DMP du projet "PAPPSO Proteomics facility"

Mélisande Blein, Olivier Langella, Filippo Rusconi

DMP du projet "PAPPSO Proteomics facility"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "INRAE - Trame Structure" fourni par INRAE - Institut national de recherche pour l'agriculture l'alimentation et l'environnement.

Renseignements sur le plan

Titre du plan	DMP du projet "PAPPSO Proteomics facility"
Version	version 1
Domaines de recherche (selon classification de l'OCDE)	Biological sciences (Natural sciences)
Langue	fra
Date de création	2021-03-17
Date de dernière modification	2021-09-15
Type d'identifiant	DOI
Licence	Creative Commons Attribution Non Commercial 4.0 International

Renseignements sur le projet

Titre du projet	PAPPSO Proteomics facility
Sources de financement	• INRAE - Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement :

Produits de recherche :

1. Raw data produced by the mass spectrometer (Jeu de données)
2. Scientific software projects developped by PAPPSO (Logiciel)

Contributeurs

Contributeurs	Affiliation	Rôles
Olivier Langella	CNRS	• Personne contact pour les données (Scientific software, Raw data)
Mélisande Blein-Nicolas	INRAE	• Coordinateur du projet • Responsable du plan de gestion de données

DMP du projet "PAPPSO Proteomics facility"

Informations sur la structure

Nom de la structure

Plateforme d'Analyse Protéomique de Paris Sud-Ouest (PAPPSO proteomics facility)

Responsabilités dans la structure

Nom, Prénom	Courriel	Rôle
Blein-Nicolas, Mélisande	melisande.blein-nicolas@inrae.fr	Responsable scientifique
Henry, Céline	celine.henry@inrae.fr	Responsable technique
Langella, Olivier	olivier.langella@universite-paris-saclay.fr	Responsable bioinformatique

Etablissement(s) tutelle(s)

Université Paris Saclay, INRAE, CNRS, AgroParisTech

Financeur(s) (*permettant l'acquisition des jeux de données – hors projet)*

PAPPSO receives funding from various sources mainly including INRAE, IBiSA, the region of Île-de-France and the University Paris-Saclay. Members of the PAPPSO team are staff members from INRAE and CNRS.

Informations sur le plan de gestion

Historique des versions

Date	n° de version	Status	Auteur	Affiliation de l'auteur *(se reporter à l'annuaire INRAE)*	Validé par	Validé le
16/06/2021	1	unpublished	Mélisande Blein-Nicolas	UMR Génétique Quantitative et Evolution - Le Moulon		

Présentation générale des données

Raw data produced by the mass spectrometer

Mode d'obtention des données

- Données générées par la structure

Raw data are unprocessed data output by the instruments (mainly mass spectrometers) hosted at PAPPSO.
We convert these raw data into an open format for mass spectrometry. The files output by the conversion process (formats: mzXML or mzML) are also stored without modification of the mass data in them.

Origine

- Observation

Instruments hosted at PAPPSO.

Nature des données

Proteomics data (LC-MS/MS analyses on protein extracts)

Format des données

.tdf (Bruker's format)
.raw (Thermo's format)
.mzXML (open format for raw data)
.mzML (open format for raw data)

Scientific software projects developped by PAPPSO

Mode d'obtention des données

- Données générées par la structure
- Données produites par un tiers

As the codes of our software are open, third parties can contribute to them and add new functionalities.

Type de données

- Software

The softwares that we develop are:

- X!TandemPipeline implemented in C++ for the peptide identification and protein inference
- MassChroq implemented in C++ for peptide quantification
- MCQR implemented in R for data analysis
- msXpertSuite implemented in C++ for modelling, simulating and analyzing ionized flying species

Nature des données

programming language (C++, Python, R, perl, Java, Javascript, PHP...)

Format des données

text files

Droits de propriété intellectuelle

Raw data produced by the mass spectrometer

Qui détiendra les droits sur les données et les autres informations créées ?

The users and their respective institutions are the owners of the produced data.

Scientific software projects developped by PAPPSO

Qui détiendra les droits sur les données et les autres informations créées ?

The institutions that paid for the development of the code have the intellectual property rights (mainly INRAE and CNRS).

Confidentialité

Raw data produced by the mass spectrometer

Identification des jeux de données contenant des données confidentielles

PAPPSO guarantees the confidentiality of all the datasets produced, as described in its charter. The decision of data publication is taken with the users.

Quelles sont les mesures prises et les normes auxquelles il est nécessaire de se conformer pour garantir cette confidentialité ?

Access to the instruments is regulated and monitored. The data are directly transmitted to the storage cluster through a secure shell (ssh) link. The server is protected by a firewall filtering connections by IP number ranges and accessible only by encrypted ssh key to the PAPPSO agents. Even if the desktop and laptop computers of the PAPPSO agents are encrypted (following the recommendations of the CNRS under the responsibility of the CSSI), it is not allowed to copy data elsewhere than on the cluster. All PAPPSO agents sign a charter for the use of computer resources specifying the confidentiality of all data produced. In the event of an incident, the data is recovered from the PAPPSO backup servers or from the regularly archived LTO tapes.

Scientific software projects developped by PAPPSO

Identification des jeux de données contenant des données confidentielles

The codes of our softwares are not confidential, but the master branches of their repositories are restricted for some actions.

Quelles sont les mesures prises et les normes auxquelles il est nécessaire de se conformer pour garantir cette confidentialité ?

The code of our software are versioned on the GitLab forge hosted by INRAE's MATHNUM department forgemia. We use the application's functionalities to manage the repositories access and grants.

Le cas échéant, comment la confidentialité de données fournies par des personnes sera garantie lorsque les données seront partagées ou rendues disponibles pour une analyse de second niveau ?

Not relevant

Partage des données

Raw data produced by the mass spectrometer

Y a t'il une obligation de partage (ou à l'inverse une interdiction ou une restriction) ?

The user is committed to publicly open the data. An embargo period can be defined with respect to the scientific community usage. The decision of data publication is taken with the user.

Quelles sont les réutilisations potentielles de ces données ?

- Re-use for new research purposes
- Re-use to apply different methods of analysis
- Support for public policy, expertise
- Trainings

La lecture des données nécessite-t-elle le recours à un logiciel ou un outil spécifique ? Si oui, lequel ?

- Bruker data (directory-based raw data format): readable by DataAnalysis software (Bruker) or free and open source PAPPPSO's implementation (PAPPSOms++ library);
- Thermo Fisher Scientific raw data: readable by Qual Browser (Thermo) or Free Style (Thermo);
- mzXML or mzML: readable by the mineXpert2, mzMine2 free and open source software or any other proteomics open source software.

Comment les données seront-elles partagées ?

The data will be deposited in the reference repositories like MASSIVE or PRIDE, on the Data INRAE portal.

Avec qui ?

- Autre
- Public data are shared with anyone (open access)
- Private data are shared with identified users (academic and/or private partners)

Sous quelle licence ?

- Autre (à préciser dans la zone d'Informations supplémentaires

A DOI will be automatically associated to the data deposited on reference repositories like MASSIVE or PRIDE. Otherwise, we encourage the users to associate a DOI to the dataset with INRAE's open data repository service. It allows to discuss the license associated to the data. By default, public data are under the CC-BY V4.0 license. We can provide support if needed.

Scientific software projects developped by PAPPSO

Y a t'il une obligation de partage (ou à l'inverse une interdiction ou une restriction) ?

All our software are licensed under the GNU GPL v3. Any modification is allowed as long as the modified code is available under the same conditions.

Quelles sont les réutilisations potentielles de ces données ?

- Improvement of existing functionalities
- Addition of new functionalities to answer new needs

La lecture des données nécessite-t-elle le recours à un logiciel ou un outil spécifique ? Si oui, lequel ?

Any software produced by PAPPSO only requires publicly available standard compilers, runtimes, operating systems to run.

Comment les données seront-elles partagées ?

All our softwares are versioned on private repositories of the GitLab forge forgemia and are available publicly or upon request.

Sous quelle licence ?

- Autre (à préciser dans la zone d'Informations supplémentaires

All our sofwares are under the GNU General Public Licence version 3.

Organisation et documentation des données

Raw data produced by the mass spectrometer

Quels méthodes et outils sont utilisés pour acquérir et traiter les données, depuis leur acquisition jusqu'à leur mise à disposition, leur archivage ou leur destruction ?
Utiliser éventuellement un lien vers un schéma illustrant les processus

Raw data are acquired by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). If necessary, they may be converted to the mzXML/mzML formats by using the msConvert software. They can be visualized and explored by using mineXpert2. Raw data can be processed by using X!TandemPipeline for peptide identification, protein inference and protein quantification based on spectral counts. When processed by using MassChroQ, together with identification results, they allow to produce peptide quantification results. Finally, quantification results can be analysed and interpreted by using MCQR. A schematic overview of the data processing is available here.

The raw data produced by the mass spectrometers (proprietary format) are archived on LTO8 tape in two copies, one stored on the Jouy site, the other on the Moulon site. The files transformed to open formats (mzXML or mzML) as well as all subsequent analysis data are stored on a distributed, fault-tolerant and resilient file system (ceph) (erasure code). To prevent manipulation errors, the data is archived with BURP software in multiple copies every day on dedicated servers (RAID5), guaranteeing a storage time of one month. The PAPPSO storage system is hosted in our secure computer room at IDEEV (controlled access, air conditioning, fire protection system). At the end of the analyses, the data stored on the cluster are archived on LTO8 tapes in 2 copies (one in Jouy, the other in IDEEV). The content of the tapes is referenced and the tapes are labelled.

Quelles métadonnées seront utilisées pour accompagner le jeu de données ? Quels seront les standards, vocabulaires, taxonomies... utilisés pour décrire et représenter les données et éléments de métadonnées ? Comment les métadonnées seront-elles produites et mises à jour ?

Metadata	Metadata origin and production mode (e.g. manual input, automatic annotation...)	Standard and associated vocabularies	Conditions or frequency update (if applicable) (e.g. change in accessibility)
Experimental design	manual input	text tabulated file format	
Protein sequence data	automatic annotation from genome sequence or RNAseq information	FASTA files	Latest available version at the moment of the analysis
Phenotyping experiment	manual input	Minimal Informations About Plant Phenotyping Experiments (MIAPPE)	
Protein functionnal annotation	automatic or manual annotations from trEMBL, SwissProt, mapMan, KEGG, GO	GO, KEGG annotations	Latest available version at the moment of the analysis

Une documentation complémentaire aux métadonnées est-elle nécessaire pour décrire les données et assurer leur réutilisabilité sur le long terme ?

Metadata	Documentation
Experimental design	self explanatory (biological or technical repetitions, experimental conditions...)
Protein sequence data	Publicly available databases
Phenotyping experiment	Publicly available biological ontologies
Protein functionnal annotation	Publicly available ontologies

Comment les fichiers de données sont-ils gérés et organisés : contrôle des versions, conventions de nommage des fichiers, organisation des fichiers

By essence, the raw data files need no version control system, because these files are only stored and never modified. Raw data files are stored on our file system using a naming convention (by site, instrument, date, experiment). This complete path is conserved from the initial acquisition to the final archive location. The intermediate files (xpip files, for example) are so-called "generated" data files, which means that they can be regenerated using the same raw data set and the same software program. We do store these intermediate files along with the raw data files. Generated files contains the name and version of software used to process data.

Quel est le processus de contrôle qualité des données ?

The quality of the proteomics data is controlled at several stages: a visual control of the chromatograms as well as numerical indicators (number of identified peptides and proteins, percentage of assigned spectra) allow to ensure that the data were correctly acquired by the mass spectrometer. In addition, graphical representations such as distribution density, boxplot or PCA are used to check the alignment of retention times, the dispersion and distribution of data in each sample, the variability between samples.

Scientific software projects developped by PAPPSO

Quels méthodes et outils sont utilisés pour acquérir et traiter les données, depuis leur acquisition jusqu'à leur mise à disposition, leur archivage ou leur destruction ?

Utiliser éventuellement un lien vers un schéma illustrant les processus

Standard text editor to produce programming language code.

Quelles métadonnées seront utilisées pour accompagner le jeu de données ? Quels seront les standards, vocabulaires, taxonomies... utilisés pour décrire et représenter les données et éléments de métadonnées ? Comment les métadonnées seront-elles produites et mises à jour ?

Our source code is documented using standards : Doxygen for C++, R documentation...

Une documentation complémentaire aux métadonnées est-elle nécessaire pour décrire les données et assurer leur réutilisabilité sur le long terme ?

Not relevant

Comment les fichiers de données sont-ils gérés et organisés : contrôle des versions, conventions de nommage des fichiers, organisation des fichiers

Version control is provided by git. We use general naming convention for R code as well as the directory name scheme.
We use the debian package policy for other sofware : each file is described and stored in standard places.

Quel est le processus de contrôle qualité des données ?

Our development process is test driven. Our debian packages can only be released if tests are successful.
We also use test datasets and controls are made for each software version to guarantee reliability.

Stockage et sécurité des données

Raw data produced by the mass spectrometer

Quels sont les types de flux empruntés par les données et les supports utilisés pour les stocker ?
(Faire éventuellement un lien vers un schéma)

* Raw data are acquired on dedicated work stations using conventional hard disks (RAID) into data files.
* Raw data files are transfered using ssh encrypted connection to the servers in IDEEV computer room. File storage is distributed on a ceph cluster (distributed file system). Ceph guarantees reliability using erasure code (protection against hardware failures) and performance.
* Raw data files are converted into Open data formats (mzML, mzXML) on the data storage system.
* So called "generated files", processed data from raw data files are also produced and stored on the ceph data storage system.
* To ensure data safety, data files backups are made every day through dedicated RAID file servers.
* Finally, raw data files and generated files are archived in the long term onto LTO8 tapes

Quelle est la volumétrie actuelle et prévisionnelle ?

In 2022, annual data production is 20Tb. Previsional production for 2024 is 30Tb.

L'entité hébergeant physiquement les données a-t-elle une politique de sécurité pour son système d'information ?
politique locale, charte des infrastructures de recherche...

Charte des infrastructures de recherche à l'Inra

Sécurité - Confidentialité : les données font-elles l'objet d'échange ou de partage avec de tiers acteurs et selon quelles modalités ? comment sont déterminés les droits d'accès aux données avant leur publication ?

As already discussed, all data produced by PAPPSO are confidential, only available to PAPPSO members.
Upon demand from data owners, raw data or generated files are sent directly or uploaded to a data repository for Open Access purpose or confidential access if needed.

Sécurité - Intégrité - Tracabilité : Quelles sont les mesures de protection mises en œuvre pour suivre la production et l'analyse des données ?

PAPPSO is certified ISO 9001. Our production processes are traced in our Quality Management System.

Scientific software projects developped by PAPPSO

Quels sont les types de flux empruntés par les données et les supports utilisés pour les stocker ?
(Faire éventuellement un lien vers un schéma)

Git repositories are used on local file systems and then pushed to the central git repository in ForgeMIA gitlab.

Quelle est la volumétrie actuelle et prévisionnelle ?

500Mb maximum per year.

Sécurité - Confidentialité : les données font-elles l'objet d'échange ou de partage avec de tiers acteurs et selon quelles modalités ? comment sont déterminés les droits d'accès aux données avant leur publication ?

Source code are available publicly under the terms of the GPLv3. Access is granted by the ForgeMIA Gitlab repository.

Sécurité - Intégrité - Tracabilité : Quelles sont les mesures de protection mises en œuvre pour suivre la production et l'analyse des

données ?

git system and debian quality assurance team.

Archivage et conservation des données

Raw data produced by the mass spectrometer

Quelles sont les données à conserver sur le moyen ou le long terme et quelles sont les données à détruire ?

All the elaborated and derived data are preserved in the long term as well as the useful raw data (for a later re-exploitation).
Only temporary files needed to generate elaborated data can be destroyed.

Sur quelle plateforme d'archivage pérenne seront archivées les données à conserver sur le long terme ? Sinon, quelles procédures seront mises en place pour la conservation à long terme ?

The data stored are archived on LTO8 tapes in 2 copies (one in Jouy, the other in Le Moulon). The content of the tapes is referenced and the tapes are labelled.

Quelle est la durée de conservation des données ?

At least 2 years.

Quelles garanties de financements couvriront les coûts associés à la conservation à long terme ?

Long term conservation is guaranteed as long as PAPPSO is financially supported.

Scientific software projects developped by PAPPSO

Quelles sont les données à conserver sur le moyen ou le long terme et quelles sont les données à détruire ?

all source codes and all versions are kept.

Sur quelle plateforme d'archivage pérenne seront archivées les données à conserver sur le long terme ? Sinon, quelles procédures seront mises en place pour la conservation à long terme ?

Long term support is provided by the ForgeMIA.

All our software are also conserved by the Debian project.

Quelle est la durée de conservation des données ?

no limit

Quelles garanties de financements couvriront les coûts associés à la conservation à long terme ?

Financial support is guaranteed by institutional support of ForgeMIA and international support of the Debian project.