:	Results of the CloudButton Serverless Benchmark. It includes plots with results of the Flops benchmark and the Storage benchmark.
Access:	Open Data
Volume:	2.2 MB
Variety:	Plots.
DOI:	https://doi.org/10.5281/zenodo.3923893

CloudButton data will not only be useful for the current and future generation of big data and cloud technologies researchers, but also big data practitioners and companies (from SMEs to multinationals) with a vested interest in new programming models for data analytics.

3 FAIR data

In general terms, research data should be **FAIR**, that is **findable**, **accessible**, **interoperable** and **reusable** [13]. These principles precede implementation choices and do not necessarily suggest any specific technology, standard, or implementation/solution.

Here, we follow the Horizon 2020 FAIR DMP template [14], 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 will update this document with information on a finer level of granularity.

3.1 Making data findable

Used data In order to ensure that the data used in the project is easily findable, we will make 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 will take special care to provide the necessary instructions, metadata and tools for locating and processing those datasets.

Produced data CloudButton is expected to deposit generated data in an open online research data repository. We have 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 have created a CloudButton community in Zenodo[1] to gather all the open data contributions of the project. The repository allows to assign specific keywords to each dataset as well as a minimum of the DataCite's Metadata Schema [15] recommended terms.

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

Source code. To make the source code open to the general public, we have created a code repository in GitHub for CloudButton[2]. GitHub is currently one of the most popular code management systems due to the advanced features and easy management that it provides to developers. This has various potential benefits to the management and dissemination of CloudButton source code: for instance,

[1] `https://zenodo.org/communities/cloudbutton`

[2] `https://github.com/cloudbutton`

GitHub is well-known across developer communities, which facilitates the access to the source code of CloudButton. 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 CloudButton (even for internal use). Additionally, we'll also make source code citable and uniquely identifiable by automatically archiving software releases in Zenodo [16].

Finally, the CloudButton web page[3] will list all project results and provide links to their respective repositories in Zenodo or GitHub.

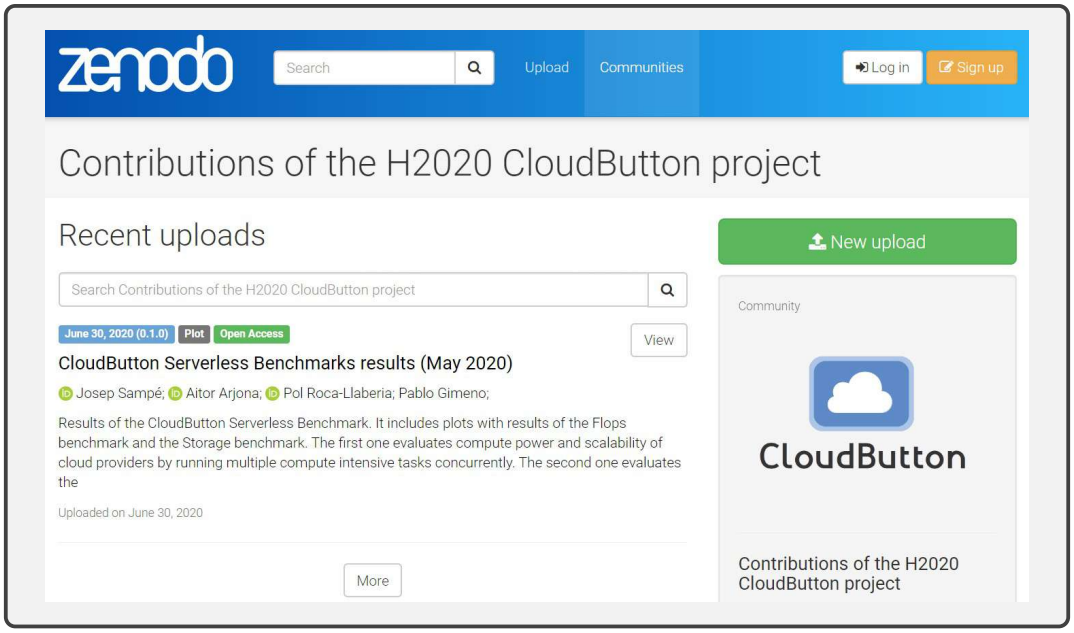

Figure 1: CloudButton community in Zenodo

3.2 Making data openly accessible

It is our intention that all data produced during the CloudButton project is openly accessible as the default. Pre-existing datasets used in the experiments are mostly public and openly available (see Table 1). The only private dataset is the UD11: Inmune System Cancers dataset from the International Cancer Genome Consortium (ICGC). ICGC data is protected due to privacy concerns (essentially, knowing the data leads to potentially identifying the donors). CloudButton beneficiaries confirmed in the consortium agreement that the background, results and any data that is provided or made available between the consortium members shall not include personal data. As a consequence, the beneficiaries agreed to not share datasets that may include human related data that could relate to an identified or identifiable living individual. The Pirbright Institute has started the procedure to gain access to the ICGC dataset. According to the CloudButton consortium agreement and the ICGC constraints, the experiments using this dataset will be private and not shared with the rest of the consortium.

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

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

codes, ISO 3166 for country codes, NUTS for region codes, ...).

As the project progresses and data is identified and collected, further information on making data interoperable will be outlined in subsequent versions of the DMP. Specifically, information on data and metadata vocabularies, standards or methodology to follow to facilitate interoperability and whether the project uses standard vocabulary for all data types present to allow interdisciplinary interoperability.

3.4 Increase data re-use (through clarifying licenses)

Data will be made accessible, and therefore available for re-use, within one month of the publication of the related peer-reviewed scientific article. Data will be shared under the Creative Commons Attribution 4.0 International Public License (CC BY 4.0) [17]. This license guarantees the widest possible re-use and redistribution while only requiring that appropriate credit is given.

As CloudButton 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 CloudButton 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 CloudButton project and making it FAIR

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

- Store and make findable the dataset in the CloudButton 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 CloudButton project:

- Store the source code under the CloudButton 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.

4 Allocation of resources

Costs related to Open Access to research data in Horizon 2020 are eligible for reimbursement during the duration of the project under the conditions defined in the H2020 Grant Agreement. The budget of the project already allocates 2,000€ per partner for costs related to provide Open Access, particularly to scientific peer-reviewed publications.

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://cloudbutton.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 CloudButton 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.

- **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 Ethical aspects

There is no sensitive ethical issue of collecting, storing, processing and archiving data raised by the research of the CloudButton project. Any potential ethical issue raised during the life of the project may be reported to the CloudButton 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 will be updated periodically. Particularly, the DMP will be 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 will be 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 2019
Data Management Plan, 2nd version	D2.4	M18	June 2020
Data Management Plan, 3rd version	D2.6	M36	December 2021

8 Conclusions

This document is the second version of the CloudButton Data Management Plan. It presents the current status of reflection within the CloudButton consortium about the research data that will be used, collected or generated. This DMP describes how the research data will be made findable, accessible, interoperable and reusable.

The next and last version of the DMP is due in December 2021. It is expected to present all the additional datasets that will be 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 will be tackled in order to provide a refined version of the Data Management Plan.

References

[1] Centro Nacional de Información Geográfica (CNIG), "Centro de Descargas." http://centrodedescargas.cnig.es/CentroDescargas/.

[2] "Sentinel-2 satellite date – ESA." https://scihub.copernicus.eu/dhus/#/home.

[3] "European Data Portal: SIGPAC datasets." https://www.europeandataportal.eu/data/en/dataset?tags=SIGPAC.

[4] "Large LiDAR point cloud datasets produced by the National Plan of Aerial Orthophotography (PNOA)." http://pnoa.ign.es/productos_lidar.

[5] "Sistema de Información Agraria de Murcia." http://siam.imida.es.

[6] Agencia Estatal de Meteorología, "Open Data." https://opendata.aemet.es.

[7] Confederación Hidrográfica del Segura, "Irrigation communities." https://www.chsegura.es/chs/cuenca/infraestructuras/postrasvaseTajoSegura/distribucion.html.

[8] European Environment Agency, "Natura 2000 End 2018 - OGC Geopackage." https://www.eea.europa.eu/data-and-maps/data/natura-10/natura-2000-spatial-data/natura-2000-spatial-lite-1.

[9] Functional Anotation of ANimal Genomes (FAANG) project, "Data portal." http://data.faang.org/home.

[10] The European Bioinformatics Institute, "European Nucleotide Archive (ENA)." https://www.ebi.ac.uk/ena.

[11] International Cancer Genome Consortium, "Data portal." https://dcc.icgc.org/.

[12] Metaspace, "Metaspace Platform for metabolite annotation of imaging mass spectrometry data." https://metaspace2020.eu/.

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

[14] European Commission, "Guidelines on FAIR Data Management in Horizon 2020." http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, 2016.

[15] DataCite Metadata Working Group, "DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.2. DataCite e.V.." http://doi.org/10.5438/rv0g-av03, 2019.

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

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