

Formation

Connaissance et prise en main des outils de traitement de données

0.1 Sommaire

Déroulement demi-journée :

- Présentation Cerise
- Bonnes pratiques
- Droits/Habilitations
- Assistance
- Onyxia/SSP Cloud

0.2 Avant-propos

Ce diaporama de formation a été rédigé dans le but d'être le support visuel des formations dispensées au [MASA](#).

Cette formation s'adresse à tous les nouveaux arrivants au SSM Agriculture qui seront amenés à manipuler des données sous Cerise ou sous Onyxia. Elle est dispensée en distanciel sur une **demi-journée**.

Ce support ne se substitue pas [aux formations R dispensées par les formateurs du MASA](#).

Il permet aux nouveaux agents ayant déjà pratiqué R dans un autre contexte de **découvrir les spécificités de Cerise et d'en faire un bon usage**.



MINISTÈRE
DE L'AGRICULTURE
ET DE LA SOUVERAINETÉ
ALIMENTAIRE

*Liberté
Égalité
Fraternité*

1 Présentation



1.1 C'est quoi Cerise ?

CERISE : Consolidation Et Restitution de l'Information Statistique

Cerise contient l'ensemble des données et des programmes R utilisés par le SSM Agriculture.

Cerise est une plateforme qui repose sur la solution “Posit Workbench” commercialisée par la société du même nom “Posit”. Elle offre une interface web pour utiliser RStudio dans un environnement multi-utilisateurs et sécurisé.

Cerise est articulé autour de 3 machines virtuelles (VM) accessibles via votre navigateur :

- VM1 : accès général à tous les agents du SSM
- VM2 : accès réservé Demesis et traitements gourmands en ressources
- VM3 : accès réservé aux opérations particulières (RA / RICA ...)

URL à faire figurer dans vos favoris : <https://rstudio.agriculture.rie.gouv.fr/>

1.2 Avantages de Cerise

Cerise présente plusieurs avantages :

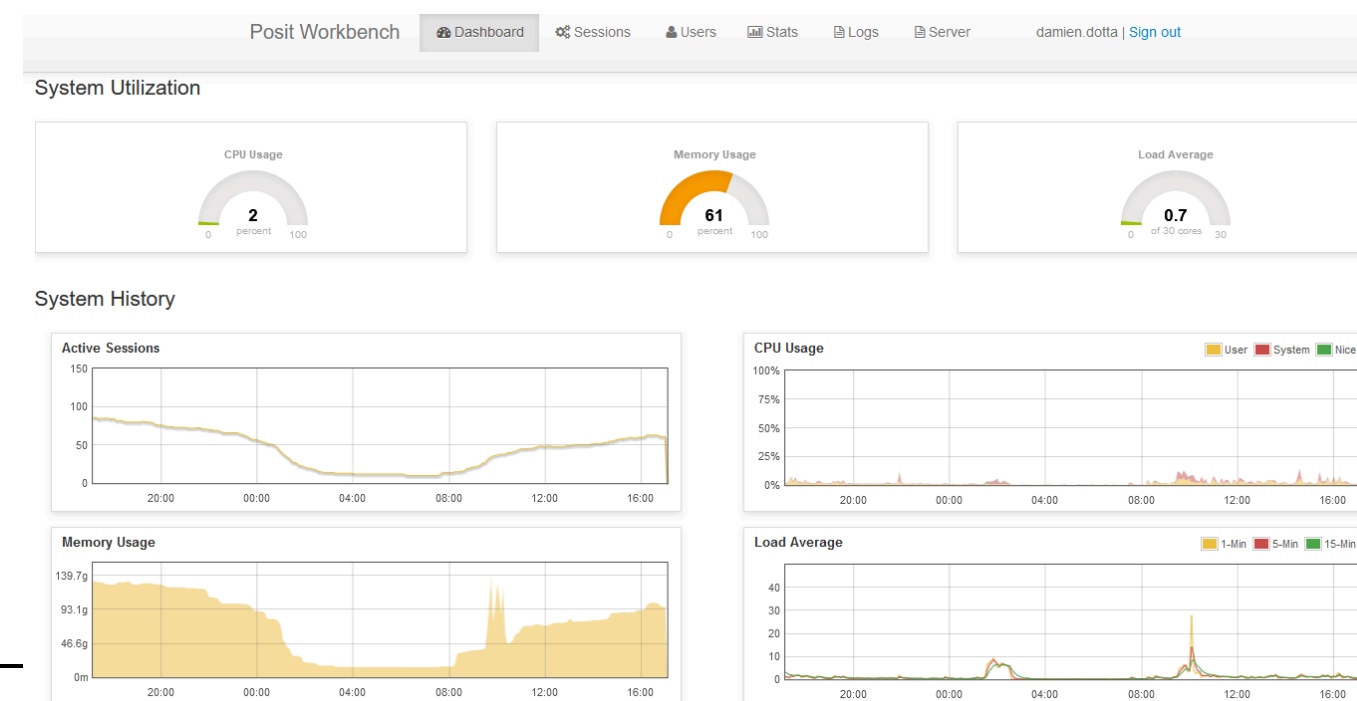
- Accès aux données du SSM Agriculture (enquêtes, nomenclatures, fichiers externes...)
- Montée de versions régulière (Workbench / R / Packages)
- Sécurité /conformité
- Support Centre de Service et assistance BQIS garantis



1.3 Ressources Cerise

- 160 Go RAM
- 30 processeurs virtuels
- Plus de 400 utilisateurs distincts
- 80 sessions simultanées en moyenne

Des admins Cerise qui vous surveillent 😊



1.4 Mises à jour Cerise

- **R**
 - Une **montée de version** annuelle de R_Base
Le DéMéSIS met à disposition 2 versions de R_base :
 - **Suppression de la version la plus ancienne présente l'année N-1 sur CERISE**
 - **Reconduction de la version la plus récente présente l'année N-1 sur CERISE**
 - **Ajout de la dernière version stable mise à disposition sur le CRAN**
- **RStudio**

Dernière version stable mise à disposition par l'éditeur (au moment de la réalisation de septembre/octobre de l'année N-1)
- **Packages**
 - La liste des packages R valides et compatibles à une date donnée sont mis à disposition dans le Data Center du MASA.
 - Une liste pré-établie de packages est préinstallée sur CERISE pour chacune des version R_Base

1.5 Cerise en 2025

Composant	Version actuelle
RStudio	2024.09 « Cranberry Hibiscus »
R_Base_Core	R 4.2.3 – 2023-03-15 R 4.4.1 – 2024-06-15
Packages	R 4.2.3 – 2023-03-15 Nombre de packages disponibles : 18839 Nombre de packages installés : 770 R 4.4.1 – 2024-06-15 Nombre de packages disponibles : 20946 Nombre de packages installés : 80

1.6 Organisation de Cerise (1/2)

Cerise est organisé en plusieurs répertoires :

- **00-Espace-Personnel** => espaces personnels des agents, accessible par l'agent uniquement
- **01-Espace-de-Partage** => lieu de partage général (programmes/formation/outils...) entre les différents acteurs
- **02-Espace-de-Production** => plateforme de stockage des données brutes collectées, ainsi que des fichiers de données et programmes issus des traitements statistiques réalisés par l'équipe projet (voir image plus loin)
- **03-Espace-de-Diffusion** => mise à disposition au sein du SSM des données issues des traitements statistiques réalisés en amont

=> Ces deux derniers espaces sont découpés par opérations statistiques

- **04-Espace-Echanges** => stockage des fichiers de données à transmettre aux autres applications du SI CASSIS (par exemple Agreste) ainsi qu'aux SI des partenaires extérieurs

1.7 Organisation de Cerise (2/2)

FilesPlotsPackagesHelpViewerPresentation

+ New Folder

+ New Blank File

+ Upload

✖ Delete

➡ Rename

⚙ More

↺

Home > CERISE

R

...

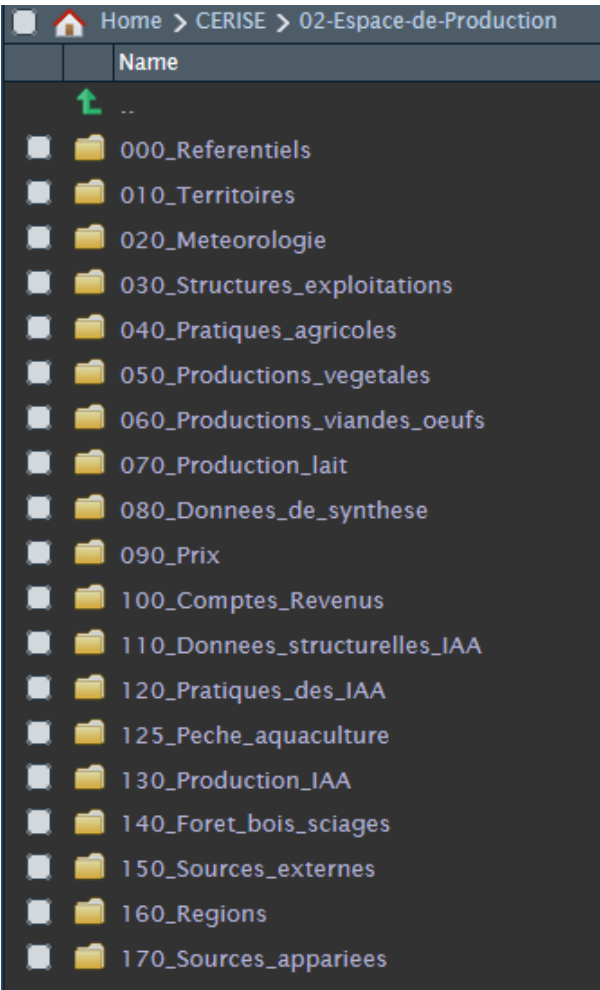
	Name	Size	Modified
↑	..		
📁	00-Espace-Personnel		
📁	01-Espace-de-Partage		
📁	02-Espace-de-Production		
📁	03-Espace-de-Diffusion		
📁	04-Espace-Echanges		

1.8 Focus sur l'espace de production (1/2)

- 1er niveau par **rubriques** (19 rubriques existantes)
 - 2ème niveau par **sources**
 - 3ème niveau par sous-répertoires millésimés

Exemple :

```
1 070_Production_lait/  
2 070_Production_lait/7010_Conj_lait  
3 070_Production_lait/7010_Conj_lait/EML_2018  
4 070_Production_lait/7010_Conj_lait/EML_2019  
5 070_Production_lait/7010_Conj_lait/EML_2020  
6 070_Production_lait/7010_Conj_lait/EML_COLLECTE  
7 070_Production_lait/7010_Conj_lait/EML_ESTIM  
8 070_Production_lait/7010_Conj_lait/Programmes  
9 ...
```



1.9 Focus sur l'espace de production (2/2)

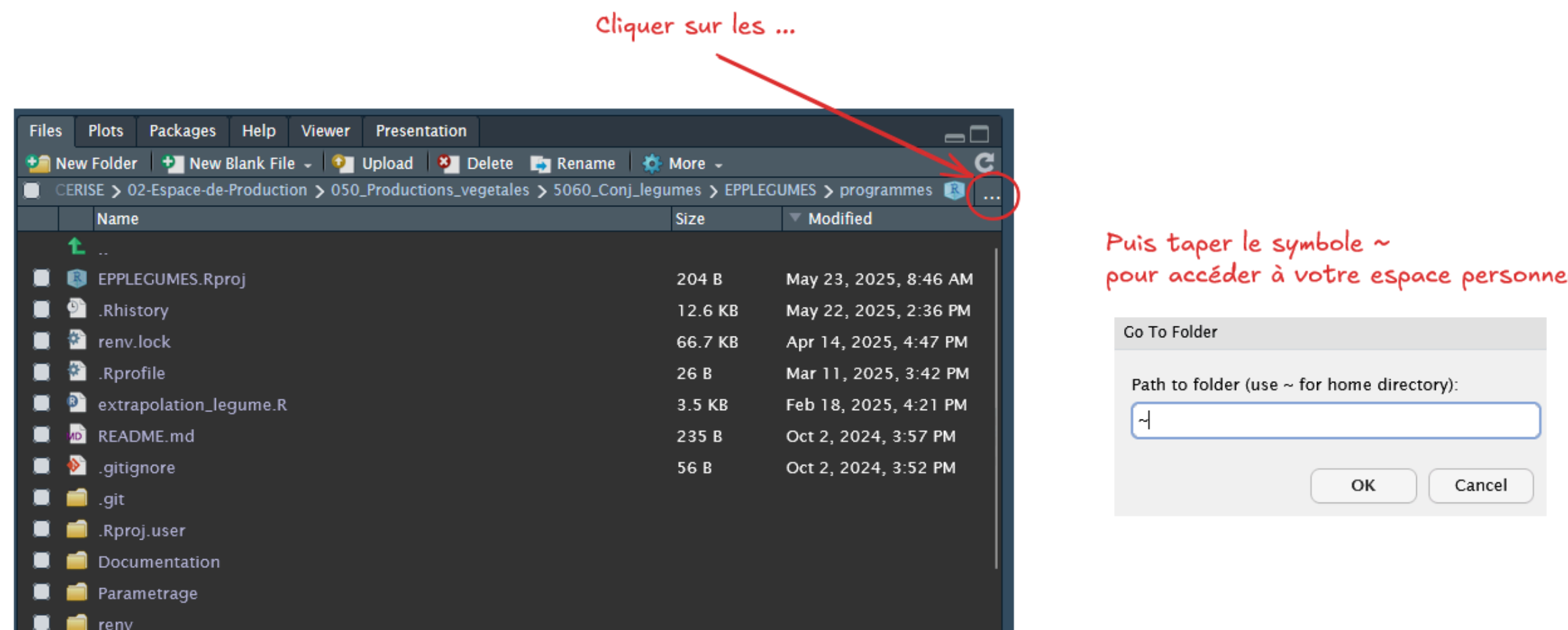
Les différents groupes d'habilitations disposent des droits suivants :

- Utilisateur standard : aucun
- Chargé d'études : lecture
- Producteur de données : lecture/écriture
- Administrateur : lecture/écriture



1.10 Accéder facilement à son espace personnel

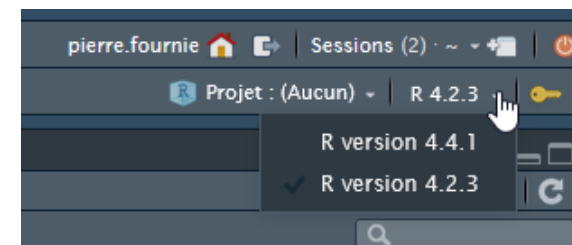
L'accès à l'espace personnel de Cerise peut être difficile lorsque vous êtes “perdus” dans l'arborescence riche de Cerise. Voici ci-dessous comment faire pour y accéder rapidement.



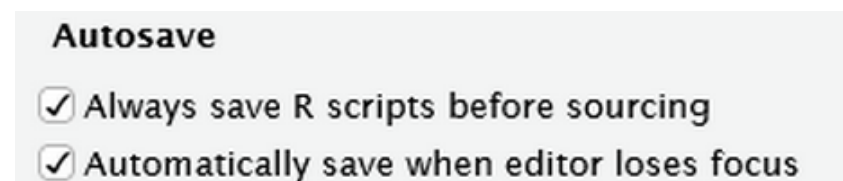
Remarque : soyez économe dans l'utilisation de votre espace personnel. A ne réservez que pour des expérimentations. ~~Pas de recopie de données...~~

1.11 Remarques et astuces sur Cerise

- Pas d'explorateur de fichiers comme Windows sous Cerise. La navigation dans l'arborescence se fait via le menu “Fichiers” ou “Files” de RStudio.
- Pour sélectionner une version R_Base, il faut la sélectionner en haut à droite dans la fenêtre RStudio



- Pour éviter le risque de perdre du code R pendant une interruption Cerise, il est recommandé de cocher ces 2 cases accessibles dans le menu de RStudio > Outils > Options globales > Onglet Sauvegarder





MINISTÈRE
DE L'AGRICULTURE
ET DE LA SOUVERAINETÉ
ALIMENTAIRE

*Liberté
Égalité
Fraternité*

2 Bonnes pratiques



2.1 Cerise, un espace partagé

Comme tout espace partagé et mutualisé, il convient d'être économe en ressources sur Cerise.

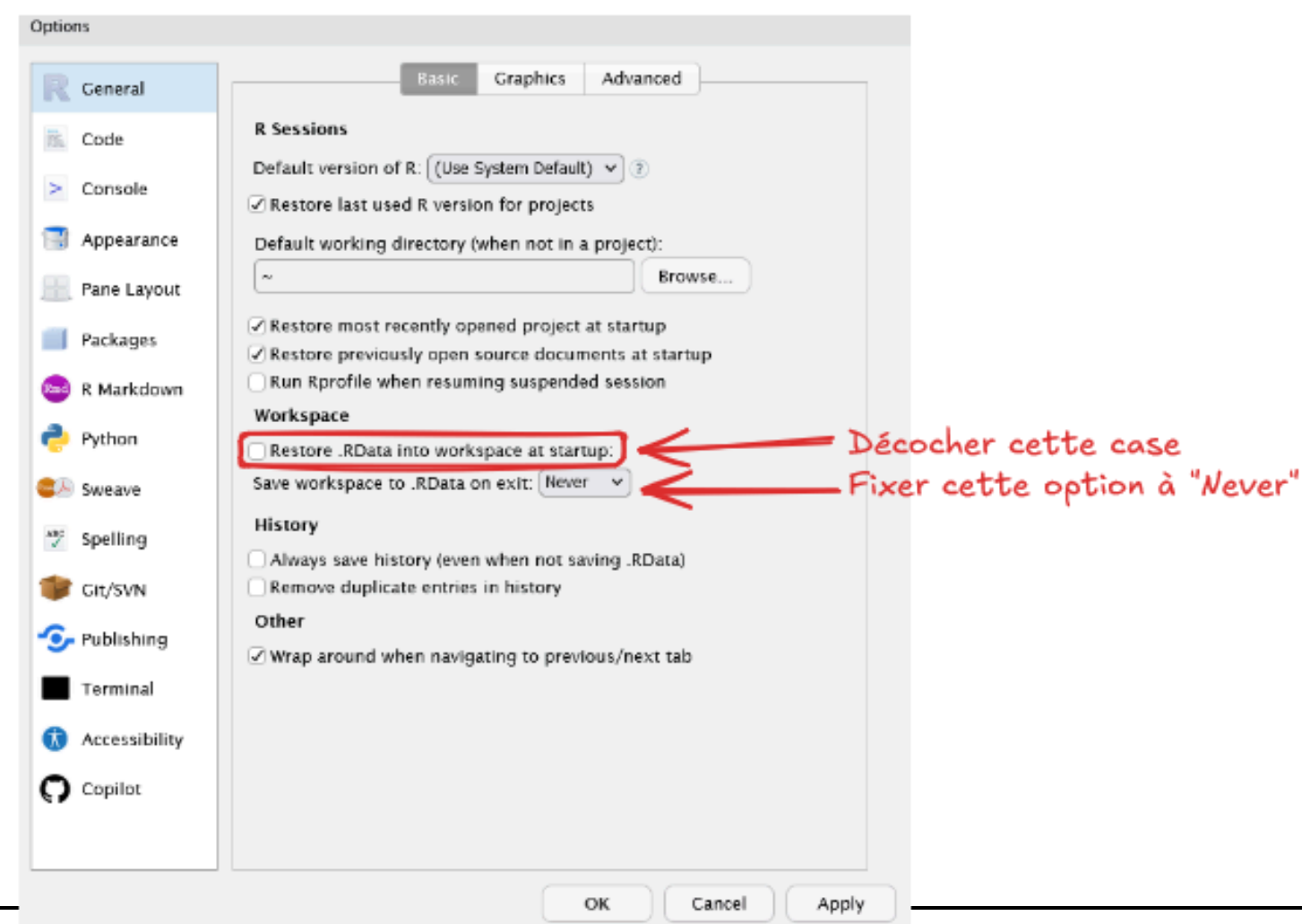
Le DEMESIS a principalement 2 métriques en tête :

- **La consommation de mémoire vive (RAM)** : utile pour stocker les données en cours de traitement dans les sessions R.
- **La consommation de CPU (processeur)** : reflète l'intensité des calculs demandés par une session R.

2.2 Gestion des ressources par les utilisateurs (1/2)

Voici quelques conseils pour limiter la consommation de mémoire sous Cerise :

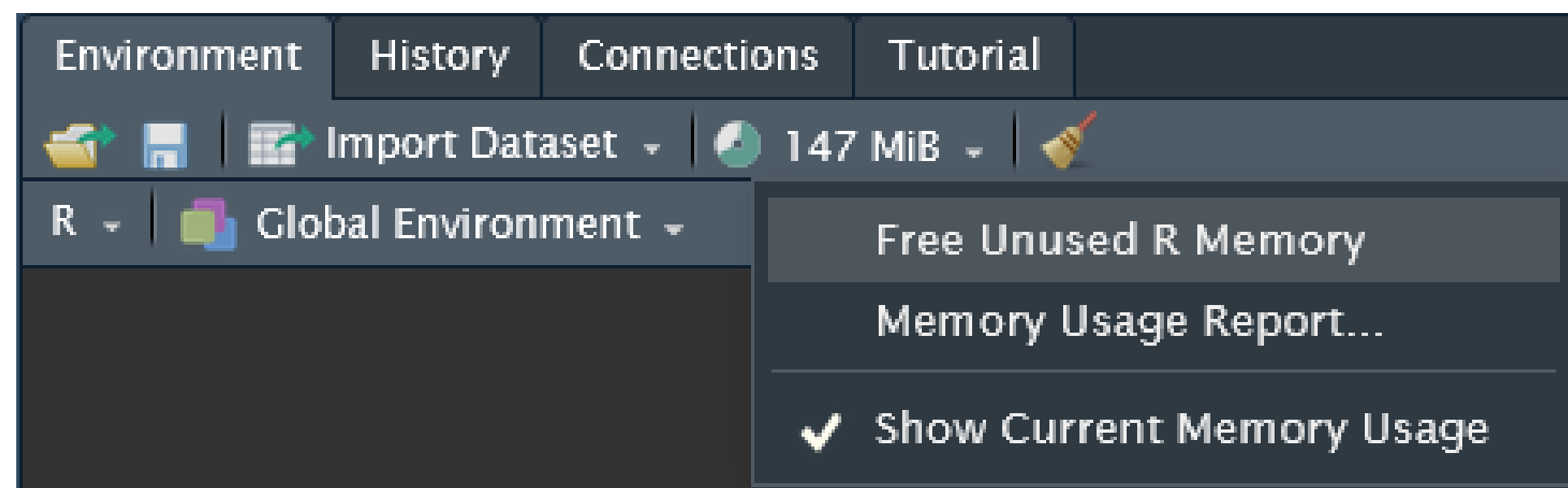
- Désactiver la sauvegarde automatique des éléments de session



2.3 Gestion des ressources par les utilisateurs (2/2)

- Nettoyer régulièrement sa mémoire vive

Utiliser la fonction `gc()` pour libérer la mémoire occupée inutilement par votre session.
Ou via l'interface de RStudio :

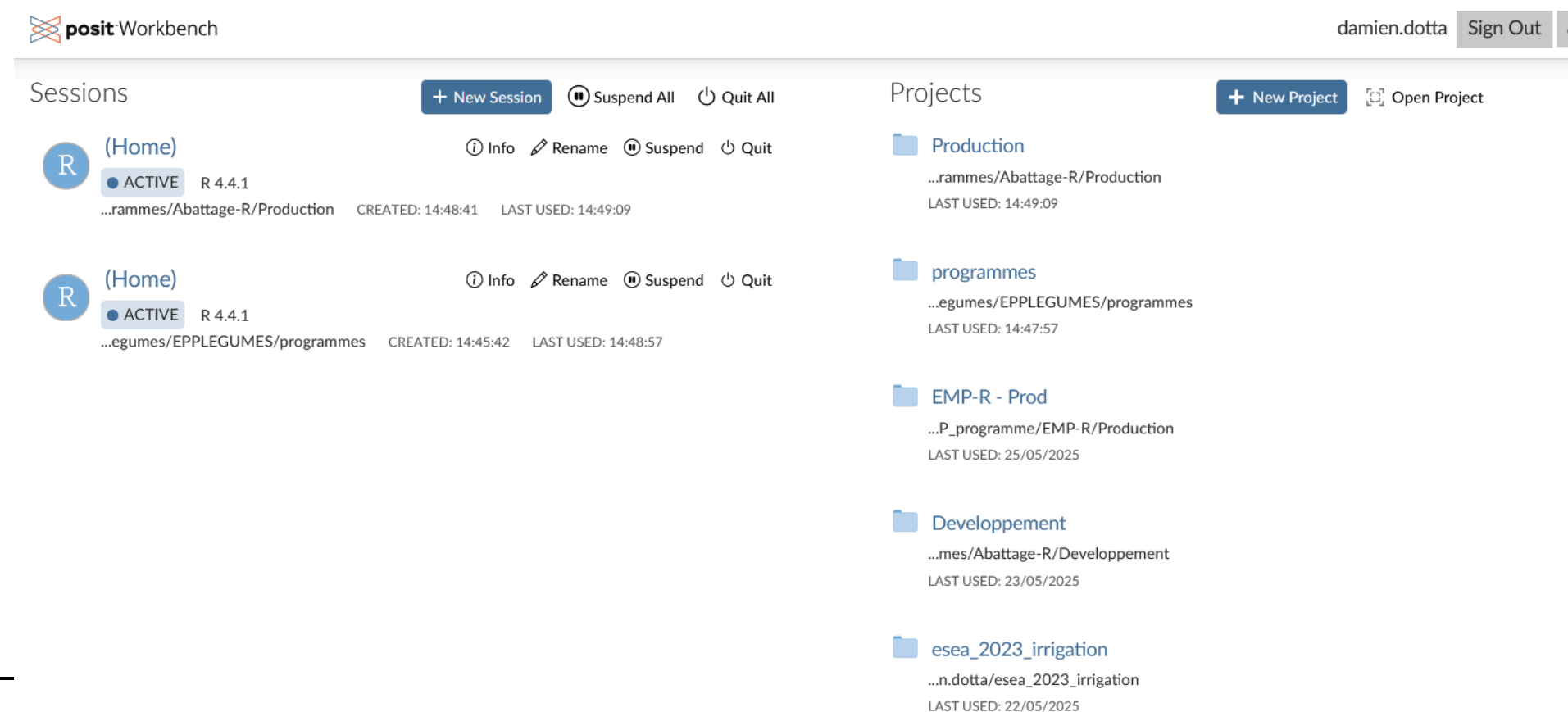


Voir [cette page d'utilité](#) pour en savoir plus.

- Pour les données volumineuses, privilégier le format de fichier Parquet (voir plus loin) Voir [ici](#) pour en savoir plus et/ou vous inscrire aux annonces de formation du BQIS.

2.4 Gestion des sessions (1/2)

- Quand vous vous connectez sur Cerise via l'adresse fournie - **si vous n'avez qu'une session d'ouverte** - Cerise vous place directement dedans (vous arrivez donc dans l'interface RStudio).
- **A partir de 2 sessions ouvertes**, lorsque vous vous connectez à Cerise, vous allez arriver sur l'écran de gestion des sessions :



2.5 Gestion des sessions (2/2)

- Dans la colonne de gauche de l'écran des sessions, vous trouverez vos sessions ouvertes.
- Dans la colonne de droite de l'écran des sessions, vous trouvez les différents projets que vous récemment utilisés.

Chaque session est indépendante des autres. Si vous avez lancé un long traitement dans une session, celle-ci est occupée et non-réactive le temps du traitement, mais vous pouvez continuer à travailler normalement dans les autres sessions.

! À retenir !

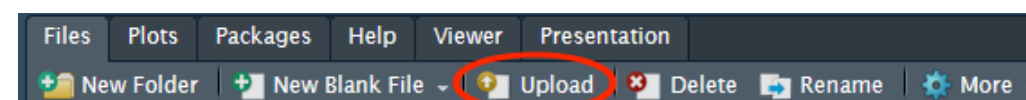
Il est important de veiller à limiter votre nombre de sessions actives (maximum 5 !) au risque de ne plus pouvoir accéder à Cerise par la suite.

Au S2 2025, il est prévu de limiter le nombre de sessions en parallèle par utilisateur et de supprimer automatiquement les sessions inactives.

2.6 Chargement/Téléchargement de fichiers Cerise (1/2)

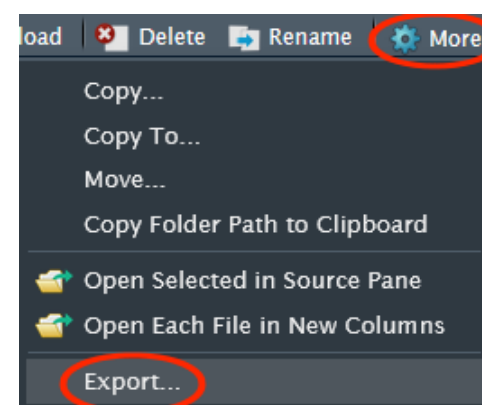
- Pour **charger** vers Cerise des fichiers depuis votre poste en local :

Cliquer sur le bouton “upload” dans l’onglet “Files”



- Pour **télécharger** depuis Cerise vers votre poste en local :

Cliquer sur la roue crantée dans l’onglet “Files”



2.7 Chargement/Téléchargement de fichiers Cerise (2/2)

- Si vous souhaitez **charger** plusieurs fichiers en même temps sous Cerise, faire un fichier ZIP. Son contenu sera ensuite automatiquement dézippé sous Cerise.
- Si vous souhaitez **télécharger** plusieurs fichiers depuis Cerise, RStudio Server va automatiquement les joindre dans un fichier ZIP sur votre poste en local.

2.8 Restauration des fichiers

- Offre de sauvegarde du centre de service (CDS)
 - Sauvegarde complète le vendredi soir
 - Sauvegarde différentielle les autres jours (Lun./Mar./Mer./Jeu. soir)
- Les sauvegardes différentielles ne sont conservées que **15 jours calendaires**
- Des demandes de restauration délicates voire impossibles :
 - Délais de remontée du besoin métier
 - Délais de prise en charge du centre de service

2.9 Versionner votre code

Une bonne pratique pour limiter les demandes de restauration de fichiers est de **versionner avec Git** vos scripts et programmes R.

Git permet :

- D'obtenir de la traçabilité
- De travailler collectivement
- De faire des revues de code
- De revenir en arrière dans le temps...

Un module de formation est disponible [à cette adresse](#), n'hésitez pas à vous y inscrire !

2.10 Utilisation du mode projet

Il est recommandé d'utiliser **le mode projet** le plus souvent possible.

Plusieurs avantages :

- Organisation claire
- Gestion des chemins d'accès portable
- Reproductibilité
- Isolation des environnements (avec `renv`)
- Intégration avec Git ...



2.11 Comparatif des formats de fichier de données

Format	Taille du fichier	Utilisation mémoire	Vitesse écriture	Vitesse lecture
Parquet	✔ Faible (colonnes compressées, binaire)	✔ Faible (lecture par lot, colonnes ciblées)	⚖ Écriture plus lente (compression + formatage)	⚡ Très rapide
RDS	✔ Moyenne à faible (compressé, un seul objet)	⚖ Modérée (lecture directe d'un objet)	✔ Rapide (compresse par défaut)	✔ Rapide (pour un seul objet)
<u>RData</u>	⚠ Moyenne à faible (compressé, contient plusieurs objets)	✖ Moyenne à élevée (charge tous les objets en mémoire)	✔ Relativement rapide	⚠ Lecture rapide mais tout est chargé (peu flexible)
CSV	⊘ Très grande (non compressé, texte brut)	⊘ Élevée (tout doit être <u>parsé</u> , conversion de type)	🕒 Rapide à écrire, peu coûteux	🐢 Lent, très coûteux en ressources

2.12 Recommandations selon le contexte

Cas d'usage	Format conseillé
Volume important, usage mutualisé, scalable	Parquet
Persistance R native, mono-objet	RDS
Sauvegarde complète d'environnement	RData
Échange simple, manuel, petit volume	CSV



2.13 Résumé des propriétés des différents formats

- CSV

Très facile à manipuler manuellement, mais inefficace pour les gros volumes.

Pas de support natif pour les types de données (dates, facteurs).

I/O très lente en R pour de grands volumes.

- Parquet

Requiert le package “arrow” en R.

Lecture partielle possible (par colonne ou ligne).

Excellent pour la mutualisation, le cloud, ou les flux inter-langages (Python, Spark...).

- RData

Sauvegarde plusieurs objets, utile pour des environnements complexes.

Inadapté pour charger sélectivement un seul objet dans un gros fichier.

- RDS

Format natif optimisé pour un seul objet R.

Très bon compromis pour la persistance simple et performante en R pur.



MINISTÈRE
DE L'AGRICULTURE
ET DE LA SOUVERAINETÉ
ALIMENTAIRE

*Liberté
Égalité
Fraternité*

3 Droits Cerise



3.1 Habilitations

La majorité des espaces présents sous Cerise sont soumis à des régimes d'habilitations.

Pour les nouveaux arrivants, ce sont les responsables hiérarchiques qui demandent les habilitations sur l'ensemble des outils.

La procédure à suivre est disponible sous Pistache [sur cette page](#).

Pour toutes les autres habilitations supplémentaires au fil de l'eau qui concernent Cerise, vous pouvez faire une demande à la BAL d'assistance : assistance.si-stat.sg@agriculture.gouv.fr



3.2 Matrice des habilitations

Groupes : 99999 : GU_SSP_SSP Utilisateur Standard

99998 : GU_SSP_CERISE_ADMIN Administrateur CERISE

99997 : GU_TEC_AGENT Automate Java

1RRRSS : GU_SSP_PROD_RRRSS Producteur de données

2RRRSS : GU_SSP_ETUD_RRRSS Chargé d'études

Avec : 1 : Producteur de données (R/W)

2 : Chargé d'étude (R)

RRR : Code de la Rubrique

SS : sous code de la source

GU_SSP_SSP

GU_SSP_CERISE_ADMIN

GU_TEC_AGENT

GU_SSP_PROD_<CODE_SOURCE>

GU_SSP_ETUD_<CODE_SOURCE>

00 – Espace-Personnel					R	R			
	Dossier personnel				<RW>				
01 – Espace-de-Partage					R	R			
	SSM								
	SSP								
		BMIS			RW	RW			
		SDSAFA			RW	RW			
		SDSRR			RW	RW			
		CEP			RW	RW			
		MDD			RW	RW			
	SRISE								
		Hauts-de-France			RW	RW			
		[... Régions ...]			RW	RW			
		Mayotte			RW	RW			
02 – Espace-de-Production									
	NOM_RUBRIQUE					RW		R	R
		NOM_SOURCE						RW	R
			nom_enquête annuelle			RW		RW	
				Univers		RW		RW	R
				Echantillons		RW		RW	R
				Données brutes		RW		RW	R
					Capibara	RW	RW	R	R
					Externe	RW		RW	R
				Données traitées		RW		RW	R
03 – Espace-de-Diffusion					R				
	NOM_RUBRIQUE				R			R	R
		NOM_SOURCE			R	RW		RW	R
			enquête annuelle		R				
				Données diffusées	R				
04 – Espace-Echanges					RW	RW			

3.3 Rendre un dossier (ou un fichier) modifiable par vos collègues

Des ACL (Access Control List) sont appliqués dans Cerise.

Il s'agit d'un mécanisme de gestion des droits qui permet de définir qui peut accéder à quelles ressources et avec quels niveaux de permissions.

Côté utilisateurs, cela implique quelques règles d'usage à suivre - pour éviter notamment la non-modification d'un dossier/fichier par vos collègues.

❗ Règle générale :

Ne pas faire “Déplacer...” des dossiers/fichiers depuis son espace personnel vers un espace de partage mais faire un “Copier vers ...”



MINISTÈRE
DE L'AGRICULTURE
ET DE LA SOUVERAINETÉ
ALIMENTAIRE

*Liberté
Égalité
Fraternité*


4 Assistance



4.1 Demander de l'aide sur Cerise - Qui ?

Les demandes d'assistances et les remontée de bugs sont à adresser à : assistance.si-stat.sg@agriculture.gouv.fr.

- Directement par mail;
- Via [cette page](#) en cliquant sur le bouton :

 [Faire une demande d'assistance](#)

4.2 Demander de l'aide sur Cerise - Comment ?

Essayez autant que possible de suivre les conseils contenus dans [cette page](#) ou [celle-ci](#) avant de poser votre question.

Votre demande sera d'autant plus vite traitée que celle-ci sera facilement reproductible par l'équipe d'assistance

4.3 Mise à disposition plateforme de test

Cerise PPRD : <https://rstudio-pprd.agriculture.rie.gouv.fr>

Objectif :

- Test des programmes sur la nouvelle version de R à venir
- Montée de version des packages et mise à jour des programmes le cas échéant

Contexte de test :

- Habilitations : Iso-production
- Système de fichiers / arborescence couramment synchronisé avec Cerise de PROD
- Tests ouverts à tous les utilisateurs Cerise





MINISTÈRE
DE L'AGRICULTURE
ET DE LA SOUVERAINETÉ
ALIMENTAIRE

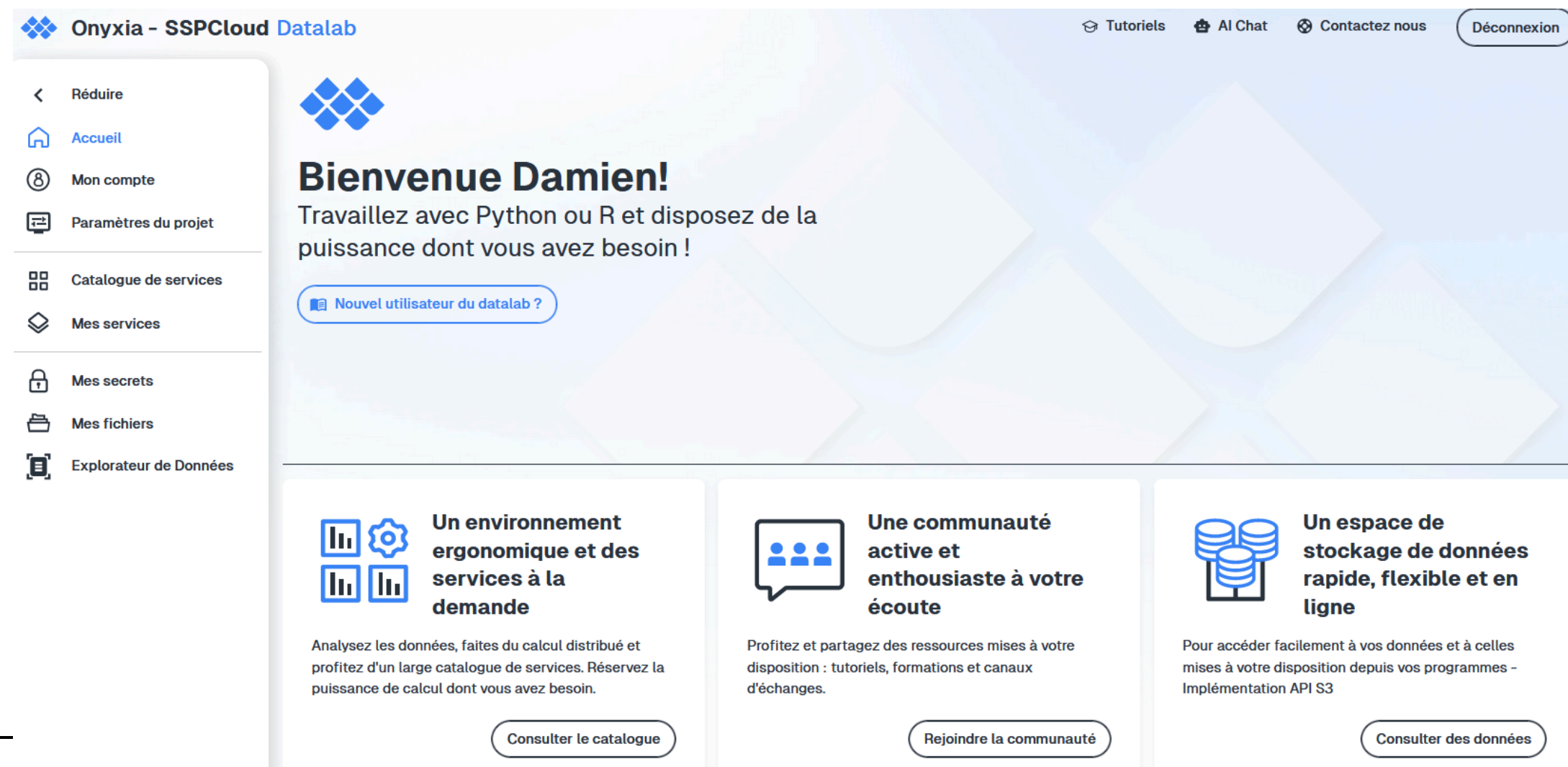
*Liberté
Égalité
Fraternité*

5 Onyxia



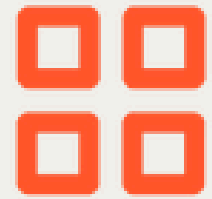
5.1 Onyxia - SSP Cloud

- Onyxia : une plateforme open source de traitement de données moderne développée par l'Insee
- SSP Cloud : une instance d'Onyxia déployée, maintenue et opérée par le SSP (Service Statistique Public)



5.2 Avantages d'Onyxia

Un Datalab dimensionné pour les usages innovants



Catalogue de services de traitement de données

Déployer des services et outils
à la demande avec un niveau
avancé de personnalisation.



Service de stockage de données

Proposer un système de stockage
d'objets distribué hautes performances
compatible avec Amazon S3.



Ressources centralisées, garanties et équilibrées

Capacité de calcul adaptée
à la plupart des utilisations.



Accès aux couches basses de l'infrastructure

Autonomie complète sur la
configuration d'environnements
avancés et spécifiques.

5.3 Remarques

- Le Datalab est une **plateform mutualisée** : les ressources utilisées par les services sont partagées entre les différents utilisateurs.
- **Pas de sauvegarde “classique” du code informatique** dans le Datalab => l'utilisation du contrôle de version avec Git est **obligatoire**.
- Même chose pour **le stockage des données** : la solution de stockage de fichiers associée au Datalab est **MinIO**, un système de stockage d'objets basé sur le cloud, compatible avec **l'API S3 d'Amazon**.

5.4 Aperçu du catalogue de services

Catalogue de services

Rechercher

Tous

Interactive services

Databases

Deprecated

Automation

Experimental

Dataviz

Homebrew

Jupyter-python

The JupyterLab IDE with Python and a collection of standard data science packages.

Lancer

Rstudio

The RStudio IDE with a collection of standard data science packages.

Lancer

Vscode-python

The Visual Studio Code IDE with Python, Julia, and a collection of standard data science packages.

Lancer

Jupyter-tensorflow

The JupyterLab IDE with Python and the deep-learning framework TensorFlow.

Lancer

Vscode-tensorflow-gpu

The VSCode IDE with Python and the deep-learning framework TensorFlow, with GPU support.

Lancer

Vscode-pyspark

The Visual Studio Code IDE with PySpark, an interface to use Apache Spark from Python.

Lancer

Jupyter-pytorch

The JupyterLab IDE with Python and the deep-learning framework PyTorch.

Lancer

Rstudio-gpu

The RStudio IDE with a collection of standard data science packages, with GPU support.

Lancer

Rstudio-sparkr

The RStudio IDE with a collection of standard data science packages. It includes SparkR, an R package that provides an interface to use Apache Spark from R.

Lancer

42 / 44

5.5 Un espace dédié à l'expérimentation

- Travail uniquement sur **des données ouvertes et non sensibles**
- Hébergement d'hackatons
- Déploiement de dataviz'
- ...



5.6 Bibliographie

- La documentation du SSP Cloud est accessible [sur cette page](#).

