



**A European infrastructure
for farmed animal genotype to phenotype research**

**Deliverable 3.1
Establish EuroFAANG Data Management Plan**

Grant agreement no°: 101094718

Due submission date

2023-06-30

Actual submission date

2023-06-30

Responsible author(s): Dr Peter W. Harrison. European Molecular Biology Laboratory.

Confidential: No

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094718. The content of this report reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it

DOCUMENT CONTROL SHEET

Deliverable name	Establish EuroFAANG Data Management Plan
Deliverable number	3.1
Partners providing input to this Deliverable	EMBL
Draft final version circulated by lead party to: On date	Coordinators and Work Package Leaders on 2023/06/28
Approved by (on date)	Coordinators and Work Package Leaders on 2023/06/29
Work package no	3
Dissemination level	Public

REVISION HISTORY

Version number	Version date	Document name	Lead partner
1.0	2023/06/28	20230628_EuroFAANG_Deliverable3.1_V1.docx	EMBL
1.1	2023/06/30	20230630_EuroFAANG_Deliverable3.1_V1.1.docx	EMBL

Changes with respect to the DoA (Description of Action)

None

Dissemination and uptake

This is a public deliverable.

Table of Contents

1. Summary	4
2. Core Report.....	4
3. Data Management Plan.....	4
3.1. Data Summary	6
3.2. FAIR data	7
3.2.1. Making data findable, including provisions for metadata	7
3.2.2. Making data openly accessible.....	7
3.2.3. Making data interoperable	8
3.2.4. Increase data re-use (through clarifying licences)	8
3.3. Allocation of resources.....	9
3.4. Data security.....	10
3.5. Ethical aspects	10
3.6. Other issues	10

1. Summary

The purpose of this deliverable was to establish the first version of the EuroFAANG Data Management Plan (DMP) that outlines the Research Infrastructures (RI) commitment to FAIR (Findable, Accessible, Interoperable and Reusable), secure, and ethical data management. This DMP was constructed in accordance with the FAIR Data Management guidance template of Horizon 2020 and in accordance with the best practices established as part of the EU Open Research Data Pilot.

The EuroFAANG research infrastructure for farmed animal genotype to phenotype (G2P) research in Europe is not generating data and research output as a funded activity. However, as an infrastructure, it supports the community generation of open data and research output for genotype to phenotype research. To this end, a key aspect of the proposal is the development of a data policy, supported by rich metadata and format standards, to ensure data generated by the infrastructure is open and FAIR, and explores required policies for the interaction and generation of data with industry. This Data Policy and access principles is due to be published in December 2023.

The EuroFAANG DMP outlines the continuing extension of existing EuroFAANG FAIR research data management practices for ongoing and future European based G2P research. Such projects validate and archive data through the FAANG/EuroFAANG Data Coordination Centre's Data Portal into the INSDC public archives at EMBL (<https://data.faang.org/projects/EuroFAANG>). This DMP therefore extends the existing DMPs and best practice for the ongoing Horizon 2020 EuroFAANG affiliated projects. The EuroFAANG DMP is a living document that will increase in granularity throughout the EuroFAANG concept development phase and future phases of the RI. The DMP will be updated periodically, and whenever significant changes to data management or data policy are developed.

2. Core Report

The below document is the archived first version of the EuroFAANG Research Infrastructure Data Management Plan that was also made available to the entire consortium. It is a living document that will be updated periodically.

3. Data Management Plan



A European infrastructure for farmed animal genotype to phenotype research

Data Management Plan

Grant agreement no°: 101094718

DMP Version 1.0

Last updated: 2023-06-28

Authored by: Dr Peter W Harrison. Genome Analysis Team Leader. European Molecular Biology Laboratory

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094718. The content of this document reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it

REVISION HISTORY

Version number	Version date	Change description	Lead institute
1.0	2023-06-28	First version of the Data Management plan covering FAIR, secure and ethical data management practice for EuroFAANG.	EMBL
1.1	2023-06-29	Updates from coordinators and work package leads. Addition to ethics section.	EMBL

3.1. Data Summary

The EuroFAANG infrastructure for farmed animal G2P research in Europe is not generating data and research output as a funded activity. However, as an infrastructure, it will be supporting the community generation of open data and research output for genotype to phenotype research. To this end, a key aspect of the RI is the development of a data policy, supported by rich metadata and format standards, to ensure all data generated by the infrastructure is open, ethical, and FAIR. All generated data will be collated in the context of existing EuroFAANG, FAANG and community datasets within the EuroFAANG Data Portal. EuroFAANG affiliated projects, European researchers and industry will continue to generate datasets during the RI concept development phase under EuroFAANG coordination. The EMBL Data Coordination Centre will ensure the reusability of generated data and research outputs by providing rich supporting metadata, detailed mandatory protocols of research and analysis methods, links to the open access analysis software and parameters that generated the data, and clear provenance and licensing. For software, EuroFAANG members contribute heavily to the nf-core community, that supports community collaboration on the development of a curated set of reusable and openly licensed analysis pipelines.

All data generated by the EuroFAANG community will have internationally recognised identifiers from the International Nucleotide Sequence Database Collaboration (INSDC). These will be issued upon submission to the EMBL public archives. All data will be submitted through a brokered submission system that will ensure compliance with FAANG/EuroFAANG metadata standards and the data format standards that it develops in collaboration with Elixir. The EuroFAANG data portal will hold all data within the trusted public archives of EMBL that are part of the Elixir Infrastructure (<https://elixir-europe.org/platforms/data/core-data-resources>). Data will follow the FAANG data sharing principles (<https://www.faang.org/data-share-principle>) that promotes open science practices, pre-publication data-sharing, collaboration, and data reuse for benefit to the community and acceleration of research. All of the raw data will be released pre-publication under Fort Lauderdale (<https://www.genome.gov/Pages/Research/WellcomeReport0303.pdf>) and Toronto (<https://www.nature.com/articles/461168a>) principles to provide maximum benefit to the community. EuroFAANG will enhance and extend existing FAANG and Elixir metadata standards and ontology vocabularies (<https://data.faang.org/ruleset/samples>). Interoperability of generated data will be ensured so that data can be effectively utilised computationally through the EuroFAANG data portal application programming interfaces.

The DMP will be updated periodically, and whenever significant changes to data management or data policy are developed.

3.2. FAIR data

3.2.1. Making data findable, including provisions for metadata

The deposition of EuroFAANG community data, supported by the EuroFAANG RI, to the EMBL-EBI public archives will ensure the generated data is highly discoverable. EuroFAANG utilises and enhances the rich global FAANG metadata standards (<https://data.faang.org/ruleset/samples#standard>). All data submissions will be validated through the FAANG validation and submission tools (<https://data.faang.org/validation/samples>). The deposition in the public archives gives every data file a unique accession. These accessions are globally recognised by the comparable archives at the National Center for Biotechnology Information (NCBI; <https://www.ncbi.nlm.nih.gov/>) and DNA Databank of Japan (DDBJ; <https://www.ddbj.nig.ac.jp/index-e.html/>). Different assay files are linked through the inclusion of the BioSamples identifier in all data submissions so that all the datasets generated on each sample can be easily grouped and accessed from downstream presentation resources. EuroFAANG will conform with the FAANG record naming conventions. The FAANG data portal utilises Data Warehouse technology to ensure that all ontology validated metadata fields are keyword searchable using its predictive search. The data portal utilises the rich ontology supported metadata to provide filters that allow a user to explore data based on species, technology, breeds, sex, material, organism part, cell type, assay type, archive, and sequencing instrument. The portal allows for EuroFAANG data to be placed in the context of global FAANG data and a range of external legacy datasets.

3.2.2. Making data openly accessible

All samples and omics data will be deposited in the EMBL-EBI public archives. These are widely recognised and approved repositories for the long-term storage of biological data and the deposition routes are established with the FAANG Data Coordination Centre (DCC), that itself is based at EMBL-EBI. Apart from the reserved right of first global publication stipulation set out in the FAANG Data Sharing statement (<https://www.faang.org/data-share-principle>), there are no restrictions on use of the data, and no data access committee is required. The following data sharing statement is available both via the websites and Application Programmatic Interfaces (machine readable) of the public archives and FAANG data portal.

"This study is part of the FAANG project, promoting rapid prepublication of data to support the research community. These data are released under Fort Lauderdale principles, as confirmed in the Toronto Statement (Toronto International Data Release Workshop. Birney et al. 2009. Pre-publication data sharing. Nature 461:168-170). Any use of this dataset must abide by the FAANG data sharing principles. Data producers reserve the right to make the first publication of a global analysis of this data. If you are unsure if you are allowed to publish on this dataset, please contact the FAANG Data Coordination Centre and FAANG consortium (email faang-dcc@ebi.ac.uk and cc_faang@iastate.edu) to enquire. The full guidelines can be found at <http://www.faang.org/data-share-principle>."

The FAANG data portal collates the files from the various underlying archives to a single access point. The FAANG API provides programmatic users with the access FTP addresses to make a

secondary call to download the data files themselves. Following coordinated developments between EuroFAANG partners, software that is built using NextFlow will be made available to the community through the nf-core initiative (<https://nf-co.re/>), a community effort to collect a curated set of pipelines built using NextFlow. No specific tools are required to access the data from the data portals or the FAANG data portal, as they will use standard accepted file formats of the public archives.

3.2.3. Making data interoperable

EuroFAANG community data will be submitted through the FAANG DCC that will ensure the data is interoperable with other global FAANG datasets and highly reusable by the wider livestock community. To ensure interoperability with all other FAANG datasets, they are all validated to conform to FAANG metadata standards (<https://data.faang.org/ruleset/samples#standard>).

To ensure interdisciplinary interoperability EuroFAANG will utilise the recommended ontologies of the FAANG metadata standards as set by the FAANG metaFAIR Committee (<https://www.faang.org/faq?name=metaFAIR>). Wherever an ontology is not possible EMBL will employ controlled lists to prevent erroneous metadata recording. The ontologies that will be utilised in the data recording includes:

OBI	https://www.ebi.ac.uk/ols/ontologies/obi
NCBI Taxonomy	https://www.ebi.ac.uk/ols/ontologies/nbitaxon
EFO	https://www.ebi.ac.uk/ols/ontologies/efo
LBO	https://www.ebi.ac.uk/ols/ontologies/lbo
PATO	https://www.ebi.ac.uk/ols/ontologies/pato
VT	https://www.ebi.ac.uk/ols/ontologies/vt
ATOL	https://www.ebi.ac.uk/ols/ontologies/atol
EOL	https://www.ebi.ac.uk/ols/ontologies/eol
UBERON	https://www.ebi.ac.uk/ols/ontologies/uberon
CL	https://www.ebi.ac.uk/ols/ontologies/cl
BTO	https://www.ebi.ac.uk/ols/ontologies/bto
CLO	https://www.ebi.ac.uk/ols/ontologies/clo
SO	https://www.ebi.ac.uk/ols/ontologies/so
GO	https://www.ebi.ac.uk/ols/ontologies/go
NCIT	https://www.ebi.ac.uk/ols/ontologies/ncit
CHEBI	https://www.ebi.ac.uk/ols/ontologies/chebi

3.2.4. Increase data re-use (through clarifying licences)

EuroFAANG community data will be publicly released in the EMBL-EBI archives at the earliest opportunity and for raw data pre-publication. This will be submitted to the archives without embargo so that it is immediately released to the public. This is in accordance with the FAANG data sharing principles (<https://www.faang.org/data-share-principle>), that is based upon the principles of the Toronto (<https://www.nature.com/articles/461168a>) and Fort Lauderdale (<https://www.genome.gov/Pages/Research/WellcomeReport0303.pdf>) agreements. This reserves the right for submitters to make the first global publication with the data, and whether a dataset has an associated publication is tracked clearly in the FAANG data portal

(<https://data.faang.org/home>). All datasets will be clearly labelled with these data sharing principles, with the following statement:

"This study is part of the FAANG project, promoting rapid prepublication of data to support the research community. These data are released under Fort Lauderdale principles, as confirmed in the Toronto Statement (Toronto International Data Release Workshop. Birney et al. 2009. Pre-publication data sharing. Nature 461:168-170). Any use of this dataset must abide by the FAANG data sharing principles. Data producers reserve the right to make the first publication of a global analysis of this data. If you are unsure if you are allowed to publish on this dataset, please contact the FAANG Data Coordination Centre and FAANG consortium (email faang-dcc@ebi.ac.uk and cc_faang@iastate.edu) to enquire. The full guidelines can be found at <http://www.faang.org/data-share-principle>."

This enables the wider community to immediately make use of the data that the EuroFAANG community produces to provide maximal value to researchers. All software developed by the consortium will be openly licensed for reuse, with the license file displayed in the root folder of all repositories.

All EuroFAANG community data will be assessed with the latest guidelines on quality assurance, comply with directives of the public archives and with any quality guidance from the FAANG coordinated action. Through the accurate recording of metadata, associated protocols and analysis software, and deposition in public archives that the data will remain available for long after the project grant ends, for the lifetime of the underlying public archives. The data will therefore be reusable by any party, at some point the datasets may be superseded by those produced on newer technologies. There will be no restriction on third party use of the data.

3.3. Allocation of resources

EMBL is responsible for the curation, storage, and preservation costs as per its remit in providing the Elixir BioSamples and European Nucleotide Archives. These archives ensure long term preservation and assurance of data beyond the availability of any community specific portals and data services. The activity of the FAANG Data Coordination centre (DCC) to conduct data management and coordination are not funded by the EuroFAANG RI project during the concept development phase. Future allocation of resources and funding for data management and coordination will be developed as part of the business case development for the RI, alongside developing concepts of transnational access to European services. The DCC will ensure that data generated under the EuroFAANG RI umbrella will conform to FAIR data principles. This will include continued enhancement of existing FAANG metadata standards, expansion to new data types, archival support tools, data portal discovery and data visualisations to improve findability, accessibility, interoperability, and reusability of EuroFAANG data. These enhancements will also benefit the entire FAANG community as improvements will apply to all FAANG data. The costs associated with ensuring EuroFAANG data is FAIR have been fully accounted for. Data management is the responsibility of the FAANG Data Coordination Centre at EMBL-EBI that is led by Peter Harrison (Work Package 3 lead).

EuroFAANG research projects and the wider community will use the EMBL-EBI public archives for the long-term preservation of generated data, AND these resources have separate long term funding that will persist into the future. The inclusion of the data within the FAANG consortium data portal (<https://data.faang.org/home>) and Ensembl browser (<https://www.ensembl.org/index.html>) also ensures the functional annotation of genomes will

remain accessible by the community in the long term, as these are likely to continue to receive separate long term funding.

3.4. Data security

The concept for the EuroFAANG RI project will work on the premise that any data generated through EuroFAANG RI services will be submitted to EMBL-EBI public archives and catalogued by the FAANG data portal. The EMBL-EBI archives are internationally recognised repositories for the long-term secure storage of scientific data. All data will be assigned a unique identifier for long term identification and preservation of the datasets. The EMBL-EBI data centres that host the public archives providing the long-term data storage are state of the art. EMBL-EBI uses three discrete Tier III plus data centres in different geographical locations to ensure long-term security. Research data is also replicated through the International Nucleotide Sequence Database Collaboration (INSDC; <http://www.insdc.org/>) agreements that sees the data replicated at the National Center for Biotechnology Information (NCBI; <https://www.ncbi.nlm.nih.gov/>) and DNA Databank of Japan (DDBJ; <https://www.ddbj.nig.ac.jp/index-e.html>) that agree to recognise each other centres accessioned datasets. Each of the three INSDC resources agree to recognise the identifiers assigned by the other members, replicate the data presentations and act as a geographical mirror of all datasets.

EMBL-EBI commits to store the data for the lifetime that the archives remain active, ensuring this data remains available to the scientific community for years to come, this is part of EMBL-EBI's core commitment to ensure public scientific data remains available through its core data resource archives.

3.5. Ethical aspects

The EuroFAANG RI will comply with all international, EU and national level legal and ethical requirements for the relevant countries in which its partners operate. Any 'omics and phenotypic data handled by the RI will not be based on human subjects and thus informed consent is not required for data sharing or storing of personal data.

The EuroFAANG RI requires that the data have been generated in accordance with the regulations and guidelines for the use of animals for scientific purposes. This includes that the appropriate approval for the animal study in question has been granted by an appropriate ethics committee and/or regulatory body and that the studies take into account the ARRIVE Essential 10 (ARRIVE guidelines doi:10.1371/journal.pbio.3000411).

As part of the operations of the RI and for conducting surveys of European institutes, researchers, and industry representatives, the EuroFAANG RI will collect and store personal data in the form of web logs, survey results, interview results and email distribution lists. The EuroFAANG RI will fully comply with General Data Protection Regulations for its activities and web services, with GDPR statements and terms of use available and clearly displayed on the EuroFAANG website (<https://eurofaang.eu/>) and FAANG/EuroFAANG data portal (<https://data.faang.org/>).

3.6. Other issues

As well as complying with H2020 procedures for data management, the EuroFAANG RI will abide by the FAIR metadata standards (<https://data.faang.org/ruleset/samples#Standard>) and data sharing policy of the global Functional Annotation of Animal Genomes (FAANG) coordinated

action (<https://www.faang.org/data-share-principle>). This statement outlines the expectations of all FAANG projects that contribute to the coordinated action in terms of data recording, archiving, and sharing. The statement includes the principles of the Toronto (<https://www.nature.com/articles/461168a>) and Fort Lauderdale (<https://www.genome.gov/Pages/Research/WellcomeReport0303.pdf>) agreements. The requirements set out in the FAANG data sharing principles do not conflict with those imposed by the EU H2020 data management principles nor those we propose for the EuroFAANG RI.