A European infrastructure
for farmed animal genotype to phenotype research

Deliverable 3.3

Report on EuroFAANG and Elixir on shared information management strategy including an update of the DMP

Grant agreement no°: 101094718

Due submission date

2024-04-30

Actual submission date

2024-06-14

Responsible author(s):

Peter Harrison (EMBL)

Emily Clark (University of Edinburgh)

Martien Groenen (Wageningen University)

Sensitive: 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	Report on EuroFAANG and Elixir on shared information management strategy including an update of the DMP
Deliverable number	3.3
Partners providing input to this Deliverable	EMBL
Draft final version circulated by lead party to: On date	Coordinators on 2024-06-11
Approved by (on date)	Coordinators on 2024-06-14
Work package no	3
Dissemination level	PU

REVISION HISTORY

Version number	Version date	Document name	Lead partner
V1	2024-06-11	20240611_EuroFAANG_Deliverable_3.3v1.docx	EMBL
V2	2024-06-14	20240611_EuroFAANG_Deliverable_3.3_final.docx	EMBL

Changes with respect to the DoA (Description of Action)

None

Dissemination and uptake

This is a public deliverable

Table of Contents

1. Executive Summary

Background	This deliverable documents the EuroFAANG and Elixir Research Infrastructure shared information management strategy. The shared strategy is being developed and enacted through the recently established ELIXIR Domestic Animal Genome and Phenome focus group that was collectively conceived by the ELIXIR, EuroFAANG and Pheno-Live infrastructures. This newly established ELIXIR focus group, that aims to become a full ELIXIR community, gathers animal science, bioinformatics, and data management experts from across Europe to drive the development of standards, services, and training in the domestic animal genotype to phenotype domain.
Objectives	This report outlines the data flow and information systems of the respective domains of the EuroFAANG and ELIXIR Research Infrastructures. This clarifies the strategies and respective roles of both infrastructures in data management and standards for domestic animals and agriculture, and how the emerging EuroFAANG and Pheno-Live Research Infrastructures can position themselves within the already established ELIXIR landmark infrastructure data framework.
Methods	The structure of this report, focussing on the data flow and information management systems of the infrastructures, was modelled on the previous interaction between ELIXIR and EMPHASIS, a Research Infrastructure dedicated to plant phenomics. This provides a roadmap for how an agricultural research infrastructure can effectively integrate with the ELIXIR life sciences infrastructure. For the development of this report and for further developments over the coming years the key interaction forum is the recently established ELIXIR Domestic Animal Genomes and Phenomes focus group that includes members of the emerging EuroFAANG and Pheno-Live Research Infrastructures and the established landmark infrastructure ELIXIR.
Results & implications	The report on the Data Standards and Information systems strategies of the EuroFAANG and ELIXIR infrastructures documents the data flows, information systems and data standards of each infrastructure. It also highlights common tasks and areas for integration between the infrastructures and proposes next steps for this integration that will be managed through establishing working groups in the ELIXIR Domestic Animal Genomes and Phenomes focus group. The EuroFAANG Data Management Plan was also updated based on the

	findings of the report. This report text will serve as a basis of the collaboration strategy between the infrastructures, whilst the EuroFAANG and Pheno-Live Research Infrastructures continue to establish themselves in Europe and as part of EFSRI. If, and once, the Research Infrastructures are established, further avenues of collaboration and Memorandums of Understanding can be explored.

2. Data Standards and Information Systems: strategies of the European infrastructures EuroFAANG and ELIXIR

Main authors:
Peter Harrison (EuroFAANG RI, ELIXIR EMBL)
Emily Clark (EuroFAANG RI; ELIXIR UK)
Martien Groenen (EuroFAANG RI, ELIXR Netherlands)

1. Background and objectives

ELIXIR, the distributed infrastructure for life-science data, is a well-established ESFRI landmark infrastructure (a research infrastructure that reached its implementation phase) dedicated to life sciences data sharing and integration (Figure 1).
ELIXIR has three main strategic priorities to enable scientists to access and analyse life science data. These lay in the areas of Cellular and Molecular Research, Biodiversity, Food Security and Pathogens as well as Human Data and Translational Research. The latter includes the ambition to provide the infrastructure to support the discovery, access, sharing and analysis of genomics data and linked phenotypic or other data on a massice scale.
EuroFAANG is an emerging pan-European infrastructure for providing domestic animal genotype to phenotype transnational access to infrastructure services, funded by the European REA INFRA-DEV programme (2023-2025; Figure 2). The EuroFAANG Research Infrastructure fills a key gap identified in the last ESFRI roadmap, namely, “A gap regarding the focus on animals in agriculture and food sub-domain. A new infrastructure or an upgrade of existing efforts are needed at EU-level in the field of food, nutrition and processing”.

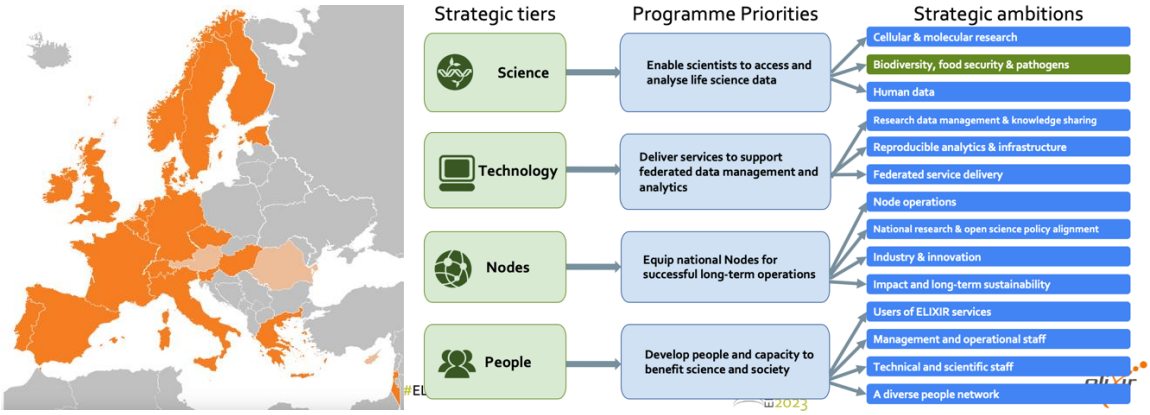

Figure 1. ELIXIR member states, strategic tiers, new programme priorities and strategic ambitions.

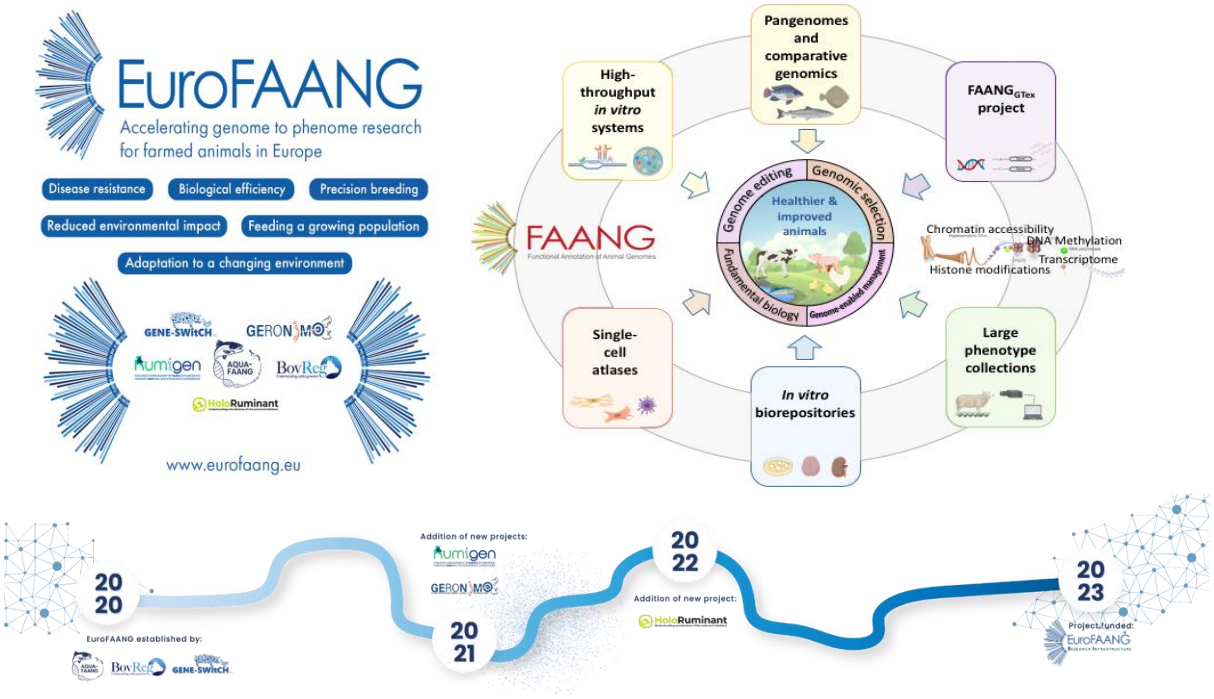

Figure 2. EuroFAANG Research Infrastructure objectives, emerging technology areas and timeline.

EuroFAANG, ELIXIR and Pheno-Live (another emerging European infrastructure focussed on animal phenomes, currently under consideration for EU REA INFRA-DEV funding) recently collectively established a focal point for animal standards and information systems as part of the ELIXIR focus group structures (https://elixir-europe.org/focus-groups). This newly established ELIXIR focus group "Domestic Animal Genomes and Phenomes", that aims to become a full ELIXIR community, focuses experts from across Europe to drive the development of standards, services, and training in the domestic animal genotype to phenotype domain. Together these infrastructures can generate readiness to meet the current challenges and demands in the Agri-Food sector in Europe, and ensure the data standards, information systems and transnational access to infrastructures are in place to meet national and international needs in food security going forward.

The purpose of this report is to outline the data flow and information systems of the respective domains of the Research Infrastructures. This clarifies the strategies and respective roles of both infrastructures in data management and standards for domestic animals and agriculture, and how the emerging EuroFAANG and Pheno-Live Research Infrastructures can position themselves within the already established ELIXIR landmark infrastructure framework. The FAANG Data Coordination Centre and FAANG infrastructure and data standards, upon which EuroFAANG RI infrastructure is based, has itself been developing and in active use for more than a decade. A key task of the focus group, and this report, therefore, is to identify commonalties and areas of future

alignment from this well-established platform to ELIXIR infrastructure and standards. Success in this area will also enable cross domain and community integration, with the wider communities in crop agriculture, biodiversity and animal health to the benefit of all.

The main objectives of the ELIXIR Domestic Animal Genome and Phenome focus group are to:

- Promote the Focus Group activities, seek membership, and scope future Community needs in different ELIXIR Nodes.
- Coordinate discussions and explore potential for data/technology solutions for addressing key issues in welfare, behaviour, health, infectious diseases, metabolism and nutritional efficiency and preservation of genetic diversity and environment.
- Develop data standards, coordination, workflows and visualisation for key developing areas of pangenomics, functional genomics, genome editing, phenotyping and biorepositories, which are needed to enhance data analyses.
- Develop FAIR (Findable, Accessible, Interoperable, Reusable) data management guidelines, standards and ontologies and promote best practices in data coordination and archiving.
- Organise tools and data training and knowledge transfer events.

The focus group in the coming year will clarify the strategies and respective roles of the infrastructures, and how EuroFAANG can integrate within the established ELIXIR information systems. This process will identify complementarities, common tasks, integrations and infrastructure gaps, and establish focus group working groups to address these areas for Domestic Animal Genomes and Phenomes. A number of these areas will be highlighted in this report and form the basis for the focus group working groups going forward and the potential for a future publication by the working group.

2. Data flow and respective domains of each infrastructure

Below in this report we define an outline of the 'domains of excellence' of each Research Infrastructure (and of their respective supporting communities) in order to better identify common tasks and points of future integration between the infrastructures.

The main domain of EuroFAANG concerning data is:

- Develop, promote and manage FAIR metadata 'omics data standards focussed on domestic animal genotype to phenotype data.
- Provide knowledge and tools (such as the Ontology Improvement Tool) for the improvement of ontologies to standardise, harmonise and enrich metadata recording.
- Establish a common data structure and access policy that ensures any data generated as part of the EuroFAANG infrastructure is coordinated, standardised

and FAIR. Coordinate the recording and archiving of data generated through transnational access to the infrastructure through pre-validated metadata and data submissions into appropriate INSDC public archives (including ELIXIR core data resources).

- Develop data and metadata standards for emerging technologies of importance in animal agriculture.
- Prototyping solutions to identified infrastructure gaps in animal agriculture genotype to phenotype space.
- Coordinate development of analysis pipelines and workflow managers for standardisation of 'omics analysis, in close coordination with the nf-core community in animal genomics.
- Facilitate access to a wealth of high-quality animal 'omics and phenome data and FAIR metadata through the EMBL-EBI hosted FAANG/EuroFAANG data portal, and the ELIXIR core data repositories resources of BioSamples, European Nucleotide Archive and Ensembl genomes browser.
- Promote open science and FAIR data principles to the scientific community and industry.

The main domain of the ELIXIR community concerning data is to:

- Enable FAIR publication of datasets to allow their use for genotype to phenotype research, and wider life science applications. ELIXIR as an infrastructure handles all types of data produced in life sciences: the whole spectrum from raw (e.g. images, time courses) to processed data (e.g. scalars representing traits).
- Enable data, tools and repositories interoperability by seeking collaboration with relevant communities to build and recommend standards, metadata and repositories.
- Allow findability and accessibility of any scientific data type, including genotype to phenotype 'omics data hosted by ELIXIR core resources and repositories. Therefore, a future EuroFAANG generated dataset could be found both through EuroFAANG Data Platform and a range of ELIXIR services hosting multidimensional genotype, phenotype and cross-referenced datasets.
- Help the building of integrative datasets that links phenotype to genotype or other cross referenced data types.
- Provide the infrastructure for the quality check of datasets, mainly at the syntactic level and to ensure the presence of minimal metadata like biological material description and complete measurement methodology traceability and provenance.
- Provide access to propose ELIXIR BioHackathon topics to prototype solutions to identified infrastructure gaps.

3. Information systems and data standards in each infrastructure

The EuroFAANG project itself isn't setting completely new data standards, instead, it focuses on aligning and extending a wealth of existing standards with the FAIR (Findable, Accessible, Interoperable, Reusable) principles for data management. EuroFAANG is extending more than ten years of global FAANG metadata standards, and it is key that these are extended to integrate with ELIXIR and the broader related domains such as crop agriculture and biodiversity. EuroFAANG emphasizes adherence to FAIR data management practice, and in particular defining clear metadata requirements to facilitate data discovery, access, and reusability. A key part that will be strengthened through the connection with ELIXIR is that EuroFAANG will involve the research community through workshops and consultations. This collaborative approach aims to develop additional standards for areas of emerging technology.

ELIXIR is organized as a distributed information system enabling FAIR principles across a federation of ELIXIR core data services and repositories. These data repositories cover the full range of datatypes of relevance to the EuroFAANG infrastructure, i.e. genomic, genetic, phenotyping and publications. Many of these ELIXIR core data resources are already employed by the EuroFAANG Research infrastructure, such as the European Nucleotide Archive and Ensembl Genomes Browser, but the integration and implementation could be further strengthened.

4. Common tasks identified between the infrastructures

Integration of EuroFAANG Data Coordination with ELIXIR Core Data resources and deposition databases

The EuroFAANG Data Coordination Centre is enhancing and extending the metadata standards that have developed over more than a decade under the global FAANG consortium. The commonality and expertise of the alignment with ELIXIR bring enhanced data richness, context and cross domain expertise. ELIXIR curates a vast collection of biological data across various species and domains. By integrating their farmed animal genomics data with relevant resources in ELIXIR, EuroFAANG can enrich their datasets with additional context and enable researchers to make more comprehensive analyses. For example, as part of EuroFAANG's data offering there is potential for a shared development of combining data from specific livestock breeds and aquatic species with ELIXIR's data on related species or relevant biological pathways. This cross-referencing can lead to deeper insights and novel discoveries within industry.

A key area is utilising ELIXIR expertise for further standardisation of data formats and interoperability. ELIXIR promotes the use of standardized data formats and interoperability best practices. By aligning their data with these standards, EuroFAANG can ensure their farmed animal genomics data seamlessly integrates with existing resources within ELIXIR. This eliminates hurdles in data exchange and facilitates

collaborative research efforts across different institutions. Researchers can easily combine EuroFAANG's data with other relevant datasets hosted by ELIXIR for more powerful analyses and becomes important for EuroFAANG's goals of transnational access with institutes that have not previously been involved in a FAANG project.

ELIXIR also provides EuroFAANG with a comprehensive toolbox of bioinformatics analysis tools and established workflows. This gives EuroFAANG researchers access to powerful resources for data analysis, visualization, and interpretation. Instead of developing their own tools from scratch, EuroFAANG can leverage ELIXIR's expertise and established infrastructure to analyse their data more efficiently and effectively. This becomes even more effective with the included collaboration with the nf-core animal genomics community that is connecting through the ELXIR focus group. A final area is enhancing data quality and trustworthiness for both academic and industry applications. By associating their data with ELIXIR's trusted resources, EuroFAANG can enhance the credibility and trustworthiness of their farmed animal genomics data within Europe, leading to wider adoption and use of the data by the research community.

Integrating Interoperability between information systems

EuroFAANG data interoperability will be integrated with offerings and solutions from the ELIXIR interoperability platform (https://elixir-europe.org/platforms/interoperability) including adapting the EuroFAANG data flow to utilise the FAIR Cookbook, RDM Kit and other FAIRification efforts from the latest Elixir programme and current ELIXIR communities. The Research Data Management Toolkit provides researchers with a one-stop shop for best practices and guidelines. The FAIR Cookbook (https://faircookbook.elixir-europe.org/content/home.html) provides a collection of "recipes" that researchers can follow to improve the FAIRness of their data. These recipes include specific steps, tools, and best practices. The cookbook also includes information on the benefits of FAIR data and the challenges of achieving FAIRness. The EuroFAANG Data Coordination Centre at EMBL will assess these offerings and how the existing and future data standards it develops can be further integrated and improved.

Data reuse will be further increased through the integration with ELIXIR core data resources and repositories, that will ensure data is seamlessly available from both the EuroFAANG Data portals and a range of ELIXIR supported solutions. ELIXIR endorsed data reuse solutions such as Research Object (RO) Crates and data integration with platforms such as Galaxy, will improve date reuse by a wider group of researchers. An RO-Crate bundles research data with its associated metadata, creating a self-contained package that facilitates sharing and future comprehension. The RO-Crate structure organizes data files alongside a machine-readable metadata document, typically in JavaScript Object Notation for Linked Data (JSON-LD) format, which comprehensively describes the data's origin, creation process, and any relevant contextual information. This standardized format promotes data transparency and enables seamless integration with data repositories and other research projects. The ELIXIR UK node, and in particular the team of Carole Goble at University of Manchester, will assist with this integration into EuroFAANG.

Aligning EuroFAANG FAIR Data standards with ELIXIR

Parallel to this, the FAANG/EuroFAANG data standards FAIRsharing.org records will be updated to take advantage of the latest developments of the platform. FAIRsharing.org is a community-recognized resource for searching and identifying data repositories relevant to specific research areas. By registering relevant datasets on FAIRsharing.org, EuroFAANG can significantly increase the visibility of their farmed animal genomics data for researchers worldwide. This can lead to wider use of the data and potential collaborations with other research groups. FAIRsharing.org goes beyond just listing data resources. It also provides information on data standards, access procedures, and usage conditions. This information is crucial for researchers who want to reuse EuroFAANG's data in their own projects. By adhering to FAIR data principles encouraged by FAIRsharing.org, EuroFAANG can ensure their data is more easily integrated with other datasets and facilitate collaborative research efforts. FAIRsharing.org often curates' data resources based on specific quality criteria. By registering their data resources on FAIRsharing.org, EuroFAANG can demonstrate their commitment to data quality and transparency. FAIRsharing.org also encourages data providers to document their data collection and processing methods. This documentation can improve the trustworthiness and credibility of EuroFAANG's data for the research community.

Specific recommendations from Elixir will be reviewed and incorporated, such as the recommendations of the FAIRification of genomic tracks, that EuroFAANG can implement through the representation of its data through Ensembl.

Integrated future training

A collaborative advanced training program between EuroFAANG's Research Infrastructure and ELIXIR could significantly boost expertise in handling and analysing large-scale farmed animal genomics data. EuroFAANG's infrastructure offers a wealth datasets, whilst ELIXIR's proficiency in bioinformatics tools and data management could provide a strong foundation for researchers to develop essential skills in data wrangling, analysis, and interpretation based on real world data examples. This joint program could encompass workshops, tutorials, and hackathons, fostering a collaborative environment where researchers can learn from EuroFAANG's data expertise and ELIXIR's bioinformatics knowledge. By combining these strengths, the program has the potential to equip researchers with the necessary tools and knowledge to unlock the full potential of big data in farmed animal genomics research.

Joint EuroFAANG ELIXIR BioHackathons

A EuroFAANG Research Infrastructure and ELIXIR joint BioHackathon could offer a valuable platform for collaborative research in farmed animal genomics. EuroFAANG's data infrastructure combining genetic and phenotypic data across various livestock species with ELIXIR's expertise in bioinformatics tools and data management, the

BioHackathon could facilitate significant progress in identified technology and data management gaps. This could include leveraging advancements in AI to data analysis to optimise breeding programmes. This collaborative effort has the potential to yield significant advancements in farmed animal genomics.

5. Next steps

1. The data standards and information systems outlined in this document will be further discussed at the ELIXIR Domestic Animal Genome and Phenome focus group, with respect to the objectives of the group and future community.
2. Infrastructure and data standards gaps and areas of integration between the infrastructures will be further explored in working groups that will be established as part of the ELIXIR Domestic Animal Genome and Phenome focus group and future community. These working groups will report back to the monthly focus group meetings (https://elixir-europe.org/focus-groups/domestic-animals-genome-phenome; First Monday of the month at 2pm CET/CEST), and to the executive groups of ELIXIR and EuroFAANG for enactment of any identified actions.
3. The connection and respective remits of the ELIXIR Domestic Animal Genome and Phenome focus group and the also recently established nf-core Animal Genomics community will be explored. The nf-core Animal Genomics community will regularly report on activity and actions as part of the ELIXIR Domestic Animal Genome and Phenome focus group monthly meetings.
4. After further amendments by the EuroFAANG executive steering group, ELIXIR leadership and the ELIXIR Domestic Animal Genome and Phenome focus group chairs, this text will serve as a basis of the collaboration strategy between the infrastructures, whilst the EuroFAANG and Pheno-Live Research Infrastructures continue to establish themselves in Europe and as part of EFSRI. If, and once, the Research Infrastructures are established, further avenues of collaboration and Memorandums of Understanding can be explored.
5. The ELIXIR Domestic Animal Genome and Phenome focus group are planning a whitepaper, and this document will form the basis of the data standards and information systems sections of the manuscript.
6. The tasks identified in this document, and as part of the ELIXIR Domestic Animal Genome and Phenome focus group will be enacted by the EMBL-EBI EuroFAANG Data Coordination centre as it integrates its data standards and infrastructure within the established wider ELIXIR life sciences processes.

3. Appendix 1 - Participants to the Elixir Domestic Animal Genome and Phenome kick off meeting held on 3rd June 2024

Alexey Sokolov (EMBL-EBI)
Endre Barta (MATE, HU)
Hervé Acloque (INRAE)
Sarah Fischer (FBN, NFDI4Biodiverstiy)
Elisabetta Giuffra (INRAE)
Daniel Fischer (LUKE)
Physilia (ELIXIR Hub)
Nina Melzer (FBN)
Cedric Notredame (CRG)
Björn Langer (CRG)
Jose Espinosa-Carrasco (CRG)
Geena Cartick (EFFAB)
Martijn Derks (WUR)
Ole Madsen (WUR)
Martien Groenen (WUR)
David MacHugh (UCD)
Temitope Ige (INRAE)
Sigbjørn Lien (NMBU)
Peter Maccallum (ELIXIR Hub)
Joseph Robertson (NMBU)
Michèle Boichard (INRAE)
Catherine Larzul (INRAE)
Klaus Wimmers (FBN)
Tina Hartwig (FBN)
Rene Baumont (INRAE)

Apologies:
John Hancock
Peter Harrison (EMBL-EBI)
Cagla Kaya (EFFAB)
Ana Granados(EFFAB)

4. Appendix 2 - Updated EuroFAANG Data Management Plan

A European infrastructure for farmed animal genotype to phenotype research

Data Management Plan

Grant agreement no°: 101094718

DMP Version 1.3

Last updated: 2024-06-11

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
1.2	2024-06-11	Updates based on deliverable 3.3 and the shared information management strategy with ELIXIR	EMBL

4.1. 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.

4.2. FAIR data

4.2.1. 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.

The FAANG/EuroFAANG data standards fairshairing.org records will be updated to take advantage of the latest developments of the platform. Specific recommendations from Elixir will be reviewed and incorporated, such as the recommendations of the FAIRification of genomic tracks, that EuroFAANG can implement through the representation of its data through Ensembl.

4.2.2. 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.

4.2.3. 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/tf?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/ncbitaxon
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

EuroFAANG data interoperability will be integrated with offerings and solutions from the ELIXIR interoperability platform (https://elixir-europe.org/platforms/interoperability) including adapting the EuroFAANG data flow to utilise the FAIR Cookbook, RDM Kit and other FAIRification efforts from the latest Elixir programme and current ELIXIR communities.

4.2.4. 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.

Data reuse will be further increased through the integration with ELIXIR core data resources and repositories, that will ensure data is seamlessly available from both the EuroFAANG Data portals and a range of ELIXIR supported solutions. ELIXIR endorsed data reuse solutions such as Research Object Crates and data integration with platforms such as Galaxy, will improve date reuse by a wider group of researchers.

4.3. 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.

4.4. 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-EBIs core commitment to ensure public scientific data remains available through its core data resource archives.

4.5. 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/).

4.6. 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.