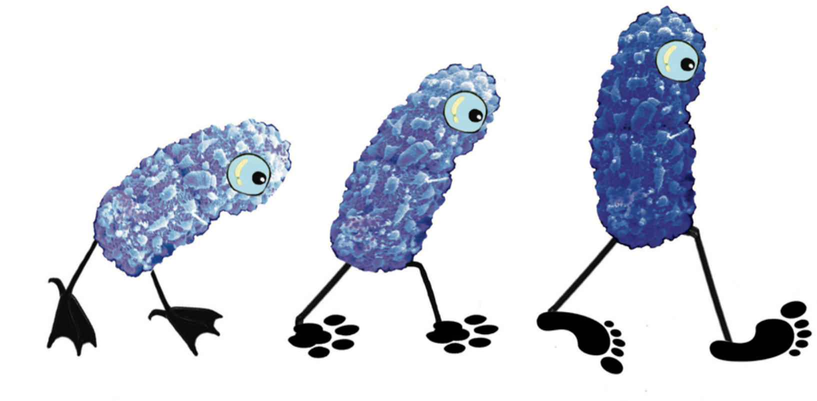

MedVetKlebs: *Klebsiella pneumoniae* from ecology to source attribution and transmission control

(A component of European Joint Programme One Health EJP, funded by the European Union's Horizon 2020 research and innovation programme
Grant Agreement No. 773830)

Data Management Plan (DMP)

Horizon 2020 FAIR DMP

Admin Details

Project Name: "MedVetKlebs - *Klebsiella pneumoniae* from ecology to source attribution and transmission control"
Principal Investigator/Researcher: Sylvain Brisse (sylvain.brisse@pasteur.fr)
Funder: European Commission (Horizon 2020)

Version information

Version number: DMP final version
Date of first version: 11.04.2019

A. DATA SUMMARY

A.1. What is the purpose of the data collection/generation and its relation to the objectives of the project?

The MedVetKlebs project aims to develop, evaluate and harmonize methods for sampling, detection, strain typing and genome-based genotyping of *Klebsiella pneumoniae*, and share these methodologies across Institutions and with the scientific community to optimize the current practices. In this context, the main data generated across this project are protocols, genomic sequences and the biological/metadata associated, modelling data and web applications.

A.2. What types and formats of data will the project generate/collect? What is the expected size of the data?

Table 1. Categories, types, format and location of the data

Category	Type	Format	Accessible in
Research Data	Biological data	MS Excel compatible files (.xlsx, .csv, .txt)	https://zenodo.org
	Molecular data (e.g. whole-genome)	.fastq; .fasta	ENA (https://www.ebi.ac.uk/ena), NCBI (https://www.ncbi.nlm.nih.gov/), BIGSdb (https://bigsdb.pasteur.fr/klebsiella/klebsiella.html)
Methodologies	Methods/Protocols	.pdf	https://www.protocols.io
	Software	.Rshinny	https://maldityper.pasteur.fr
	Statistics/Modelling data	.dta (stata); .Rdata (R)	https://zenodo.org, Github

Generated data are discriminated in **Table 1**. The expected size the data is of ~3Tb.

A.3. What is the origin of the data? Will you re-use any existing data and how?

No old data will be used in this project. All deliverables from the project will come from original data acquired during this period.

A.4. To whom might it be useful ('data utility')?

The data produced within the project are useful for medical, veterinary and environmental microbiology laboratories, European public health and safety authorities, broader scientific community and general public.

B. DATA FINDABLE

MedVetKlebs, compliant with H2020 open-access policy, will deposit generated and collected data in open online research data repositories. Please see **Table 1** for more details.

B.1. What naming conventions do you follow? Do you provide clear version numbers?

For metadata, protocols and dataset names, we will define naming convention consisting in 3 mandatory parts:

- A prefix, indicating if it is a protocol or metadata;
- A root consisting of a short and meaningful name of the protocol/metadata;
- A suffix indicating the date of the last upload into the Repository in YYMMDD format.

Each of these parts are separated by an underscore: _
E.g. Prot_KlebsiellaIsolationFromMeat_190321.pdf

For strain and genome names, we will define naming conventions consisting in 4 mandatory parts:

- A prefix, indicating the acronym of the project;
- A root consisting of:
  - the code of the partner of the project;
  - a suffix indicating the origin of the isolate;
  - a 3 digits number indicating the number of the isolate.

Each of these parts are separated by a dash: -
E.g. MVK-01E001

B.2. Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

The chosen repositories (please see **Table 1** for more details) allow researchers to deposit both publications and related research data and linking them to these through persistent identifiers (such as Digital Object Identifiers).

B.3. What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Metadata are based on ZENODO’s metadata, including the title, creator, date, contributor, description, keywords, format, resource type, etc.

B.4. Will search keywords be provided that optimize possibilities for re-use?

Specific keywords based on the metadata files (e.g. OHEJP, MVK, by source, etc.) will optimize the search across datasets in order to provide possibilities to easily re-use the data.

C. DATA ACCESSIBLE

C.1.Which data produced and/or used in the project will be made openly available as the default?

By default, the data produced or used in the project will be made openly available through different repositories (please see **Table 1** for more details).

C.2. How will the data be made accessible (e.g. by deposition in a repository)?

Data and results will be communicated to the scientific community, decision makers and general public through publications in scientific journals, presentations at conferences and through open-access data repositories (please see **Table 1** for more details).

C.3. What methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

For most data, only standard software tools, *e.g.* web browsers, pdf file readers, and text and tabular readers, will be needed. However, in the case of genomic and modelling data, specialized tools will be required to access the data; these tools are standard bioinformatics tools and are openly available.

C.4. Where will the data and associated metadata, documentation and code be deposited? Have you explored appropriate arrangements with the identified repository?

Please see **Table 1** for more details.

C.5. If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Not applicable.

C.6. If there are restrictions on use, how will access be provided? Is there a need for a data access committee? Are there well described conditions for access (i.e. a machine-readable license)? How will the identity of the person accessing the data be ascertained?

Not applicable.

D. DATA INTEROPERABLE

D.1. Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc?

Data and metadata will be stored and shared in standard formats (**Table 1**) and will thereby be interoperable.

D.2. What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable? Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary?

MedVetKlebs aims to use metadata standards in use within the One Health surveillance domain in order to use a common vocabulary, codes list and mapping of pre-defined values for harmonizing the descriptions of metadata and data.

E. DATA RE-USABLE

E.1.Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

Public data will be available from open repositories (**Table 1**), and therefore reusable by third parties, even after the end of the project.

E.2. How will the data be licensed to permit the widest re-use possible?

The public repositories chosen to store and allow the re-use of the data use CC-BY license (https://creativecommons.org/licenses/).

E.3. When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

For research data, open access will be provided once the associated research manuscript is published or available in open repositories. The embargo period envisaged will be of approximately two years after the end of the project, in order to guarantee the publication of the results before giving access to the data for others to use.

E.4. How long is it intended that the data remains re-usable?

No end date is envisaged for the re-use of the data. All the data will be stored in public repositories and retained for the lifetime of the repository.

E.5. Are data quality assurance processes described?

Some research data will be deposited in repositories providing curation appropriate to the data (ENA, NCBI, BIGSdb), or will be curated by consortium members ahead of deposition (for protocols or metadata).

F. RESSOURCES

F.1. What are the costs for making data FAIR in your project? F.2. How will these be covered?

Research data and documentation will be available through free repositories (**Table 1**) so no additional costs to guarantee the open access data are envisaged.

F.3. Who will be responsible for data management in your project?

Sylvain Brisse and Carla Rodrigues (both at Institut Pasteur, Paris, France).

F.4. Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

All the research data and protocols (**Table 1)** will be available in free repositories for which the preservation shall be retained for the lifetime of the repository.

G. SECURITY

G.1. What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

In addition to being made available in the aforementioned repositories (**Table 1**), all the data is also stored on the servers of the IT Department of the Institut Pasteur (Paris, France) in a space dedicated to the project. The servers are redundant on two separate sites to ensure availability, and data is backed up periodically according to best practices.

G.2. Is the data safely stored in certified repositories for long term preservation and curation?

Certified repositories in the life sciences are rare. Zenodo has embarked on a certification process. NCBI, ENA and BIGSdb are public reference repositories in the field, which is why they were chosen for the project. Regarding long term preservation, the items will be retained for the lifetime of the repositories and backups will be ensured by the system.

H. ETHICS

H.1. Are there any ethical or legal issues that can have an impact on data sharing? H.2. Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?

Not applicable.