DATA MANAGEMENT PLAN

General Information

This following document is the Data Management Plan (DMP) of the consortium project **S-KELOID** *Understanding Keloid Disorders: A multi-scale in vitro/in vivo/in silico approach towards digital twins of skin organoids on the chip,* - thereafter termed as '*consortium*' - funded by the Agence National de la Recherche (ANR) (France) and the Fond National de la Recherche (FNR) (Luxembourg), for a duration of 48 months.

Grant number ANR-21-CE45-0025-01

Partner institutions involved in the consortium: Université Franche-Comté (UFC), Centre National de la Recherche Scientifique (CNRS), Centre d'investigation clinique (CIC) of Besançon, Institut National de la Santé de la Recherche Médicale (INSERM), Université de Bourgogne (UB), Université de Bourgogne Franche-Comté (UBFC), University of Luxembourg (UL).

Board (i.e., steering committee) members:

- Raluca Eftimie, UFC, Laboratoire de Mathématique de Besançon (LMB), Unité Mixte de Recherche (UMR) CNRS n°6623 , UBFC
    - Scientific coordinator for France
- Stéphane Bordas, UL, Department of Engineering (DoE)
    - Scientific coordinator for Luxembourg
- Jérôme Chambert, UFC, Institut de Franche Comté Electronique Mécanique Thermique et Optique – Sciences et Technologies (FEMTO-ST), UMR CNRS n° 6174 , UBFC
- Gwenaël Rolin, CIC, INSERM 1431
- Franz Chouly, UB, Institut de Mathématiques de Bourgogne (IMB), UMR CNRS n°5584 , UBFC
    - From July 2023: on leave and not associated anymore with IMB.

Members of the consortium:

- UFC, LMB : Alexei Lozinski , O. Adebayo (PhD Student)
- UFC, FEMTO-ST: Emmanuelle Jacquet, Arnaud Lejeune
- CIC: Brice Chatelain, Thomas Lihoreau , Yasmin El Mahi (PhD Student)
- UL, DoE: Lars Beex, Jack Hale, Stéphane Urcun

1. Data description and collection or re-use of existing data

1.1 New data collection methodologies and re-use of data

At CIC and FEMTO-ST:

- 2D and 3D cells co-culture, imaging from microscopy and real-time microscopy.
- Patient data: collected by the CIC under the French legislation as well as EU General Data Protection Regulation (GDPR) and good clinical practices (GCPs), including anonymization by the CIC team. The anonymised data will be shared under a data sharing agreement by the CIC (or any other patient recruitment centre) and others members of the consortium. Each patient in the study will be informed of the study and will provide their consent. Tissue data, healthy and keloid, contain medical history, photography, CT scans and biopsy.

At UFC and UL, numerical simulations of mathematical models (published in open access articles), produced by Matlab, Python and C++ codes, will be composed of images and raw data (as .csv, .txt files). Moreover, UFC members will also collect and re-use data available through various published studies and open databases (for preliminary parametrisation of their mathematical models).

Data provided by a current thesis, belonging to the project "Fibrolution" funded by the 2019-2022 research call « Envergure » of Région Bourgogne Franche-Comté, will be used in the consortium. Clinical data previously collected by the CIC and the FEMTO-ST[1] [1]of healthy and keloid tissue will be used in the consortium. Codes of the project « Fibrolution » will be adapted and re-used by the consortium.

All data added to the consortium data collection will contain their source, methodology, date and sensitivity nature (in vitro/in vivo human or not, in silico).

1.2 Description of the data collection

Expected volume of the database: 2TB x 6 month-based archiving during 48 months = 16TB.

[1]J. Chambert, et al. (2019). Multimodal investigation of a keloid scar by combining mechanical tests in vivo with diverse imaging techniques. J. Mech. Behav. Biomed. Mater, 99: 206-215 https:// clinicaltrials.gov/ct2/show/NCT03312166

Expected volume of the published data: To be defined within the next few months, when publications of articles will be expected. According to the UL guideline of good scientific practices, the published data will be stored for 10 years.

The data collection will contain:

- 2D and 3D images of in vitro experiments: tif preferentially, else jpg ; pdf
- 2D and 3D images of in vivo keloids: mp4 preferentially, else avi ; tif preferentially, else jpg ; nifti (open source format for clinical imaging)
- Medical reports, histological reports: pdf (the original txt document will be stored in CIC)
- Data arrays in .csv, .xml resulting from the experimental data conversion, for modeling purpose
- Codes in Octave/Matlab, Python and C++ and ascii text files.
- Data for visualization (.vtk, .xdmf, .msh) resulting from the modelisation.

2. Documentation and data quality

2.1 Database standard

The consortium will use the Metadata Encoding and Transmission Standard (METS), belonging to the Machine-Readable Cataloging Standards (MARC). The data will be presented in a xml file following the METS :

• **METS header**: the METS document itself, such as its creator, editor, etc.

• **Descriptive Metadata**: May contain internally embedded metadata or point to metadata external to the METS document. Multiple instances of both internal and external descriptive metadata may be included.

• **Administrative Metadata**: Provides information regarding how files were created and stored, intellectual property rights, information regarding the provenance of files comprising the digital library object (such as master/derivative relationships, migrations, and transformations). As with descriptive metadata, administrative metadata may be internally encoded or external to the METS document.

• **Structural Map**: Outlines a hierarchical structure for the digital library object, and links the elements of that structure to associated content files and metadata. The Structural Map is the only section required for all METS documents. For the S- Keloid project, they will 3 hierarchical level: 1. The data source (UFC, FEMTO-ST, CIC Besançon, UB, UL) 2. Methodology: 2D in vitro, 3D in vitro, healthy in vivo, keloid in vivo, Matlab in silico... with all the required refinement. 3. Data: each project belonging to a methodology will be stored in its own folder. This folder will be marked by its level of sensitivity (published, private, confidential) determining its level of access.

• **Structural Links**: Allows METS creators to record the existence of hyperlinks between nodes in the Structural Map. For instance, the folder containing numerical results of UL or UFC that require, as inputs, the CIC data number X should point to the CIC/Methodology/ProjectNumber_X

Currently there is not a unique metadata standard in the community of the consortium, the partners of the consortium belonging to different scientific fields, mathematics, engineering, biology and health.

2.2 Organisation of the database and documentation

1. The first hierarchical level of the database will be the **data source** (~/UFC, ~/FEMTO-ST, ~/CIC, ~/UB, ~/UL), that is to say, the partner institution involved in the consortium that produced the data.
2. Inside each data source, the second hierarchical level of the database will be the **methodology**: for instance, ~/CIC/2D_in_vitro, ~/CIC/3D_in_vitro, ~/CIC/Real_time_microscopy, ~/FEMTO-ST/Uniaxial_healthy_tissue, ~/FEMTO-ST/ Uniaxial_keloid, ~/UFC/Continuous_field, ~/UFC/Agent_based, ~UL/Uncertainty_quantification, ~/UL/Global_sensitivity_analysis... with all the required refinement.

   In order to enable the re-use of these data: **Each methodology folder should be accompanied by a README.txt file explaining the methodology**

3. Inside each methodology folder, the last hierarchical level of the database is the **data**: each project belonging to a specific methodology will be stored in its own folder. If needed, each data folder should contain a file with its structural links (i.e. the links pointing to other data or code nodes required to process the data in the considered folder). Each data folder could be marked with a sensitive warning, depending on the data nature (in vivo) and a specific encryption could be considered.

   Depending on the data nature, file headers, code books, or lab notebooks may be included. Regarding codes, software, libraries and packages versions will be added.

2.3 Data quality

All experimental data will be under processes of calibration, repeated samples and measurements, and standardised data capture as described in the grant proposal ANR-21-CE45-0025-01

Each member of the consortium will be responsible for the consistency and documentation of their own data.

All published data will be peer reviewed.

3. Storage and backup during the research process

Data storage of the consortium will be composed of two locations:

• a secured datahub center where all the data of the consortium will be stored following the organisation described section 2.2. The consortium choses the Biocore datahub facilities (UL – Luxembourg Centre for Systems Biomedicine (LCSB)– Elixir: https://elixir-luxembourg.org/) to be the secured datahub center.

• a personal storage for each member of the consortium provided by the IT support services of their own institution (UFC, CIC, UL, CNRS). For FEMTO-ST members, FAIR principles are ensured using the portal dat@ubfc. Data would be archived at dmarc.femto-st.fr supported by the FEMTO-ST lab IT service

• The LmB and IMB members will use the storage options available on their CNRS PLM accounts

Regarding exclusively mathematical results, a back-up of these data will be stored in computers of IMB and FEMTO-ST.

At least, new data produced by the consortium or update of already existing data will be uploaded to the secure datahub center **every month**. A global backup of consortium data will be archived **every six months**.

4 Legal and ethical requirements, codes of conduct

4.1 Patient data

Every patient's informed consent will be gained by CIC and FEMTO-ST when collecting the data. **Patient data collected by the consortium will be anonymised without exception**. Only the members of consortium belonging to the CIC will have access to the non-anonymised data (stored exclusively at the CIC), to perform the anonymization process. They could also access to the non-anonymised data for the purpose of patient follow-up, and only during for the duration of this follow-up.

These data will be first stored in the facilities provided by the IT support services of their own institution (CIC, CNRS) under the French legislation for clinical data. The raw data of the clinical studies will be stored in the hospital database, and the pre-treated data for the project will be transferred to the secured datahub. The secured datahub will be accessible by password via secured plugin only for authorised users (the members of the consortium). Regarding specifically the clinical data, a second level of encryption may be considered.

4.2 Data ownership

This topic is mostly covered by the consortium agreement (CA). Regarding new data, results or knowledge production, the CA distinguishes '*Résultats propres*', produced by only one partner institution of the consortium and '*Résultat communs*', produce by at least two partner institutions of the consortium.

Regarding the data ownership of the *Résultats propres*, the CA article 7.2 states: '*Les Résultats Propres sont la propriété de la Partie qui les a générés*' and regarding specifically new patents in the *Résultats propres: 'Les éventuels Brevets Nouveaux et les autres titres de propriété intellectuelle sur lesdits Résultats seroNote that software/libraries/packages versions would have to be added.nt déposés à ses seuls frais, à son seul nom et à sa seule initiative*'.

Regarding the data ownership of the *Résultats communs*, the CA article 8.2.3 states: '*Chaque Partie Copropriétaire (y compris ses Affiliées), après accord des autres Parties Copropriétaires, est libre d'Exploiter ou de faire Exploiter les Résultats Communs dont elle est copropriétaire conformément au règlement de copropriété ou d'indivision existant entre les Parties Copropriétaires*' and regarding specifically new patents in the *Résultats communs,* CA article 7.3.1.1 states: '*Les Parties Copropriétaires des Résultats Communs décideront si ces derniers doivent faire l'objet de demandes de brevet déposées à leurs noms conjoints, et désigneront parmi elles le mandataire qui sera chargé d'effectuer les formalités de dépôt et de maintien en vigueur. Elles pourront aussi décider de désigner un tiers pour effectuer ces formalités*'.

Regarding the use of data of *Résultats propres* by other partners or the use of *Résultat communs* by one of the partner, the CA article 8.1 states: '*Chaque Partie propriétaire ou Copropriétaire d'une Connaissance Propre, d'un Résultat Propre et/ou d'une quote-part de Résultat Commun peut librement en disposer aux fins de l'exécution du Projet. Pendant la durée du Projet, chaque Partie propriétaire ou Copropriétaire d'une Connaissance Propre, d'un Résultat Propre et/ou d'une quote-part de Résultat Commun concède, un droit d'Utilisation, non exclusif et non transférable, de ses Connaissances Propres sous réserve des droits des tiers, Résultats Propres et/ou quote-part de Résultats Commun aux autres Parties non-propriétaires ou non-Copropriétaires, sans contrepartie financière, sur demande écrite de celles-ci lorsqu'elles leur sont nécessaires pour exécuter leur Part du Projet.*'

Regarding the specific case of software as *Résultat Propre* or *Résultat commun*, the CA article 8.5.1 states: '*Pendant la durée du présent Projet, chaque Partie (co)propriétaire de Logiciel de Base, Dérivés ou Communs, nécessaire à une autre Partie pour l'exécution de sa Part du Projet, concède un droit d'Utilisation, non exclusif et non transférable, de ses Logiciels de Base, Dérivés ou Communs aux autres Parties non-(co)propriétaires, sans contrepartie financière, sur demande écrite de celles-ci lorsqu'elles leur sont nécessaires pour exécuter leur Part du Projet*'.

Regarding confidential data, CA article 9.1.1 states: '*Toutes les Informations Confidentielles et leurs reproductions, transmises par une Partie à une autre Partie, resteront la propriété de la Partie Emettrice sous réserve des droits des tiers et devront*

être restituées à cette dernière ou détruites sur sa demande, à l'exception d'une copie qui pourra être conservée à des seules fins d'archivage'.

All the published data will be open access, under Creative Commons licence : CC0 for the data and CC BY 4.0 for the documents.

4.3 Ethical issues and codes

Documents of the studies containing clinical information will be submitted to competent authorities and ethical committee. The research will be carried-out with respect to the French legislation in effect, in particular the sections regarding research involving a human person, as described in the French public health code, the laws of bioethics, laws regarding freedom of information, the Helsinki declaration, as well as the Good Clinical Practices.

Further guidance on ethical issues regarding research involving a human person, as well as for all other ethical issues related to data treatment, research and research integrity, is available through the "*Comité d'Éthique pour la Recherche (CER) d'Université Bourgogne Franche-Comté*" and their published "*Charte de fonctionnement".*

5 Data sharing and long-term preservation

5.1 Data storage and long-term preservation

All publications and their related data will be open access, on public repositories such as Github (for codes, the DOI being generated by Zenodo), ArXiv or BiorXiv (for preprints), under Creative Commons License. According to the UL open access policy, the consortium will choose editors that provide open access published articles.

Before publication, or regarding data for internal use only, they will be stored in the secure data center Biocore, Luxembourg, and will be accessible only to the members of the consortium.

Regarding the published data, it will be archived in the secure data center Biocore, Luxembourg, **for a duration of 10 years after publication**. Data will be accessible upon reasonable request if the public repository would not be accessible for some reason. The members of the consortium responsible for the data management, Gwenaël Rolin and Stéphane Urcun, will be in charge of maintaining this access. In the case of impossibility, their direct supervisors will be in charge.

Regarding the unpublished *Résultats propres* data, the institution owning the data will decide the duration of storage and assume the cost by its own means. Regarding the unpublished *Résultats communs* data, the duration of the storage will be decided unanimously by the board members, and the cost will be supported by the consortium.

5.2 Data accessibility

The data stored at secure data center Biocore, Luxembourg, will be accessible by Aspera, a browser plugin secured by password.

The published data will be accessible at public repositories such as Github, FigShare, ArXiv, HAL, ResearchGate or BiorXiv. To facilitate the re-use of these data, these public repositories provide a persistent identifier, such as DOI for ArXiv and BiorXiv.

6. Data management responsabilities and resources

6.1 Responsabilities

Each partner of the consortium is responsible for the monthly upload of its own new data on the secure data center. Each partner of the consortium is responsible for its own data production, quality, and metadata, including administrative metadata and structural links described section 2.1 .

The secure data center Biocore is responsible for the integrity and accessibility of the data of the consortium stored in its facilities. If needed, Biocore can provide a data stewart to help the members of the consortium to access, store and maintain their data.

The consortium is responsible for archiving and versioning of the global data structure stored in the Biocore facilities.

For any request related to the data management of the consortium, the members of the consortium may refer to Gwenaël Rolin and Stéphane Urcun, in charge of the coordination of the data management in the consortium.

6.2 Resources

Each member of the consortium should spend sufficient time for his/her own data curation. Gwenaël Rolin and Stéphane Urcun will spend up to 10% of their full time for the data management of the consortium.

Regarding the Biocore facilities, with the exception of specific demands that require new hardware material, the secure data center is free of charge.

Author: Stéphane Urcun wrote the first version of the DMP, the 24th of May 2022.

R. Eftimie updated the current second version of the DMP on 21st December 2023.