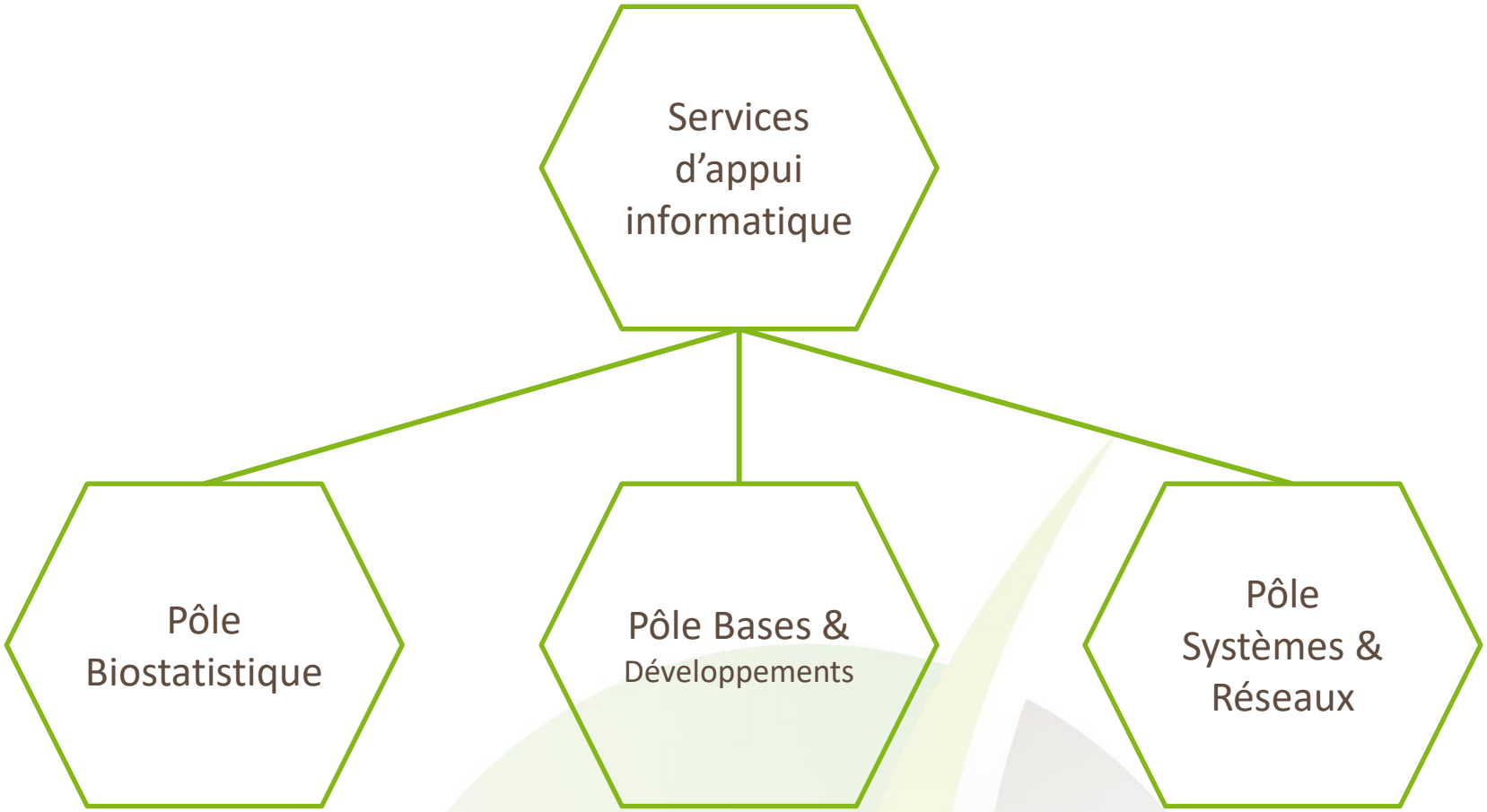
Sécurisation des analyses statistiques avec R : Retour d'expérience

Aurore Philibert
Julien Dugas

Pôle biostatistique du GEVES



GEVES | Groupe d'Étude et de contrôle des Variétés Et des Semences

Pôle biostatistique du GEVES

- Le pôle Biostatistique :
  - Est constitué de 4 personnes
  - A pour mission d'apporter conseil en interne dans le choix des méthodes et des outils statistiques les plus appropriés. Ainsi que de leur mise en œuvre.
  - Intervient dans le cadre de projet opérationnels (en routine) et de recherche
  - Doit maintenir les programmes et applications Shiny développées par le pôle tout au long de leur utilisation (aide aux utilisateurs, gestion des bugs,…)

- Le pôle Bases & Développements :
  - Est constitué de 8 personnes + des prestataires (développeurs) ponctuels
  - A pour mission de gérer les bases de données de l'entreprise
  - A pour mission de développer et maintenir de nombreux applicatifs métiers (gestion et analyse des données aux champs, gestion des données de laboratoires, …)

- Le pôle Systèmes & Réseaux :
  - Est constitué de 4 personnes
  - A pour mission principale de gérer l'infrastructure depuis les serveurs jusqu'aux terminaux en passant par les réseaux et l'assistance aux utilisateurs

GEVES | Groupe d'Étude et de contrôle des Variétés Et des Semences

Objectifs du pôle biostatistique

- Sécurisation des analyses statistiques développées (et à maintenir)
  - Quand une analyse statistique est lancée, tout le monde lance la même version du script R, du logiciel R et des packages
  - Chaque analyse statistique utilise une version spécifique de R et de ses packages
  - Ne faire aucune installation de R sur le poste des utilisateurs
  - Garantir un temps de calcul identique pour tous les utilisateurs
  - Permettre le lancement des scripts R via des applicatifs virtualisés (accessibles par internet, développés en .net)
  - Pouvoir retrouver la version de l'analyse statistique lancée pour produire un résultat précédent
  - Faciliter la montée de version R/packages
  - Rendre possible et rapide le transfert de maintenance d'une chaîne statistique d'un.e statisticien.ne à un.e autre
- Automne 2022 -> aucun de ces critères n'était intégralement respecté

Démarche

- Pour remplir ces objectifs, deux chantiers en parallèle ont été mis en œuvre :
  - Création d'un environnement commun de développement R
  - Création d'une plateforme de biostatistique

Environnement de développement R (1/3)

- Charte d'écriture des programmes R :

  Permettre la reprise d'un script par une autre personne

  Travail en interne du pôle pour définir les conventions de nommage, la façon de coder, etc. Historiquement les scripts R étaient codés de façon très imbriquée ce qui rendait difficile le débogage et la reprise du code par un.e autre statisticien.ne

  Rapide, a nécessité seulement une réunion. Devra être revu dans quelques temps si besoin d'ajustement

- Tests unitaires :

  Permet de monter de version de R très rapidement

  Apprentissage du fonctionnement des tests unitaires dans R en autonomie + 1 formation

  Le concept et la mise en œuvre sont apparus très intuitif. Un temps est à prévoir dans tous les projets pour leur implémentation. Difficultés rencontrées sur les graphiques (chaque élément du graphique doit être identique ? L'objet doit être le même ? Etc.)

Environnement de développement R (2/3)

- Généralisation de l'utilisation de Git :

  Permet un suivi des modifications des scripts R et d'accéder aux versions précédentes ou actuelles par une autre personne

  Apprentissage autonome de Git + 1 formation générale

  Mis en place plus ou moins facilement. Nécessite une formation pour certains et besoin de l'utiliser souvent pour rendre le process automatique. Possibilité d'étiquetage/label (ex : n° version)

- Suivi des versions :

  Permettre de gérer l'historique

  Apprentissage en interne du pôle pour donner des numéros de version à nos scripts, comme en informatique, afin de pouvoir suivre les évolutions

  En cours d'implémentation, beaucoup de questions autour de « est-ce une modification majeure ou mineure ? ». Finalement le plus important était d'avoir systématiquement un numéro de version différent à chaque nouvelle mise en production.

Environnement de développement R (3/3)

- Documentation/Vignette :
  - Permet de garder une trace de :
    - Que contient le script
    - Quelles analyses statistiques sont faîtes et pourquoi ce choix
    - Suivre les modifications (exemple : changement de réglementation)
    - En faisant le lien avec le code R pour permettre une reprise par quelqu'un d'autre plus facilement (exemple : la sélection des données se fait dans la fonction qui s'appelle …)
  - Chaque personne du pôle a écrit une vignette sur son application Shiny ou script R puis organisation d'une relecture de chacun puis discussion commune sur les différentes façons de faire
  - *Indispensable pour la compréhension des analyses développées précédemment et à maintenir. Assez facile lorsque l'on est « dans le sujet ». Deux formats ont été testés : la vignette qui correspond au package ou le format R Markdown html.*

    Pour les applications Shiny, deux formats possibles testés : tout en un ou bien une vignette qui décrit le fonctionnement de l'application et une autre qui détaille le code et les analyses qui sont plus en détails derrière chaque « bouton »

Création d'une plateforme de biostatistique

Permet de :

- Centraliser toutes les analyses statistiques au même endroit
- Avoir une seule version « active » par analyse
- Suivre les versions de chaque analyse dans le temps
- Chaque analyse statistique est lancée avec la même version de R, des packages et des scripts
- Mettre à disposition les analyses statistiques :
  - via les applications métiers développées par le pôle Bases & Développements
  - via des applications web développées par le pôle Biostatistique (Shiny)

Formation en autodidacte sur les technologies à utiliser (Docker) + Besoin des compétences des 2 autres pôles informatiques pour :

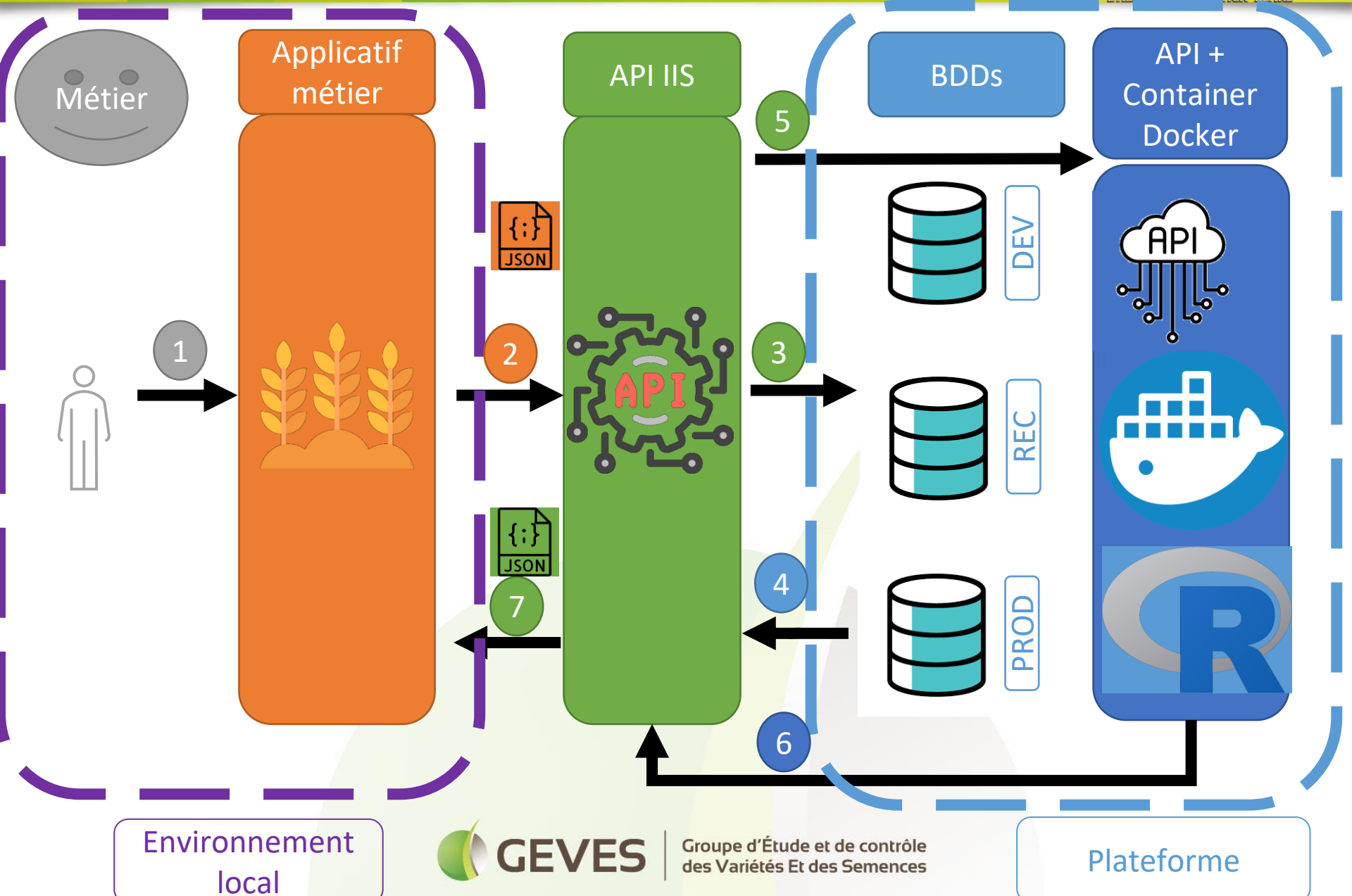
- Monter le serveur et installer Docker (S&R)
- Développer une API qui relie les applicatifs métiers à la plateforme (B&D).
- Pour les scripts R, nécessité de créer une base de données de suivi des utilisations de la plateforme.
- Format d'échange de données (entrée/sortie) avec l'applicatif métier passé du csv au json

A nécessité de nombreux échanges entre les trois pôles. Avec des problèmes de compréhension car vocabulaire différent, habitudes d'outils utilisés différents, etc. Pour une mise en production fin 2023 ! Utilisée pour toutes les applications Shiny. Transfert en cours pour le lancement des scripts R

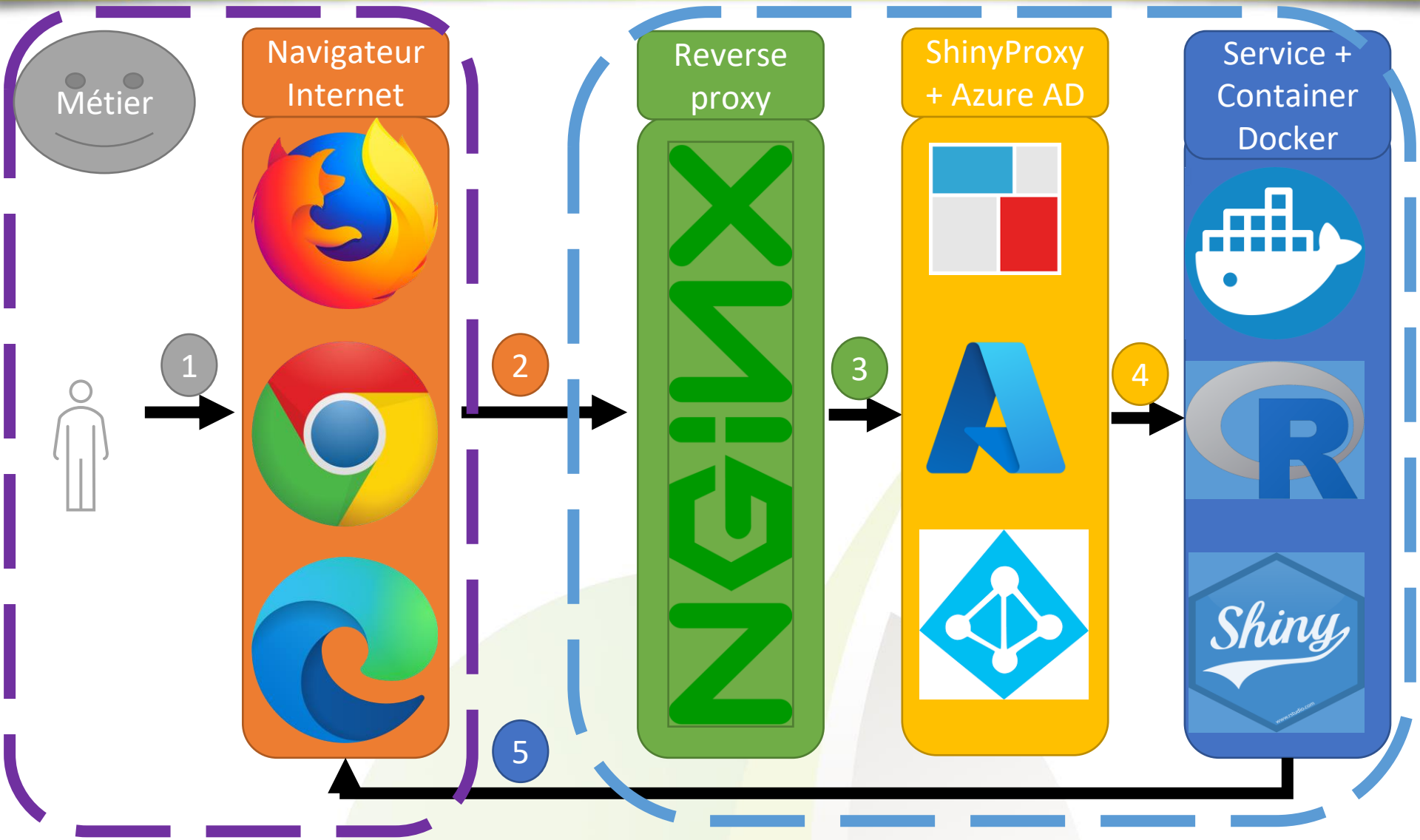
Fonctionnement de la plateforme

- Trouver une technologie qui :
  - Réponde aux objectifs de la plateforme (centralisation, gestion de versions (R, packages, scripts), maintenance, indépendance du poste utilisateur)
  - Soit bien documentée et connue pour une prise en main rapide et efficace
  - Puisse être facilement appelée (API déjà développée) par les applicatifs métiers
  - Soit peu onéreuse et nécessite peu de ressources matérielles
- Choix des technologies Docker, Linux, ShinyProxy :
  - Choix de Docker : environnements d'utilisation indépendants du système hôte pouvant être facilement créés et déposés sur des serveurs de calculs (centralisation)
  - Choix de ShinyProxy : gestion facilitée de multi-utilisateurs sur plusieurs applications (containerisation des lancements utilisateurs des applications R-Shiny)
  - Choix de Linux : bonne compatibilité avec Docker et ShinyProxy, stabilité
  - Technologies éprouvées et répandues avec des documentations très détaillées
  - Appel de l'API Docker en concordance avec les appels API déjà développés au pôle Bases et Développements et appel facilité à la plateforme en interne au pôle
  - Aucun coût, uniquement sur les serveurs utilisés (taille, nombre de cœurs CPU, etc.)
- Formation en autodidacte, mode d'emploi plateforme, tests de performance et de stabilité, aller-retours entre les 3 pôles SI, difficultés de connexion entre technos, passage par plusieurs solutions insatisfaisantes (docker compose, ShinyServer)

Fonctionnement de la plateforme : scripts R



Fonctionnement de la plateforme : R-Shiny



Environnement local

GEVES | Groupe d'Étude et de contrôle des Variétés Et des Semences

Plateforme

Merci pour votre attention !

Et n'hésitez pas à venir nous voir pour discuter de vos propres retours d'expériences et/ou questions !

GEVES | Groupe d'Étude et de contrôle des Variétés Et des Semences