A European infrastructure
for farmed animal genotype to phenotype research

Deliverable 3.5

Report on current data infrastructure gaps and proposals for required data infrastructure development including an update of the DMP

Grant agreement no°: 101094718
Due submission date
2025-June-30
Actual submission date
2025-July-01

Responsible author(s):

Emily Clark (EMBL-EBI)
Alexey Sokolov (EMBL-EBI)

eclark@ebi.ac.uk

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	Report on current data infrastructure gaps and proposals for required data infrastructure development including an update of the DMP
Deliverable number	3.5
Partners providing input to this Deliverable	EMBl-EBI, INRAE, WU, UEDIN
Draft final version circulated by lead party to: On date	FBN 2025-07-01
Approved by (on date)	FBN 2025-07-01
Work package no	3
Dissemination level	Public (PU)

REVISION HISTORY

Version number	Version date	Document name	Lead partner
Vs1	2025-06-30	D3.5_EuroFAANG_Infrastructure_Gaps_Updated_DMP_v1	EMBL-EBI
Vs2			
Vs3			
Vs4			
Vs5			

Changes with respect to the DoA (Description of Action)

None

Dissemination and uptake

This is a public deliverable

Table of Contents

1. Introduction

The purpose of this deliverable is to provide a report on the current data infrastructure gaps and proposals for required data infrastructure development including an update of the Data Management Plan (DMP).

Developing coordinated, secure and high-quality data management, structure, interoperability and integration between the EuroFAANG research infrastructure (EuroFAANG RI) and the EuroFAANG Data Coordination Centre (DCC) is key for its success. Our aim was to identify during EuroFAANG RI concept development to identify gaps in existing FAANG data structure and infrastructure that needed to be filled to support the developing EuroFAANG meta infrastructures (WP4-6), access requirements (WP2) and interoperability with other existing and developing infrastructures (WP7).

For this deliverable we explored gaps in the data infrastructure in 3 keys ways:

- By examining current operability and functionality issues affecting the current FAANG Data Portal (https://data.faang.org/home) and updating these.
- Participating in WP4, WP5 and WP6 working groups to identify gaps related to in-vitro systems, genome editing and industry stakeholder interaction with the portal.
- Engaging with the global FAANG community to identify wider community defined gaps in the existing infrastructure and associated training requirements.
- Understanding current and expected capacity of the Data Portal in terms of numbers of submitted datasets.
- Interacting with animal phenotyping researchers for development of the GenoPHEnix ESFRI proposal and updated DMP.
- Building the ELIXIR Domestic Animal Genomes and Phenomes Focus Group for WP7 and managing synergies between the two infrastructures for data and metadata submissions into data repositories such as ENA and BioSamples.

We also provide an update of the DMP (Deliverable 3.1), also incorporating the data access policy and principles (Deliverable 3.2). This update is based on discussions for the GenoPHEnix project to incorporate animal phenotyping data in the FAANG Data Portal infrastructure as a proposed infrastructure development.

2. Current Data Portal Infrastructure

The FAANG Data Portal infrastructure plays a key role for EuroFAANG by supporting high-quality functional annotation of animal genomes, through open FAIR sharing of data, complete with standardised rich metadata. As described in Harrison et al. 2021 [1]: *FAANG projects produce a standardised set of multi-omic assays with resulting data placed into a range of specialised open data archives. To ensure this data is easily findable and accessible by the community, the portal automatically identifies and collates all submitted FAANG data into a single easily searchable resource. The Data Portal supports direct download from the multiple underlying archives to* *enable seamless access to all FAANG data from within the portal itself. The portal provides a range of predefined filters, powerful predictive search, and a catalogue of sampling and analysis protocols and automatically identifies publications associated with any dataset. To ensure all FAANG data submissions are high-quality, the portal includes powerful contextual metadata validation and data submissions brokering to the underlying EMBL-EBI archives.*

The portal has flexibility to incorporate new technical infrastructure to effectively deliver new data types and technologies to best fit the needs of the domestic animal science community. In this deliverable we described the updates of the current infrastructure that we have undertaken during the current EuroFAANG RI project and pose suggestions as to how the infrastructure can be developed in the future.

3. Updates to FAANG Data Portal Infrastructure for EuroFAANG

Throughout the project we have been investigating current operability and functionality issues affecting the FAANG Data Portal and looking at ways to mitigate these. Based on community discussions around user accessibility and local knowledge of cloud based and programming software changes have made three key improvements to the FAANG Data Portal infrastructure to fill major gaps:

1. **Transitioning to cloud:**
   - The transition to cloud is essential for modernizing FAANG IT infrastructure, enabling greater scalability, flexibility, and cost efficiency. All core FAANG services have now been successfully migrated to the cloud, with stateless components, such as front-end and back-end deployed on GCP Cloud Run, and the database layer hosted on Elastic Cloud to ensure high availability and performance.
2. **Moving from Angular to Python Dash:**
   - Moving from Angular to Python Dash prioritizes rapid development and iteration for data-focused applications and significantly speeds up delivery by eliminating the need for a separate front-end stack. Shifting to Python Dash simplifies integration with our existing Python-based services. It reduces the need for context switching and allows for faster development of data-driven features by enabling a shared codebase between front-end visualizations and back-end logic.
3. **Adding new search query functionality:**
   - Elastic search enables fast and intelligent free-text search across our website, enhancing the user experience by allowing users to quickly find relevant information through natural language queries. Integrating it provides a scalable, high-performance search layer that supports advanced filtering and relevance ranking out of the box. Global search functionality deployed in production - https://data.faang.org/globalsearch.

These updates have significantly enhanced the functionality and inoperability of the FAANG Data Portal filling significant gaps in data infrastructure and user experience. They will be continously maintained and updated for the duration of the EuroFAANG RI project.

4. Current capacity of the FAANG Data Portal

Understanding the amount of data that the Data Portal for the EuroFAANG RI will be required to manage is important to determine whether there are any gaps in the infrastructure in terms of data ingestion capacity, either now or in the future.

To quantify the amount of data that the FAANG Data Portal would be expected to handle as an e-infrastructure, we have provided the following three examples; 1) For AquaFAANG**,** one of the H2020 EuroFAANG projects, the size of data produced, between 2021 and 2013, was 61.3 Tb in total which gives an example of the total output from a large research and innovation project that would need to be managed by the EuroFAANG RI; 2) At CIGENE, one of the organisations within EuroFAANG (via partner NMBU) the amount of data generated in 2021 and 2022 was 9Tb and 7Tb respectively which provides an approximation of the potential size of genomic data produced by each partner annually; 3) Table 1 provides an example of the number and size of the genomic datasets for farmed animals that were deposited in the European Nucleotide Archive (a core data service delivered and maintained by EMBL-EBI) in 2024.

Table 1: Datasets deposited in the ENA in 2024 for farmed animals

Scientific Name	Common Name	Tax ID	Total datasets for 2024	European datasets for 2024 (approximation)
Gallus gallus	chicken	9031	210	42
Sus scrofa	pig	9823	269	54
Bos taurus	domestic cattle	9913	262	52
Ovis aries	sheep	9940	150	30
Equus caballus	horse	9796	40	8
Capra hircus	goat	9925	136	27
Oncorhynchus mykiss	rainbow trout	8022	29	6
Salmo salar	Atlantic salmon	8030	25	5
Sparus aurata	gilthead seabream	8175	8	2
Bos indicus	zebu cattle	9915	13	3
Scophthalmus maximus	turbot	52904	10	2
Bubalus bubalis	water buffalo	89462	11	2
Oryctolagus cuniculus	rabbit	9986	28	6
Rangifer tarandus	reindeer	9870	2	0
Total			1193	239

In terms of computational capacity a large-scale genomics project such as AquaFAANG with ~60Tb would require 10 worker nodes (each with 8CPU and 32 GB of RAM) of computational capacity. The expected total annual submission of genomic data to the data services provided by the EuroFAANG RI would likely also total ~64Tb (assuming 8 partners within the EuroFAANG RI each produce 8Tb of genomic data annually).

The submission/validation tool, for samples, experiments and analysis which is currently available via the FAANG Data Portal (https://data.faang.org/validation/samples) is essential to make sure the data and metadata submissions are brokered smoothly to BioSamples and the European Nucleotide Archive. The Data Portal infrastructure can currently manage the level of ingestion of data described above but would need continued support for maintenance and development to ensure the portal infrastructure, including the submission and validation tools, scaled with amount of data being submitted. This would be an on-going process that would be monitored over the lifetime of the EuroFAANG RI, and any necessary infrastructure upgrades would be made proactively according to expected demand.

5. Gaps in the data infrastructure identified by interactions with other work packages and global FAANG

Through shared meetings with work packages 4, 5 and 6. Three key gaps in the current data infrastructure were identified:

1. Lack of suitable metadata fields and ontologies for in vitro systems (addressed in Deliverable 4.4)
2. Capacity to provide catalogues of single guide RNAs and other CRISPR libraries through the FAANG Data Portal (identified and described in Deliverable 5.1).
3. Ability to broker encrypted data through the FAANG Data Portal (identified and described in Deliverable 6.1)

These points are discussed in the specific deliverables they are relevant to. In brief:

- We plan to run a virtual workshop before the end of the project to update the metadata and ontologies for in vitro models building on the work we have already done with work package 4 to develop these. We also provided a metadata training session at the G2P in a Dish workshop to help facilitate data and metadata submissions by the research community working on in vitro systems for domestic animals.
- To provide the catalogues of CRISPR libraries we would aim to work with the Broad Institute and with networks of collaborator such as the Engineering Biology (EB2) Hub, at the Roslin Institute (UEDIN) by providing links from the FAANG Data Portal.
- Sharing of encrypted data is more complex due to considerations around FAIR and OPEN access to data as outlined in the EuroFAANG Data Portal but with work package 6 we are working with a colleague at UC Davis through the AG2PI project to determine how this might be possible. In general though encrypted data could be deposited and access via the Data Portal in the same way as standard datasets so would not require any significant infrastructure development.

In addition, we were also able to discuss gaps in infrastructure with the global FAANG community who identified a lack of training materials for metadata and data submissions as a significant gap in the current infrastructure. To address this we are making training videos based on the training we delivered at the G2P in a Dish workshop and will have these ready by the end of the project. We also plan to run a webinar to talk through the process of submission and make this available to accompany the videos. We will make the videos available via the Data Portal and can extend outreach by also making them available through the EMBL-EBI outreach and training video archives.

6. Expanding the FAANG Data Portal infrastructure for phenotyping data

During concept development for the EuroFAANG RI project a significant gap was identified in animal phenotyping. The ability to collect and utilize manual on-farm data for breeding purposes revolutionized the beef, dairy, poultry, and pig industries, has led to huge gains in productivity and efficiency and systematically collected phenotype data available for genetic evaluations are among the most extensive data types in livestock [3]. Similar opportunities for tremendous gains may be possible with the utilization of sensors and other high-throughput phenotyping technologies to improve farmed animal health, welfare and productivity [3]. However, there are several caveats to how phenotyping data is managed that are prohibitive to achieving these gains.

Phenotyping data is currently:

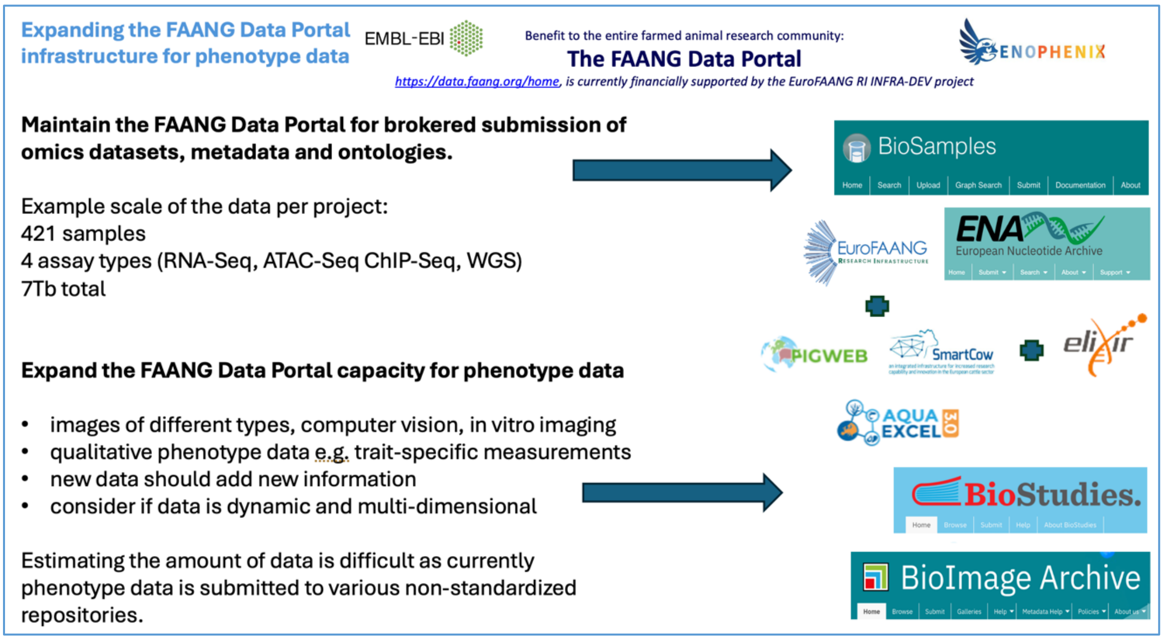
- Expensive to generate.
- Under-utilised.
- Often kept proprietary.
- Owned by farmers and breeding companies, not research institutions.
- Lacking rich metadata e.g. ICAR and Livestock Production Trait Ontology.
- Requires manual curation.

The FAANG Data Portal currently provides a highly standardised data submission and management system for omics data https://data.faang.org/home [1]. As such for genomic data, the data infrastructure is well established. However, this is not the case for phenotype data. An important part of the EU's ambition to be at the forefront of a data-driven society is open access to data produced at all scientific and commercial levels (https://tinyurl.com/yck5vcs8). Reusability of data and biospecimens (e.g. meat, milk, skin, blood, hair, faeces, urine, etc.) derived from animal experiments is crucial as it strongly supports the 3Rs and the economic efficiency of research. Despite these efforts, many issues (e.g. lack of standardised data including methodologies, SOPs and ontologies; fragmented information from farm animal biobanks and biosamples; lack of best practices and use cases for data sharing) are blocking progress towards data-driven animal agriculture [2]. Developing the concept of the EuroFAANG RI towards the integrated genome and phenome data structure proposed for GenoPHEnix will ensure that standardised data for phenotypes, biosamples and their metadata are machine-readable to improve their potential reusability, while applying common ontologies for interoperability. This consolidated effort underpins future data infrastructure development for farmed animals, and will require input from industry stakeholders, particularly in managing access to commercial data.

Phenotype data can be:

- Qualitative e.g. presence or absence of a behaviour, or a trait like horned or polled.
- Quantitative e.g. measurements of body size, carcass quality, feed efficiency etc.
- Image based e.g. computer vision, tracking animals to measure behavioral phenotypes.
- Sensor based e.g. collecting millions of data points per day for methane emission, feed intake, movement etc.
- These data types improve animal feeding, breeding and management.

Figure 1 shows how we would envisage submission of phenotype data to core EMBL-EBI ELIXIR database resources including BioStudies and the BioImage archive. This would require expanding the FAANG Data Portal to support multimodal submissions across data deposition resources. It builds on the foundation provided the EuroFAANG RI project for submission of omics data to the ENA and BioSamples and information on current management of phenotype data from the INFRAIA projects PigWeb (https://www.pigweb.eu), AQUAExcel 3.0 (https://aquaexcel.eu) and SmartCow (https://smartcow.eu).



Expanding the FAANG Data Portal infrastructure for phenotype data

Benefit to the entire farmed animal research community:

The FAANG Data Portal

https://data.faang.org/home, is currently financially supported by the EuroFAANG RI INFRA-DEV project

Maintain the FAANG Data Portal for brokered submission of omics datasets, metadata and ontologies.

Example scale of the data per project:
421 samples
4 assay types (RNA-Seq, ATAC-Seq ChIP-Seq, WGS)
7Tb total

Expand the FAANG Data Portal capacity for phenotype data

- images of different types, computer vision, in vitro imaging
- qualitative phenotype data e.g. trait-specific measurements
- new data should add new information
- consider if data is dynamic and multi-dimensional

Estimating the amount of data is difficult as currently phenotype data is submitted to various non-standardized repositories.

EMBL-EBI
GENOPHENIX
BioSamples
Home | Search | Upload | Graph Search | Submit | Documentation | About
EuroFAANG Research Infrastructure
ENA European Nucleotide Archive
PIGWEB
SmartCow
elixir
AQUA EXCEL 3.0
BioStudies.
BioImage Archive

Figure 1: Schematic showing how we would envisage submission of phenotype data to core EMBL-EBI ELIXIR database resources including BioStudies and the BioImage archive for the GenoPHEnix RI proposal.

Phenotype data, due to the large heterogeneity of formats, is currently stored in diverse databases or file systems. As such it is more difficult to estimate the size of data that would need to be injested by the Data Portal. If in text (or equivalent) format, the amount of data represents a few Gb per year. The most demanding data storage are from scanner/video recordings with can represent dozens of Tb per year, expecting to highly increase in the next few years. As an example at INRAE, where all data collected is organised in several interconnected databases, including routine and experiment recording, the amount of routine data available is 60 Gb (with an increase of

2 Gb per year), 210 Gb for experimental data (with billions of recordings) and 75 Tb for video recording (with an increase of 25 Tb per year).

The scale of this data, particularly for the image data and data coming from sensors and including 100's of millions of data points may be unmanageable through the current infrastructure. As such discussion is needed as to which phenotype data should be deposited for submission. New data being added should be dynamic, new data should provide new information so novel datasets could be prioritised as could data from populations of animals with multiple layers of data, e.g. from highly phenotyped animals with both multi-omic, genotype and phenotype information. These discussions are currently on-going within the community and facilitated through the ELIXIR Domestic Animals Genome and Phenome Focus Group and other initiatives including GenoPHEnix (https://genophenix-ri.eu) ESFRI project development and the EU-LI-PHE Cost Action (https://eu-li-phe.eu) which is focused on animal phenotyping and included a working group on Genome to Phenome Integration led by Emily Clark.

In summary to provide develop the data infrastructure to incorporate phenotyping data as part of the GenoPHEnix RI project proposal we would:

- Potentially rebrand the FAANG Data Portal as the GenoPHEnix data portal.
- Support a highly standardised data submission system for animal agriculture & aquaculture data.
- Expand the FAANG portal, to include new data types, including phenotypic measurements and image data.
- Bring new EMBL-EBI/ELIXIR data archives into the portal infrastructure.
- Connect other European initiatives e.g. ELIXIR - Focus Group for domestic animal genomes and phenomes and the EU-LI-PHE Cost Action.

For this deliverable we have also updated the DMP (D3.1) and data access policy (D3.2) to add incorporation of phenotyping in the Data Portal as a significant infrastructure development (Annex 1). The current DMP (D3.1) remains fit for purpose for omics datasets and the EuroFAANG RI project.

7. Conclusions

In conclusion we were able to identify some operability and functionality issues related to the existing data infrastructure early in the EuroFAANG RI project and have modernised and updated the FAANG Data Portal transitioning to cloud, moving from Angular to Python Dash and supporting increased search term functionality. Based on the estimates we were able to produce the Data Portal can handle the amount of omics data currently being generated by FAANG projects. Working with work packages 4,5 and 6 we have identified capacity that could be added to the Data Portal infrastructure, including, providing catalogues of CRISPR single guide RNAs, sharing encrypted data and updating the ontologies and metadata standards for in vitro systems. Through participation in global FAANG activities we identified the need to provide training videos for metadata and data submission and are working to provide these before the end of the project. The decision to merge the EuroFAANG RI project with the phenotyping INFRAIAs for the

GenoPHEnix ESFRI proposal meant that we were able to provide a new DMP and Data Access Policy developing the data infrastructure for animal phenotyping data. The shared outcomes of this deliverable provide a strong foundation for the next stage of EuroFAANG and GenoPHEnix shared data infrastructure ensuring sustainability beyond the end of the current funded EuroFAANG RI project.

This deliverable can be updated if further infrastructure gaps are identified before the end of the project. Further updates to the DMP and Data Access Policy may also be required.

8. References

1. Harrison PW, Sokolov A, Nayak A, Fan J, Zerbino D, Cochrane G, Flicek P. The FAANG Data Portal: Global, Open-Access, "FAIR", and Richly Validated Genotype to Phenotype Data for High-Quality Functional Annotation of Animal Genomes. Front Genet. 2021 Jun 17;12:639238.
2. Chamorro-Padial J, Virgili-Gomá J, Gil R, Teixidó M, García R. Agriculture data sharing review. Heliyon. 2024 Dec 27;11(1):e41109.
3. Koltes JE, Cole JB, Clemmens R, Dilger RN, Kramer LM, Lunney JK, McCue ME, McKay SD, Mateescu RG, Murdoch BM, Reuter R, Rexroad CE, Rosa GJM, Serão NVL, White SN, Woodward-Greene MJ, Worku M, Zhang H, Reecy JM. A Vision for Development and Utilization of High-Throughput Phenotyping and Big Data Analytics in Livestock. Front Genet. 2019 Dec 17;10:1197. doi: 10.3389/fgene.2019.01197.

9. Annex 1 – Updated DMP for infrastructure development for phenotyping data

Data Management Plan and Data Access Policy for GenoPHEnix – Developing the existing data infrastructure for EuroFAANG to include phenotyping data

Adapted by Emily Clark (EMBL-EBI) and Catherine Larzul (INRAE) from the original EuroFAANG RI Project Data Management Plan (Deliverable 3.1) and EuroFAANG Access Policy and Data Principles (Deliverable 3.2) written by Peter Harrison (EMBL-EBI).

Introduction

The GenoPHEnix Research Infrastructure (RI) is focused on open science and FAIR data principles according to the data sharing principles outlined by the global FAANG consortium, ELIXIR standards and existing European policies. This policy is in line with the Formalised Data Policy and Access Principles, published for the EuroFAANG Concept Development Project in 2024 (Deliverable 3.2 - approved 14/01/24) and Data Management Plan (Deliverable 3.1 – approved 30/06/25) which was developed in accordance with EU Horizon Europe guidelines, EU focus on open science and the FAIR data sharing principles.

As for the EuroFAANG Data Management Plan and the Transnational Access Policy, the Data Management Plan for GenoPHEnix will be a living document that will be updated regularly throughout the lifecycle of the GenoPHEnix research infrastructure to reflect the evolving needs of the infrastructure and the communities it supports. The data policy will also develop based on the work undertaken within the GenoPHEnix Consortium and the outcomes around data generation, processing, and transnational access, which are focused on generation of omics data, genome editing, consolidating biobanking efforts, phenotype recording, and interaction with other key EU Research Infrastructures and services, including Elixir.

This policy is designed to ensure open and FAIR access to high quality data to ensure exploitation of results and enable effective genotype to phenotype research in farmed animals by the European research community and globally.

The GenoPHEnix Research Infrastructure (referred to as "GenoPHEnix-RI" hereinafter) is committed to open science principles and recognizes the critical role of research data in advancing knowledge and fostering innovation. The principles and procedures governing access to and reuse of data generated by and in collaboration with the GenoPHEnix-RI are in accordance with the FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles and EU open science policies. It includes expectations for handling data generated by the GenoPHEnix Research Infrastructure, and the principles of accessing data generated by the infrastructure.

By implementing an open data access policy, the GenoPHEnix RI aims to contribute to a more transparent and collaborative research ecosystem advancing knowledge discovery, accelerating scientific progress, and ensuring exploitation of results, to enable effective genotype to phenotype research in farmed animals by the European research community and globally.

1. The GenoPHEnix Research Infrastructure

The GenoPHEnix Research Infrastructure (RI) aims to create a pan-European, multi-species, multidisciplinary platform for animal genetic resources, phenotyping, and sustainable breeding, focused on improving animal health, welfare, resilience, and efficiency. GenoPHEnix will support cutting-edge research in farm animal science while contributing to Replacement, Reduction and Refinement (3Rs) principles in animal experimentation and help the livestock and fish sector to develop more sustainable production systems. GenoPHEnix RI will improve the leadership and excellence of the European farmed animal research community by achieving: 1. A shared capacity for deep phenotyping of farmed animals in complementary environmental conditions, with identification of biomarkers for breeding and management; 2. Standardisation and FAIRisation of animal to genome and phenome data; 3. Expanded biobanking capabilities for animal genetic resources and cellular models; 4. Synergies with existing projects and other RIs to consolidate European farmed animal research.

2. Open access policy

As an infrastructure, GenoPHEnix will be supporting the community generation of open data and research output for genome and phenome interaction for animal farming. Open access to GenoPHEnix data is crucial both for verification of results and to ensure community re-use. The GenoPHEnix RI promotes open access to and reuse of research data, following the "as open as possible, as closed as necessary" principle. By default, data generated by the GenoPHEnix RI will be openly shared unless legitimate reasons warrant a more restricted access, such as commercially sensitive proprietary data.

The scope of legitimate restriction is being explored in detail by the GenoPHEnix RI as part of its Access Policy principles. This document outlines that the infrastructure in principle strives to be open, but the inclusion of industrial resources and data with Intellectual Property rights also offers clear advantages and increases research and innovation in farmed animals through fostering collaboration between industry and research communities. As principles of the GenoPHEnix access policy applies to wide virtual access from business users, the requirements for data access restrictions will be put in place to provide the necessary and legitimate protection of Intellectual Property rights, whilst ensuring overall value for the wider research community and EU open science requirements of publicly funded research.

Data Providers

Research data which are created by the GenoPHEnix RI are owned by the partner who generates the data and maintained under its responsibilities. Raw data will be curated and preserved by partners following their internal procedures.

Results, processed data and aggregated raw data sets will be stored in public repositories and made available through the GenoPHEnix data portal until at least five years after the end of the project.

Data types

Data collected in animal experiments can be numerical data, images, sounds, spectra, documents, texts. They can be recorded automatically with devices or manually. On the overall, a catalogue of traits recorded on animals, or on their products, would list several thousands of terms. Data are complemented by metadata that describe the research data. This may include submitter information, contact data, time of creation. It defines the data formats used and contains the context that led to the data being generated. Additionally, metadata should include animal information (e.g. sex, breed) settings of experimental sensors or instruments, environmental conditions, comments, and measurement uncertainties.

Data description may include:

- Characteristics of experiments including design, protocol and general organisation
- Resource description: information of tissue/organoïds/cell/animal, genotype, farming or aquafarming system,
- Facilities: installations, sensors, cameras or any specific devices
- Trait recovery work-flows: sensor and image analysis methods and software tools used to extract traits from raw images, spectra, signals, or other types of data
- Phenotypic data at sample, individual or group level, (e.g., weights, feed consumption, behaviour)
- Environmental conditions as collected by sensors (e.g. air temperature or hygrometry)
- Date and description of management or observation events

For phenotypic data, the preferred formats are:

- For tabular data: CSV, TXT
- For semi-structured data: JSON, XML
- For Textual data: TXT, unicode encoding.
- For images: mp4, jpg

For genomic data, the preferred formats are:

- For raw reads: Primarily submitted as FASTQ files, which store base calls and quality scores for each read.

- For aligned reads: Can be submitted as BAM (Binary Alignment Map) files, which represent aligned reads to a reference genome.
- For compressed formats: CRAM (CRAMmed Alignment Map) is a compressed version of BAM, often used for large datasets.
- For Metadata: Provided in a structured format like a spreadsheet or text file, detailing sample information, sequencing platform, and experimental conditions.

Where possible, metadata will be described according ontologies developed by INRAE (https://www.atol-ontology.com/) including the Animal Trait Ontology for Livestock (ATOL, 3686 traits, last update 14/10/2024) the Animal Health Ontology for Livestock (AHOL, 718 traits, last update 05/12/2024) and the Environmental Ontology for Livestock (EOL, 584 terms, last update 09/10/2024) or by Iowa State University (Vertebrate Trait Ontology, 3920 classes, last update 04/03/2025,https://bioportal.bioontology.org/ontologies/VT). Ontologies are updated regularly to include new trait references and according to the set of metadata standards described by the FAANG consortium (Harrison et al. 2018 Anim Genet. 2018 Oct 12;49(6):520–526).

Current and projected volumes

To describe the expected amount of data that GenoPHEnix would handle, we have provided the following four examples; 1) For AquaFAANG, one of the H2020 EuroFAANG projects, the size of data produced, between 2021 and 2013, was 61.3 Tb in total which gives an example of the total output from a large research and innovation project that would need to be managed by the GenoPHEnix RI; 2) At CIGENE, one of the organisations within GenoPHEnix, the amount of data generated in 2021 and 2022 was 9Tb and 7Tb respectively which provides an approximation of the potential size of genomic data produced by each partner annually; 3) Table x provides an example of the number and size of the genomic datasets for farmed animals that were deposited in the European Nucleotide Archive (a core data service delivered and maintained by EMBL-EBI) in 2024; 4) At INRAE, which organised all data collected in several interconnected databases, including routine and experiment recording, the amount of routine data available is 60 Gb (with an increase of 2 Gb per year), 210 Gb for experimental data (with billions of recordings) and 75 Tb for video recording (with an increase of 25 Tb per year).

Due to the large heterogeneity of formats, phenotypic data are stored in different databases or file systems. If there are in text (or equivalent) format, the amount of data represents a few Gb per year. The most demanding data storage are from scanner/video recordings which can represent dozens of Tb per year and is expected to highly increase in the next few years.

Types of Use

A: Retrieve data stored at the infrastructure and use locally
B: Obtain biological resources stored at the infrastructure and use locally
C: Online access to and use of the compute resources of the infrastructure
D: On site access to and use of the compute resources of the infrastructure
E: On site access to and use of the experimental facilities of the infrastructure

Types of Users

1: Academic user fully open access
2: Academic user partial open access (e.g. for use of the infrastructure within an ongoing collaboration between an academic and industrial partner)
3: Industrial user fully open access
4: Industrial user partial open access
5: Industrial user, private use only

3. Data Sharing Policy

The GenoPHEnix RI will follow the global GenoPHEnix Access policy principles applicable to wide virtual access. Fully open access users are required to follow the GenoPHEnix wide virtual access principles below, and all data consumers will need to follow the provisions against each dataset as recorded in the public archives and GenoPHEnix Data Portal.

Exceptions to this policy will be granted based on the types of use and types of users as defined by the Transnational Access (TNA) Policy. The process for granting this, for example to protect Intellectual Property rights for industry users of the RI, will be clarified and recorded here once these processes for TNA are further developed during the preparatory phase of the GenoPHEnix RI, as the RI establishes its processes for access and types of access.

For the latest version of this policy please see https://www.faang.org/data-share-principle, but for clarity the version as per the time of publishing of this document is recorded below.

The FAANG Data Sharing Statement

Version 2.0 (December 1, 2021) Definitions **Archive** means one of the archives hosted at the EMBL-EBI, NCBI or DDBJ. These include the ENA, Genbank, ArrayExpress and Geo. A full list of the FAANG recommended archives is available as part of the FAANG metadata recommendations. **Submission** means data and metadata submission to one of the FAANG recommended Archives. **FAANG member** means an individual who has signed up to the FAANG consortium through the FAANG website and agreed to the FAANG core principles. **Data** means any assay or metadata generated for or associated with FAANG experiments. **Analysis** means any computational process where raw assay data is aligned, transformed or combined to produce a new product. **Primary analysis** results consist of sample level analysis such as alignment to a reference genome or quantification of signal in the assay. **Integrated analysis** results represent analyses which draw together data from multiple samples and/or experiments such as genome segmentation or differential analysis results. **Internal** means data that is only accessible via the FAANG private shared storage. **Private** shared storage means a storage space hosted at EMBL-EBI that has access limited to agreed persons by the data provider **Public** means all data is available through the FAANG public data portal and underlying public archives, without embargo and is accessible to everyone.

This document describes the principles of data sharing for the FAANG consortium. Any queries about this document should be sent to faang@iastate.edu and faang-dcc@ebi.ac.uk.

FAANG believes that pre-publication data-sharing, collaboration and data reuse is for everyone's benefit and is strongly encouraged.

For FAANG data consumers:

FAANG data are released under the Fort Lauderdale and Toronto principles 1,2. FAANG data creators reserve the right to first publication of the results obtained from using a dataset in genome wide analysis (see box 1 for clarifying examples). The publications made on any dataset can be checked on the FAANG Data Portal (https://data.faang.org/). If you are unsure if you are allowed to publish on a dataset, please contact the FAANG Data Coordination Centre and FAANG consortium (email faang-dcc@ebi.ac.uk and cc faang@iastate.edu to enquire.)

When using FAANG data you should **cite relevant publications and preprints from the data creators as well as all of the data accession numbers (e.g. PRJEB19199) in the main body of the publication** (not in the supplementary materials).

The FAANG consortium is producing high quality and well-annotated datasets to support the community in generating a powerful genome to phenome resource and promotes rapid dissemination of data to accelerate research. FAANG datasets are high quality, focus on a standardised set of multi-omic assays, are accompanied by rich validated metadata, phenotypic information and detailed protocols.

Examples of **permitted use**, that must include **citation of relevant publications or preprints from the data creators and the dataset accession numbers** in the resulting manuscript:

Any researcher may download sequence data and/or derived bed files from the data portal, map these data to a genome and may derive results from these mapped data to address limited questions in their own research projects such as:

1. Is a specific set of genes expressed in a distinct tissue or set of tissues?

2. Is a locus, or pathway impacted by a particular histone mark?

3. Are particular SNV allele(s) present in the FAANG dataset?

4. What functional elements are present in a genomic region of interest for a particular trait?

Examples of **prohibited use** without prior publication from the data creators or **permission from the author**:

What is prohibited is the publication either on-line, or in the peer reviewed literature, of the results of a genome wide analysis of these data. Examples include but are not limited to:

1. Publishing on-line or in the peer reviewed literature a genome wide gene annotation file (gtf or bed) detailing transcription and isoform variation for the species' genome.

2. Publishing on-line or in the peer reviewed literature a genome wide survey of allele specific expression of transcripts and isoforms.

3. Publishing on-line or in the peer reviewed literature results derived from an integrated analysis of these data with other datasets for a genome wide study.
The above examples are not an exhaustive list, **if in doubt, please contact the FAANG Data Coordination Centre and FAANG consortium (email faang-dcc@ebi.ac.uk and cc faang-contact@animalgenome.org**). provide these data pre-publication to encourage data reuse for maximal benefit to the community.

The FAANG Steering Committee commits to report to journal editors and the laboratories involved **any event that disregards the rights of data creators** (including biological measurements as well as analysis of such measurements).

Fostering collaboration through joint data analyses is also highly encouraged so you are invited to contact data creators directly (or via faang-dcc@ebi.ac.uk), or seek collaborative partners amongst the FAANG working groups and membership.

For FAANG data producers:

FAANG recognizes that rapid sharing of the sample metadata and raw data generated by the consortium with the wider community is a priority. FAANG aims to ensure that everyone can benefit from the data created by FAANG to aid their own research as rapidly as is possible.

- All sample metadata and raw data produced for a FAANG associated project will be submitted to the public archives, without any hold until publication date, as soon as possible after sampling or data generation and initial quality control checks.

- All primary and integrated analysis results produced for a FAANG associated project are also encouraged to be made public prior to publication without embargo. However, it is acceptable that primary and integrated analysis results are kept private until publication, as long as the sample metadata and raw data have been made public.

- All FAANG public data are released under Fort Lauderdale and Toronto principles 1,2. The FAANG website, dataset descriptions and Data Portal have clear data reuse statements. The FAANG submission guidelines describe the suggested statement to include with your dataset submissions (https://dcc-documentation.readthedocs.io/en/latest/experiment/ena_template/).

- The Data Portal has developed mechanisms to clearly identify which datasets are unpublished and which have at least one publication.

For FAANG primary and integrated analyses not made available in archives pre-publication, FAANG recognizes the need to enable and promote collaboration amongst consortium and community members. FAANG therefore provides functionality for primary and integrated analyses to be privately shared between FAANG members in private shared storage hosted at the EMBL-EBI.

This requires an agreement between the two parties and that all have agreed to **the Fort Lauderdale and Toronto principles** 1,2.

Only FAANG data can be submitted to the FAANG Data Portal.

All members of FAANG can and will continue to do experimental and analysis work outside of FAANG and the other data generated is not required to meet the same data sharing expectations. Software and analysis pipelines developed by FAANG consortium members are strongly encouraged to be released under permissive open source software licenses wherever possible, such as Apache 2.0.

The FAANG Steering Committee commits to report to journal editors and the laboratories involved any event that disregards the rights of data creators (including biological measurements as well as analysis of such measurements).

REFERENCES:

1. **Fort Lauderdale principles**: Reaffirmation and Extension of NHGRI Rapid Data Release Policies: Large-scale Sequencing and Other Community Resource Projects.

2. **Toronto International Data Release Workshop**: Rapid release of prepublication data has served the field of genomics well. Attendees at a workshop in Toronto recommend extending the practice to other biological data sets.

Version 2.0 Update approved by the FAANG steering committee on 1st December 2021; Original approved on 26th May 2015.

4. FAIR data policy

The new GenoPHEnix Data Coordination Centre and GenoPHEnix Data Portal, which are based on the framework provided by the FAANG Data Coordination Centre and Data Portal, will ensure that data generated in the RI meets the highest possible FAIR principles.

This will ensure that the data will be:

- **Findable:** Data should be easy to find for both humans and machines. GenoPHEnix RI will use consistent naming conventions, providing clear and accurate metadata, and registering data in discoverable public repositories.

- **Accessible:** Data should be accessible to anyone in the research community (noting that it is expected that some generated data from industry may carry additional restrictions to protect Intellectual Property). GenoPHEnix RI will provide open access to data whenever possible and ensure that users provide standard formats and mandatory protocols for data access.

- **Interoperable:** Data should be interoperable with other data sets. GenoPHEnix RI will ensure the use of a common data model and ontologies and providing documentation that describes the data in a way that others can easily understand.

- **Reusable:** Data should be reusable for other purposes than the original study. This means ensuring clear information about the provenance of the data, extensive rich metadata, and open licenses that allow others to reuse the data whenever possible.

To achieve this, users of the GenoPHEnix RI must comply with the below provisions laid out for Brokered submission requirements and metadata standards and Data access procedures For phenotypic data, FAIR data guidelines for pig research can be used as a reference (https://doi.org/10.5281/zenodo.14765747).

5. Long term data management and preservation of data

The GenoPHEnix RI is committed to long-term data preservation and stewardship. Data will be properly documented, archived, and backed up to ensure its integrity and future accessibility. Genomic data generated within the GenoPHEnix infrastructure will have internationally recognised identifiers of the International Nucleotide Sequence Database Collaboration (INSDC). Submission through the GenoPHEnix data portal ensures that data is publicly archived in an INSDC archive. These will be issued upon submission to the EMBL-EBI 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.

6. Brokered submission requirements and metadata standards

GenoPHEnix data will be submitted through the GenoPHEnix brokered submission system that will ensure compliance with GenoPHEnix's metadata standards. For the FAANG portal, the data submission system and detailed instructions are available here https://data.faang.org/validation/samples. Further developments will be discussed within the ELIXIR Focus Group on 'Domestic Animal Genomes and Phenomes'.

All users of the GenoPHEnix RI will be required to meet the minimum GenoPHEnix metadata standards and encouraged to submit metadata as richly as possible. The brokered submission system validates and ensures the minimum standards are met and also suggests further improvements to enrich metadata reporting. This provides a submission process for samples, raw datasets and analysed datasets, with ontologies required to be supplied for many fields. Submissions also require mandatory detailed protocols and links to data workflows used to generate data.

Detailed instructions on making a GenoPHEnix submission for omics data are provided here https://dcc-documentation.readthedocs.io/en/latest/.

7. Access procedures

Data access will be facilitated through the dedicated GenoPHEnix data portal or directly from public archives. Researchers accessing data need to adhere to the terms and conditions of data use, including citation requirements and limitations on redistribution. In the case that GenoPHEnix will coordinate more restricted datasets, pertaining to the outcomes of TNA design, the GenoPHEnix will evaluate data access requests promptly and fairly, considering the potential benefits and risks of sharing. Applicants will be notified of the decision and provided with instructions for accessing the data if their request is approved.

For phenotypic data, following data access methods will also be considered acceptable:

- Direct file download from a public URL, with or without prior user authentication.
- Programmatic access administered through an adequately documented RESTful API residing on a public URL, with or without prior user authentication.
- API documentation must include OpenAPI and/or GraphQL specification.
- API access can be subject to bandwidth and/or quota limitations according to the hosting party's service policies, although these access limitations must be clearly presented to the user.
- Access through a SPARQL endpoint residing on a public URL and adequately docmented.
- File download from Public ftp server redesign on a public address.

8. Data Re-use and Citation

Users are encouraged to reuse data from the GenoPHEnix RI for further research and innovation. Proper citation of the data is mandatory to acknowledge the data creators and contribute to responsible research practices. Data derived from the use of the infrastructure that is submitted to public repositories should acknowledge the GenoPHEnix RI and include the INSDC project identifier in the main text of the publication. Publications, be it in peer reviewed journals, presentations at conferences or in papers directed to the public, should also acknowledge the use of the GenoPHEnix RI.

9. Role of GenoPHEnix Data Coordination Centre

Data will be collated in the context of existing GenoPHEnix, FAANG and community datasets in the GenoPHEnix Data Portal will be supported by the EMBL-EBI Data Coordination Centre pride themselves on ensuring 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. The EMBL-EBI Data Coordination Centre will maintain the data standards, submission infrastructure and Data Portal.

10. Monitoring and Review

This policy will be reviewed and updated periodically to reflect evolving open science practices, GenoPHEnix RI data management and legal frameworks. User feedback and input are welcome to ensure this policy remains relevant and effective. The policy is under active development whilst the RI is established, and it is expected that many sections will be updated based on the recommendations of GenoPHEnix RI think tanks and workshop events.

11. Contact and Support

For any questions regarding data access, data management, or this policy, please contact the EMBL GenoPHEnix RI Data Coordination Centre at faang-dcc@ebi.ac.uk.

12. Additional Considerations

Please note that this data policy may be supplemented by specific data use clauses for complex or sensitive data as part of TNA agreements between the GenoPHEnix RI and users. Industry users in particular should contact the GenoPHEnix RI to discuss specific data access and data generation requirements.

In case it is anticipated that the use of the GenoPHEnix infrastructure could result in data, resources or procedures that can be protected with IP rights, the use and ownership of such IP rights need to be agreed before commencing the TNA. Any background knowledge (e.g. information, know-how, data, or material, including any IP Rights pertaining to such knowledge) that is held by the user prior to use of the GenoPHEnix RI also needs to be described in specific clauses in the TNA agreement before commencing with the TNA.

The GenoPHEnix RI may collaborate with other research infrastructures and data repositories to facilitate data sharing and interoperability.

User training and resources will be available to support researchers in effectively accessing and reusing data from the GenoPHEnix RI, please consult the GenoPHEnix website for further information (https://genophenix.eu/).

Development of this data management plan and data access policy was funded partially through the EuroFAANG RI concept development 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 consortium's view. The European Commission is not responsible for any use that may be made of the information it contains.