UNDERSTANDING THE CONNECTIVITY AND DYNAMICS OF AVIAN INFLUENZA

D6.5 Data management plan

Version 1, 2023-10-30

deliverable:	D6.5
Due date:	M6
Submission Date:	2023-10-30
Delivered by:	Friedrich-Loeffler-Institut, Coordinator

Dissemination level

☒ Public

☐ Confidential, only for members of the consortium (including the Commission Services)

KAPPA-FLU has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101084171. It was funded within work program topic HORIZON-CL6-2022-FARM2FORK-02-03.

TABLE OF CONTENTS

INTRODUCTION AND SUMMARY

The Horizon Europe Model Grant Agreement requires that a data management plan ('DMP') is established and regularly updated. The following DMP plan was coordinated with all partners of the project.

1. DATA SUMMARY

1.1. WPS 1 TO 6

The majority of the data generated for KAPPA-FLU is in the form of files from:

- software and scripts developed for the use of specific laboratory equipment,
- scripts developed for epidemiological analysis, economic analysis, and statistical modelling,
- free or commercial office applications,
- free software available on the Internet used for data processing.

Data from scientific publications available through PubMed and Scopus are used as well as data from different partners already available but not yet published.

Secondary data not always on the public domain will be used for purpose of epidemiological analysis and statistical modelling. Data sources include governmental agencies involved in outbreak response and industry data.

The data generated for the WPs are numerical, textual and image-based (example formats: xlsx, txt, docx, pptx, pdf, jpeg, tiff, pptx and other licensed formats; up to several megabytes per file). Each of the partners involved in the project ensures the local or cloud backup of these data. Access to these data is secured. The FLI pursues a professional backup strategy.

KAPPA-FLU will be using data generated in previous related projects that form part of the background of the current project. This project will generate a diverse array of new data, such as sequencing data pertaining to the viruses studied and their hosts through technologies such as pathogen sequencing, RNASeq, scRNASeq. We will also generate data using other, related technologies, such as flow cytometry, ELISA-technology, PCR-data as well as imaging (e.g. immunofluorescence). Where similar data have been obtained previously, we will use such as baseline and for comparison. The purpose of generating data is to obtain sequence information regarding the pathogens studies to resolve their nature and to discern the host response.

We assume the raw data, that will be generated, are in the range of 10TB and the analysed data to be significantly less.

The data will derive from in-house work, using samples studied in the project or refer to previously in-house or published data.

1.2. TYPES AND FORMATS OF DATA TO BE GENERATED/ COLLECTED

KAPPA-FLU will collect and generate a range of data based on which targeted prevention and control strategies can be designed, as summarized in the table below. These data will largely originate from public databases, databases available from partner's own projects and research activities in work packages 1 – 3. Work package 5 will make use of secondary data, not always on the public domain, for purpose of epidemiological analysis, economic analysis of control strategies and statistical modelling. Data sources

include governmental agencies involved in outbreak response and industry data. Access and use of data not on the public domain will be governed by data sharing agreements.

Standards for collecting, curating and preserving the data are integrated in the research activities of KAPPA-FLU to ensure data is kept securely and protected from inappropriate use or disclosure. Where appropriate, KAPPA-FLU will leverage existing standards, and continue to develop existing ones for data representation, for example for sequence-based experiments. The research data generated by KAPPA-FLU will be curated in such a way that the formats will be compatible with publicly available data repositories.

For more details see Table 2 in the attachment.

1.3. RE-USE OF EXISTING DATA

In KAPPA-FLU, partners will make use of specific existing datasets collected from public databases, databases available from partner's own projects and research activities, or datasets made available by one or more partners to the consortium. The re-use of existing data is necessary to value and integrate obtained results into the current state of knowledge. These include the following data, sources, and re-use:

Table 1: Overview on data, sources, and planned re-use

Data/ source	Re-use	Expected size of data
Virus sequences / Genbank	Creation of phylogenetic trees	Order of MB
Virus sequences / GISAID	Creation of phylogenetic trees	Order of MB
Animal trial data / Pubmed, Web of Science, unpublished data of partners	Evolution of genetic and phenotypic characteristics of viruses in different avian hosts	Order of MB

Before making datasets available amongst partners or via open access data repositories, KAPPA-FLU will review any agreement(s) with third parties to evaluate any restrictive use of the data by the consortium partners, and evaluate whether or not ethical and other restrictions on further use apply. In addition, whether permissions for use of the data are needed from third parties will be evaluated.

2. FAIR DATA

2.1. MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA

Datasets will be deposited in sustainable and searchable data repositories, either publicly available/open access (e.g., Zenodo), or secure data repositories hosted by the respective institutions for sensitive data.

Persistent digital identifiers (Digital Object Identifiers) will be used for publications arising from the data, and linked to the ORCID and online research profiles of the scientists involved in the creation of the dataset. Sequence data follow the ENA-filed persistent and unique identifiers system of the European Nucleotide Archive.

Standard metadata schemes and controlled vocabulary will be used where available (e.g., for sequence data, ENA provides a standard checklist including metadata of samples and experimental levels with controlled vocabulary). International naming conventions will be applied to isolate and strain names.

Active datasets will be shared between collaborators using file sharing services (e.g., SharePoint provided by FLI) so that all project partners edit the same document if required. Changes to previous document versions will be marked (e.g. by using Track Changes mode) and different versions of datasets or documents will be assigned an identifying suffix, e.g., v1, v2 etc.

2.2. MAKING DATA ACCESSIBLE

All relevant research data will be made publicly available, apart from sensitive data, e.g., any data sets that is subject to GDPR regulations non-disclosure agreements, or details of animal experiments. Open access data and associated metadata/documentation will be shared in sustainable publicly available data repositories. Public databases like the European Nucleotide Archive repository for sequence data or GitHub for model algorithms and codes will be used. All data will be available from these databases without the need for special software and in standard data formats. Appropriate data sets are reviewed for the possibility of a data publication. At the end of the project, datasets will be deposited in a secure repository so that they remain reuseable for at least 10 years from the project end date. Under German law, all laboratory records for viral GMOs have to be preserved for 30 years.

2.3. MAKING DATA INTEROPERABLE

To ensure the interoperability of data, where appropriate, KAPPA-FLU will leverage existing standards. KAPPA-FLU will make the generated and collected research data available via appropriate, dedicated, and, where possible, open access archives and database repositories. In addition, KAPPA-FLU will upload basic datasets in standardized forms in a primary database as required by the journal in which the consortium partners publish their results.

As the project activities and the consortium partners span multiple scientific disciplines, partners will agree on standard vocabularies for all data types present in KAPPA-FLU generated and/or collected data sets, as well as on common terminology, key words, units and formats to be applied to the data and other documents.

A standard vocabulary for all data types will be used. Data files will be deposited using open formats allowing interoperability between systems and applications and preventing data loss (e.g., plain text, comma-separated values for spreadsheets, fastq/fasta for sequence information, binary alignment maps or sequence alignment maps for raw genome data). Where proprietary file formats are inescapable, they will be selected to be accessible by different operating systems and with different software packages.

2.4. WILL YOUR DATA INCLUDE QUALIFIED REFERENCES TO OTHER DATA? INCREASE DATA RE-USE

Open access data will be licensed for reuse under the terms of the Creative Commons Attribution 4.0 (CC BY 4.0) license. In general, data will be made available for reuse at the time of publication of project outcomes, with links to the dataset (DOI) included in published articles. Embargoes on data access may be applied during the process of applying for patents.

The data quality is ensured by different measures. These include validation of the sample, experimental and sample replication, comparison with results of similar studies and control of systematic distortion. Data will be structured and labelled in a logical and consistent manner, using file formats compatible with the widest possible reuse by different software, systems and applications.

As open formats are used for data archiving, the data will remain re-usable. At the end of the project, datasets will be deposited in a secure repository so that they remain reusable for at least 10 years from the project end date.

Readme files will be generated - where necessary - to provide documentation.

3. OTHER RESEARCH OUTPUTS

Newly generated protocols and workflows will be published as part of FAIR Data.

4. ALLOCATION OF RESOURCES

Data security is country specific and the partners/work package leaders in each country are responsible for the data management within their legislative context. External data repositories may also be used for data sharing, if appropriate.

Long term preservation will result in no additional costs other than repository charges for data submission, if any.

5. DATA SECURITY

Each partner in KAPPA-FLU is responsible for data stored locally and for complying with their own standard operating procedures (SOPs), and any KAPPA-FLU specific SOP that may be drawn up if deemed necessary, as well as for complying to the relevant legal and ethical requirements. The following general terms regarding data protection will be followed by the partners in KAPPA-FLU with regard to raw and final research data which is not stored in open repositories:

- The amount of data collected is relevant and not excessive;
- Data are stored securely;
- Data are fairly and lawfully processed;
- Data accuracy is ensured (i.e. all reasonable efforts to ensure data accuracy are undertaken);
- Data are used only in ways that are compatible with the original consent or agreement;
- Relevant national and international regulations regarding data protection will be applied.

Data storage facilities will be maintained in accordance with manufacturers' guidelines. Data will be backed up at regular intervals, and stored safely and securely, in accordance with the consortium partners' organizational policy.

During the project, active research data will be saved and stored on secure institutional servers with backup on a regular basis, hosted by the respective partner institutions. There will be restricted access, login credentials, and user verification. Access to sensitive data is granted only for project members with clearance through non-disclosure agreements. Where data has to be shared with collaborators, it will be contained within password protected or controlled access files and shared via e.g. ShareCloud (provided

by FLI), a dedicated project website with a secure server or other secure cloud services. Sensitive data will be encrypted before transfer or sharing. Data transfer is secured via HTTPS protocol.

The final collated datasets used in publications will be deposited in sustainable and searchable data repositories, either publicly available/open access (e.g., Zenodo), or secure data repositories hosted by the respective institutions for sensitive data.

6. Ethics

Animal experiments and components of the project involving collection of primary data through questionnaires or interviews will be authorized by the relevant committees in accordance with EU and national law.

External Independent Ethics Advisors will be consulted at least on the following points: processing of personal data, animal studies, environmental risks, transfer of personal data and material between EU and non-EU countries and benefit-sharing actions, gain-of-function research/DURC issues. In case of ethical issues data sharing will follow the recommendations of the board.

7. Other issues

N/A

ATTACHEMENT

Table 2: Description of data types

Type of data	Origin of data	Format	Expected size of data
Virus sequences	Public databases (Genbank: https://www.ncbi.nlm.nih.gov/genbank/ or GISAID: http://platform.gisaid.org), work packages 1 to 3)	FASTA	Depends on amount of metadata (order of MB)
Wild bird tracking data	Public database (https://www.movebank.org/), work package 1	.csv, Movebank format	Depends on number of birds and duration of tracking (order or GB)
Animal trial data (tissue tropism of viruses, excretion dynamics of viruses, virulence of viruses in infected birds)	Public databases (PubMed: https://www.ncbi.nlm.nih.gov/pubmed or Web of Science: http://apps.webofknowledge.com/), work packages 1 to 3	Word, Excel	Order of MB
Outbreak data including date and place as well as control measures.	National agencies involved in outbreak response and international agencies aggregating data from national agencies (e.g. EFSA).		Order of MB
Data on potential determinants for disease incursion and spread.	Repositories of climatic, demographic, poultry production data and socioeconomic indicators.		Order of MB
Poultry industry data on production parameters, costs and impact of outbreaks as well as of surveillance and control activities.	Poultry companies and national agencies involved in outbreak response.		Order of MB
Primary data from interviews / questionnaires with stakeholders	Stakeholders from the poultry sector and national agencies involved in outbreak response / animal health / public health.		Order of MB