# ReproVIP project DMP

Sorina Pop, Axel Bonnet, Carole Frindel, Hélène Ratiney, Frédéric Cervenansky

# "Reproducibility with VIP" project DMP

*Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - DMP template (english)" fourni par Agence nationale de la recherche (ANR).*

## Plan Details

| | |
|---|---|
| **Plan title** | "Reproducibility with VIP" project DMP |
| **Version** | Final version |
| **Fields of science and technology (from OECD classification)** | Health sciences, Computer and information sciences |
| **Language** | eng |
| **Creation date** | 2022-06-08 |
| **Last modification date** | 2024-06-28 |
| **Identifier** | ANR-21-CE45-0024 |
| **Identifier type** | local identifier |
| **License** | **Name** Creative Commons Attribution 4.0 International<br>**URL** http://spdx.org/licenses/CC-BY-4.0.json |

## Project Details

| | |
|---|---|
| **Project title** | Reproducibility with VIP |
| **Acronym** | ReproVIP |
| **Abstract** | In the last few years, there has been a growing awareness of reproducibility concerns in many areas of science. In a recent study, the analysis of a single neuroimaging dataset by 70 independent analysis teams reveals substantial variability in reported results, with high levels of disagreement across teams of their outcomes on a majority of tested pre-defined hypotheses. Despite the increase in awareness and a growing number of projects tackling this lack of reproducibility and proposing various tools to improve it, researchers still lack an integrated, end to end solution, providing a good level of reproducibility at a reasonable effort. In this context, ReproVIP aims at evaluating and improving the reproducibility of scientific results obtained with the Virtual Imaging Platform (VIP) in the field of medical imaging. We will focus on a reproducibility level ensuring that the code produces the same result when executed with the same set of inputs and that an investigator is able to reobtain the published results. We will investigate reproducibility at three levels: (L1) the code itself, and in particular different versions of the same code, (L2) the execution environment, such as the operating system and code dependencies, parallel executions and the use of distributed infrastructures and (L3) the exploration process, from the beginning of the study and until the final published results. At Creatis, since 2011, we have developed and deployed VIP, a web portal for medical simulation and image data analysis. By effectively leveraging the computing and storage resources of the EGI federation,VIP offers its users high-level services enabling them to easily execute medical imaging applications on a large scale computing infrastructure. In 2021, VIP counts more than |

1200 registered users and about 20applications. In the last few years, VIP has addressed interoperability and reproducibility concerns, in the larger scope of a FAIR (Findable, Accessible, Interoperable, Reusable) approach to scientific data analysis. By implementing the CARMIN API and by using the Boutiques cross-platform framework for applications, VIP provides interoperability with existing platforms, which contributes to reproducibility. VIP provides us with a strong experience and a solid set of users and applications based on which we will tackle the lack of reproducibility L1, L2 and L3 described above. In order to reconstruct and interpret medical images, researchers make use of numerous image processing algorithms. Each processing step, from the raw image to the final decision, has its specific parameters and may come from a large number of different software packages and dependencies. As a result, the barrier to entry for non-expert users is high and can easily lead to processing pipelines quickly put together that are non-reproducible. Our final aim is to provide an integrated, end to end solution, allowing researchers to launch reproducible executions in a transparent manner. The proposed solutions for evaluating and improving reproducibility will be integrated in VIP and demonstrated on two scientific use-cases sharing a common set of processing tools for MRI image processing and addressing two different challenges: (i) optimising the MRI acquisition protocol w.r.t. to the signal to noise ratio (SNR) and (ii)optimising a processing pipeline for stroke prediction.

**Funding**

- French National Research Agency : ANR-21-CE45-0024

**Start date** 2022-02-01

**End date** 2024-01-31

**Partners**

- Institut Pluridisciplinaire Hubert Curien - IPHC (UMR 7178)
- Concordia University / Big Data Infrastructures for Neuroinformatics

**Research outputs :**

1. Preclinical data used in task 3.1: Optimizing the MRI acquisition protocol (Dataset)
2. UPENN brain images used in task 3.2: Optimizing a processing pipeline for skull base tumor segmentation (Dataset)

**Contributors**

| Name | Affiliation | Roles |
|---|---|---|
| POP Sorina | CENTRE DE RECHERCHE EN ACQUISITION ET TRAITEMENT DIMAGES POUR LA SANTE - 200717526Z | • DMP manager<br>• Personne contact pour les données (Preclinical data , Tumor segmentation)<br>• Project coordinator |

# "Reproducibility with VIP" project DMP

## Preclinical data used in task 3.1: Optimizing the MRI acquisition protocol

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

**1a. How will new data be collected or produced and/or how will existing data be re-used?**

In the first part of the project, we will re-use preclinical data previously acquired in a study financed by Neurodis. Data concerns 30 rats followed in time during 6 months.

In the second part of the project, we plan to acquire new preclinical data on healthy subjects.

All the data has been and will be acquired on the PILoT imaging platform at CREATIS, then stored in a Girder warehouse at CREATIS (https://pilot-warehouse.creatis.insa-lyon.fr/).

---

**1b. What data (for example the kind, formats, and volumes), will be collected or produced?**

For this use-case, data come mainly in two formats: (i) DICOM standard (including raw data and metadata) and (ii) a Bruker (constructor) proprietary format, containing a text (metadata) and a binary file.

Data acquired for one subject may include multiple acquisitions (at different times during 6 months) and may go up to a size of 1 GB. The data collected correspond to the following sequence acquisition: localized MR spectroscopy at short echo time (STEAM, TE 3ms) and DTI acquisition (EPI, 30 directions) in the rat brain at 11.7T

The processing of the data will involve the following format transformation: Bruker format to mrui file format and LCModel file (.RAW) format for the MR spectrosocpy data, Dicom to MRTRix3 file format (.mih)

Quantification results of spectroscopy data are stored in text and json file formats.

---

### 2. Documentation and data quality

**2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?**

Metadata are extracted from the DICOM headers, but are also collected at acquisition time based on information provided by the operator. They are stored both in text files and associated to data through the Girder metadata mechanism as described in the README available in the Girder collection.

These pieces of information, linked to imaging data, such as weight, age, duration of scan under anesthesia, will contribute to data production for reproducibility and reuse and participation in population imaging or replication studies in preclinical domain.

---

**2b. What data quality control measures will be used?**

Data quality will be assessed at processing time, using the usual metrics and assessments such as SNR and linewidth for spectra.

---

### 3. Storage and backup during the research process

### 3a. How will data and metadata be stored and backed up during the research?

CREATIS will provide and manage a database infrastructure specifically developed to conduct studies on medical imaging. The warehouse, based on the Girder technology, will allow unique, private, secure and selective access. This access to the warehouse will be available through the web and through a REST API. CREATIS expertise is based on previous developments that use the warehouse technology for different national and international cohorts and actions.

The back-up system will ensure the availability of data during the time of the project, with a backup on a weekly basis. The back-up procedure follows RSSI guidelines from CNRS, such as storing backups at two distinct locations.

---

### 3b. How will data security and protection of sensitive data be taken care during the research

The Girder warehouse provides secure access through a web interface and a RESTful API. Access rights are configured according to the data protection requirements of each data collection.

For this use-case, these is no sensitive data (subjects are rats).

---

## 4. Legal and ethical requirements, code of conduct

### 4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

For this use-case, there is no personal data (subjects are rats).

---

### 4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

The CNRS owns the data.

Data access is currently restricted to Neurodis and ReproVIP members. They will be distributed under a CC-BY license after the publication of the related research work currently ongoing.

---

### 4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

Ethical issues and code of conduct w.r.t. acquisition protocols on animal are handled by the PILoT imaging platform.

The project relies on two authorizations APAFIS#25081-2020050712321671 v1 and APAFIS#30495-2018121414354904 v6 respectively for the use of an animal model of multiple sclerosis and healthy rats, and respectively delivered the 11th of May 2020 and the 25th of march 2021 by the French ministry of education, research and innovation to collect data from small animal in vivo imaging.

---

## 5. Data sharing and long-term preservation

### 5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Data is stored and will be shared through the Girder web portal. They will be distributed under a CC-BY license after the publication of the related research work currently ongoing.

The scripts used for the processing of the data have been archived on Software Heritage :

swh:1:dir:c2557ab530729af3e49da720e1c625bf78dbaea2;origin=https://gitlab.in2p3.fr/reprovipgroup/isbi-spectro;visit=swh:1:snp:3da05cf4c0e01821b5e847940f3919c086d6f219;anchor=swh:1:rev:903c476ef242ce2e7ae55de901e8770b32c0f51c

---

**5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?**

Data is stored on the Creatis Girder warehouse (https://pilot-warehouse.creatis.insa-lyon.fr/).
We will keep all raw/acquired data for a minimum duration of 5 years.

---

**5c. What methods or software tools are needed to access and use data?**

Data is stored on the Creatis Girder warehouse (https://pilot-warehouse.creatis.insa-lyon.fr/).
Data stored in the Girder warehouse is accessible through a web interface and a RESTful API.

---

**5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?**

This is work in progress, but we envisage to associate a DOI to data collections stored on Girder.

---

## 6. Data management responsibilities and resources

**6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?**

Frédéric Cervenansky: Research Engineer at CREATIS and administrator of the Girder platform.

---

**6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?**

The dedicated warehouse will be managed by Frédéric Cervenansky, member of the ReproVIP project. This activity has been planned in the project and we estimate it at approximately 16 work hours. The warehouse and backups are services provided by the CREATIS laboratory for internal projects at no additional cost.

# UPENN brain images used in task 3.2: Optimizing a processing pipeline for skull base tumor segmentation

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

**1a. How will new data be collected or produced and/or how will existing data be re-used?**

This use case will use existing data from two sources:

- the **BRATS challenge** in the form of a pre-trained model: https://cbica.github.io/CaPTk/seg_DL.html
- **UPENN images** https://www.cancerimagingarchive.net/collection/upenn-gbm/ as input of the Brats pre-processing pipeline, to test the transfer of the model to new pre-processed data according to different versions of a processing pipeline.

---

**1b. What data (for example the kind, formats, and volumes), will be collected or produced?**

The original UPENN-GBM dataset (pointed in 1a.) consists of 630 patients diagnosed with de novo gliobastoma. Each patient has multi-parametric magnetic resonance imaging scans, including four types of structural MRI scans: native T1 weighted (T1), post-contrast T1 (T1-GD), native T2-weighted (T2), and T2 fluid attenuated inversion recovery (T2-FL) scans.

From this dataset, we selected a subset of 191 complete patients and used the four raw images provided in DICOM format, which we then converted to NIfTI format required for the Brats pre-processing pipeline. The resulting dataset corresponds to a volume of 4.8 Go and is available on the UPENN-GBM-ReproVIP Girder collection (https://myriad.creatis.insa-lyon.fr).

---

## 2. Documentation and data quality

**2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?**

The UPENN-GBM-ReproVIP Girder collection provides a short description pointing both to the original data and to the GitLab repository used for the selection of the 191 patients along with the Dicom to NIfTI conversion.

---

**2b. What data quality control measures will be used?**

We verified the original data w.r.t. completeness (all image types) and duplicates through https://gitlab.in2p3.fr/MDL/reprovip-wp3-use-case/-/blob/master/formating_dataset.ipynb#6-remove-patients-that-have-duplicates-mistales-or-are-incomplete

---

## 3. Storage and backup during the research process

**3a. How will data and metadata be stored and backed up during the research?**

CREATIS will provide and manage a database infrastructure specifically developed to conduct studies on medical imaging. The warehouse, based on the Girder technology, will allow unique, private, secure and selective access. This access to the warehouse will be available through the web and through a REST API. CREATIS expertise is based on previous developments that use the warehouse technology for different national and international cohorts and actions.

The back-up system will ensure the availability of data during the time of the project, with a backup on a weekly basis. The back-up procedure follows RSSI guidelines from CNRS, such as storing backups at two distinct locations.

---

**3b. How will data security and protection of sensitive data be taken care during the research**

For this use-case, we re-used data that were already anonymized.

The Girder warehouse provides secure access through a web interface and a RESTful API : https://myriad.creatis.insa-lyon.fr

## 4. Legal and ethical requirements, code of conduct

**4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?**

This research study was conducted retrospectively using human subject data made available in open access from the publicly available repository of The Cancer Imaging Archive at https://doi.org/10.7937/TCIA.709X-DN49.
Ethical approval was not required as confirmed by the license attached with the open access data.

**4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?**

This research study was conducted retrospectively using human subject data made available in open access from the publicly available (CC-BY 4.0 licence) repository of The Cancer Imaging Archive at https://doi.org/10.7937/TCIA.709X-DN49.

**4c. What ethical issues and codes of conduct are there, and how will they be taken into account?**

This research study was conducted retrospectively using human subject data made available in open access from the publicly available repository of The Cancer Imaging Archive at https://doi.org/10.7937/TCIA.709X-DN49.
Ethical approval was not required as confirmed by the license attached with the open access data.

## 5. Data sharing and long-term preservation

**5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?**

We store and share openly through the UPENN-GBM-ReproVIP Girder collection the 191 Nifti files we produced out of the original UPENN-GBM dataset(CC-BY 4.0 licence).
The model is available in the docker images provided by the CBICA team at https://hub.docker.com/r/cbica/captk.
The scripts used for the processing of the data have been archived on Software Heritage :
swh:1:dir:fbe26c4e306667fba4ab9e8280318043cd732ab6;origin=https://gitlab.in2p3.fr/reprovipgroup/tumor-segmentation-use-case;visit=swh:1:snp:ac2249b9f90c7f6f47df20e4407f271f61c138e4;anchor=swh:1:rev:eedb2d1a70f2d98b9d0371b8279108e6cbb3a183

**5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?**

Data is stored on the Creatis Girder warehouse and preserved for a minimum duration of 5 years.

**5c. What methods or software tools are needed to access and use data?**

Data is stored on the Creatis Girder warehouse.

Data stored in the Girder warehouse is accessible through a web interface and a RESTful API.

**5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?**

Data is stored on the Creatis Girder warehouse.
This is work in progress, but we envisage to associate a DOI to data collections stored on Girder.

## 6. Data management responsibilities and resources

**6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?**

Frédéric Cervenansky: Research Engineer at CREATIS and administrator of the Girder platform.

**6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?**

The dedicated warehouse will be managed by Frédéric Cervenansky, member of the ReproVIP project. This activity has been planned in the project and we estimate it at approximately 16 work hours. The warehouse and backups are services provided by the CREATIS laboratory for internal projects at no additional cost.