Schéma Directeur Informatique ODR 2021-2024

Yann Darsel

SCHEMA DIRECTEUR INFORMATIQUE 2021-2024

03 MARS 2021 – VERSION 0.2

Observatoire du Développement Rural
Yann Darsel

Sommaire :

Schéma Directeur Informatique - Notions

Le présent document décrit l'évolution souhaitée du **Système intégré d'Information sur les Systèmes et Politiques Agricoles SISPA**[1] en termes de ressources informatiques (RH, logiciels, matériels, règles d'organisation, développements) nécessaires pour l'accompagnement des projets stratégiques de l'Unité de Service de l'Observatoire du Développement Rural (US-ODR).
Il ne concerne pas les aspects bureautiques et administratifs de l'Unité.

Ce Schéma Directeur Informatique (SDI) permet, en s'inscrivant dans le cadre du projet d'unité, et à partir de l'état des lieux du système d'information, de définir une cible à moyen terme, d'identifier, de qualifier et de planifier les projets qui contribuent à atteindre cette cible, en fonction de besoins identifiés par les acteurs.

Les orientations d'évolution du système d'information concernent la période couverte par le projet d'Unité, de 2021 à 2024. Elles seront revues à fréquence régulière (tous les 6 mois) par le pôle informatique en relation avec le Groupe de Travail Informatique (GTI).

Dans certains domaines, plusieurs scénarios sont présentés dans ce SDI en tenant compte des contraintes et des priorités. Il s'agira de choisir la meilleure trajectoire.

Ce SDI s'accompagne d'une organisation qui permet de :

- Fixer les règles de qualification des projets (niveau de priorité),
- Formaliser l'évaluation des nouveaux projets (périmètre, arbitrage).

« Les orientations d'évolution du système d'information concernent une période de 4 années, de 2021 à 2024. Elles seront revues tous les 6 mois. »

[1] **Projet unité de service Observatoire du Développement Rural (US-ODR) pour 2019-2024**

Le Système d'Information de l'Observatoire du Développement Rural

Aujourd'hui, le système d'information[2] de l'Unité de Services ODR est une plateforme **orientée services**, accessible via internet qui propose aux utilisateurs des outils de traitements statistiques et de cartographie en ligne.
C'est une plateforme **collaborative** mise à disposition de partenaires de l'INRAE pour créer des « **observatoires** » caractérisés par des communautés d'utilisateurs partageant et échangeant des documents et données à travers des services informationnels.

Le premier partenariat, avec l'Agence de Services et de Paiement (ASP) et le Ministère de l'Agriculture, a entrainé en 2006 la création d'un « Observatoire des programmes communautaires de développement rural ». Les **données primaires**[3], issues des dossiers administratifs liés aux aides communautaires du second pilier de la PAC, via les outils de gestion ISIS ou OSIRIS, sont déposées par l'ASP et traitées par l'US-ODR pour créer des **indicateurs**[4] **.** Ces derniers peuvent être exploités par différents utilisateurs autorisés pour la conception de modèles analytiques et de méthodologies d'évaluation de la Politique Agricole Commune (PAC).

D'autres partenariats ont vu le jour depuis, notamment :
- avec la Caisse Centrale de la Mutualité Sociale Agricole (CCMSA), permettant la mise en place de « Tableaux de bord de l'emploi agricole ».
- avec l'actuel Ministère de la Transition Ecologique (MTE, précédemment Ministère de l'Écologie et du Développement durable) et l'OFB, qui fournissent des couches d'informations géographiques et remobilisent les indicateurs dans différents cadres (suivi des pollutions diffuses et autres impacts de l'agriculture sur l'environnement).
- avec l'Institut National de l'Origine et de la qualité (INAO), pour développer et maintenir l'« Observatoire territorial des Signes d'Identification de la Qualité et de l'Origine » OT-SIQO.

Le système d'information actuel de l'US-ODR est une plateforme collaborative mise à disposition de partenaires de l'INRAE pour créer des « observatoires ».

Les données primaires, confidentielles pour la plupart, restent sous le contrôle de leurs propriétaires et sont utilisables par des chercheurs de l'INRAE, ou d'autres utilisateurs sous forme d'échantillons anonymes.

[2] https://odr.inrae.fr
[3] Données déposées par les partenaires fondateurs et/ou les tiers agréés.
[4] Information à laquelle un observateur peut attribuer un sens. Il s'agit donc d'une information élaborée contrairement à une simple donnée.

Les indicateurs élaborés à l'aide de la plateforme sont publiés dans des « dossiers thématiques » offrant des outils interactifs de consultation sous forme de cartes et/ou de tableaux, pour des utilisateurs autorisés.

Le système d'information de l'US-ODR a également été utilisé pour préparer et diffuser d'autres catégories de données (météo, Registre Parcellaire Graphique RPG, données sur les ventes de produits phytosanitaires,…) avec le concours des ingénieurs de l'US-ODR.

Le SI permet donc aux agents de l'US-ODR de produire des indicateurs originaux et réaliser des travaux d'expertise qui viennent enrichir des recherches transversales larges comme par exemple la place des signes de qualité dans le développement de formes d'agriculture durables, ou encore les transformations qu'induisent les politiques de développement rural sur la dynamique de l'emploi.
L'équipe ODR, grâce à sa connaissance approfondie des données, à son expérience en méthodologie de traitements de données, propose des solutions adaptées aux besoins des utilisateurs à travers l'amélioration continue du SI.

Organisation fonctionnelle des données :

Les données produites par le SI de l'US-ODR sont des indicateurs spatialisés partageant la même plateforme informatique et distribués parmi 5 réseaux[5] structurants / domaines thématiques:

- Le domaine des travaux d'évaluation des Programmes de Développement Rural PDR (**Réseau Evaluation**). Ces informations permettent le suivi, l'analyse et l'évaluation des politiques rurales territorialisées et des agrosystèmes, sur la base d'une convention multi partenariale (2015-2024) entre l'INRAE et le MAA, l'ASP, l'ARF, le MTES, la CCMSA et l'INAO.
- Le domaine de l'emploi agricole (**Réseau Emploi**) en collaboration avec la CCMSA. Ces informations permettent d'améliorer la connaissance des populations agricoles salariées et non-salariées, avec comme objectif l'analyse des conditions de développement d'emplois dans les territoires ruraux.
- Le domaine des signes d'identification de la qualité et de l'origine (**Réseau Qualité**) en lien avec l'INAO. Ces informations permettent la réalisation de tableaux d'indicateurs et, avec les données géolocalisées sur les opérateurs et autres statistiques économiques, de cartes valorisées au niveau d'un produit, d'une filière, d'une région ou d'un autre niveau géographique. Les Signes d'Identification de la Qualité et de l'Origine (SIQO) comprennent les signes de qualité européens : Appellation d'Origine Protégée (AOP), Indication Géographique Protégée (IGP), Spécialité Traditionnelle Garantie (STG), Agriculture Biologique (AB), et les signes français : Appellation d'Origine Contrôlée (AOC) et Label Rouge.
- Le domaine d'évaluation et de modélisation des systèmes agricoles (**Réseau Systèmes Agricoles**), qui se base notamment sur la valorisation du Registre Parcellaire Graphique (RPG), produit par l'ASP et l'Institut Géographique National (IGN). Ce sont des données sur les modes d'occupation agricole de sol, la diversité des cultures et les séquences de cultures au sein de territoires ou au sein des exploitations agricoles, au niveau géographique des îlots culturaux (îlots RPG), particulièrement fin. Ces informations permettent de détailler et géo-référencer les systèmes de culture (nature des cultures et leurs séquences) et les systèmes de production (exploitations agricoles françaises), grâce à l'acquisition et l'utilisation des différentes versions du RPG et d'autres sources de données : la Banque Nationale des Ventes de produits phytopharmaceutiques par les Distributeurs (BNVD), les enquêtes du Service de la Statistique et de la Prospective (SSP), entre autres.

[5] **Guide de l'utilisation avancée du site de l'ODR et de l'outil Carto Dynamique**

- Le domaine des initiatives d'observatoires locaux et régionaux (**Réseau Territoires**). La finalité de ce domaine, resté dans un état de développement initial, consiste en une utilisation transversale des données issues des 4 précédents réseaux avec une approche de départ exclusivement territoriale (géographique).

Ces 5 réseaux sont eux-mêmes divisés en programmes. Les programmes comportent des projets et les projets des dossiers. Les dossiers regroupent un ensemble de ressources sur un thème donné : fiche méthodologique, tableaux d'indicateurs, cartes, etc.

En supplément de ces réseaux structurant l'activité de l'Unité, un réseau dit « recherche », hors partenariats, a pour objectif de fournir un espace de partage de données, de résultats et de documents provenant d'expertises menées par l'US-ODR en collaboration avec des équipes de recherche, via un accès réservé.

Nature des données du SI de l'US-ODR

- Les données géoréférencées : les informations sont soit directement territoriales, sous la forme de données géocodées (se rapportant à des limites administratives communales, départementales, régions par exemple), soit elles-mêmes situées sur un territoire, sous la forme de données géolocalisées (bénéficiaires d'aides PAC, par exemple). Le principal référentiel géographique est celui des communes.
  Un référentiel géographique particulier est issu du Registre Parcellaire Graphique fourni par l'ASP puis par l'IGN qui recense les parcelles agricoles des exploitations sujettes à des aides PAC. Il définit les contours graphiques des îlots et parcelles de cultures dessinés par les exploitants qui déclarent les cultures pratiquées et leurs surfaces.
  Un RPG complété (version 2 en cours) contient, en complément, les parcelles des exploitations spécialisées dans des cultures non aidées (viticultures, arboricultures, etc.) afin d'assurer une information géo-référencée exhaustive sur l'occupation du sol agricole.

- Les données non géographiques (administratives) du SI de l'ODR sont très nombreuses et variées: ce sont des données structurées, expertisées et documentées sur les différentes mesures de la PAC (depuis 1996), la démographie agricole (depuis 2002), l'occupation agricole des sols (depuis 2006), les opérateurs engagés dans des signes de qualité (depuis 2011), etc.

- Un catalogue de métadonnées décrivant les bases de données portant sur les thèmes de l'agriculture, et de la performance économique et environnementale des systèmes de culture plus particulièrement, est aussi hébergé dans le SI de l'Unité, en lien avec la GIS GCHP2E et la mission Base de Données du département. Une application permet de le parcourir : ***Agrilogue***[6].

[6] **https://odr.inra.fr/agrilogue/**

Analyse des ressources logicielles actuelles

La plateforme web (https://odr.inrae.fr/intranet/carto_joomla/index.php) et notamment l'application originale et initiale ***CartoDynamique***[7] a été développée en langages ***PHP*** (version 5) et ***JavaScript***, en intégrant l'outil de WebMapping *MapServer*.
CartoDynamique permet de présenter des informations à toute échelle géographique et éventuellement enrichies par des couches d'information géographique publiques telles que l'hydrographie, les routes, les forêts.
Les accès (libres ou limités) et certaines notifications par e-mail sont gérés avec le système de gestion de contenu (ou Content Management System CMS) ***Joomla*** (version 3.7.5) outre la production de documentions.
Un second CMS, ***Drupal*** (version 7.54), a été employé pour concevoir un catalogue de métadonnées[8] (*Agrilogue*).
La mise en page Joomla est complétée par un une plateforme de collaboration et de documentation WikiMedia : WikiODR
Explo-SIQO est une application web intégrée dans la plateforme, qui repose sur le langage ***R*** et sa bibliothèque ***Shiny***. Cet outil est actuellement en accès libre du fait du caractère succins de l'authentification-utilisateur via *Shiny*.
De même, une « application de mise à disposition des résultats du calcul de l'indicateur de résultat R2 » (Version 2.1, pour le Programme de Développement Rural), intégrée dans la plateforme, repose sur le langage ***R*** et sa bibliothèque ***Shiny.***

Les tâches planifiées du SI de l'ODR sont gérées via l'utilitaire linux ***crontab*** ou via le client web ***Jenkins***[9]. Ces tâches planifiées permettent par exemple la sauvegarde des programmes et données.

La surveillance de l'espace disque des serveurs ODR, Gemini et VIRGO est actuellement effectuée à l'aide d'un outil développé en PHP par le pôle informatique : ***Freedisk***

Enfin, une application de collecte de l'activité des utilisateurs[10] en vue de produire des statistiques d'utilisation de la plateforme a été développée avec le langage ***R*** et sa bibliothèque ***Shiny.***

Les Systèmes de Gestion de Bases de Données (SGBD) du SI de l'ODR

Les bases de données sont gérées avec 2 SGBD : ***MySQL*** et ***PostgreSQL***. Elles rassemblent plusieurs milliers de tables. Un inventaire des bases de données par domaine est en cours.
Les données géo-localisées disponibles sur grille (Raster) ou sous forme vectorielles ne peuvent être gérées qu'à travers des scripts utilisant l'extension ***Postgis***[11] du SGDB *PostgreSQL*. Pour être exploitables par l'application *CartoDynamique*, elles peuvent être attribuées à des données géocodées grâce à des opérations simples (exemple :

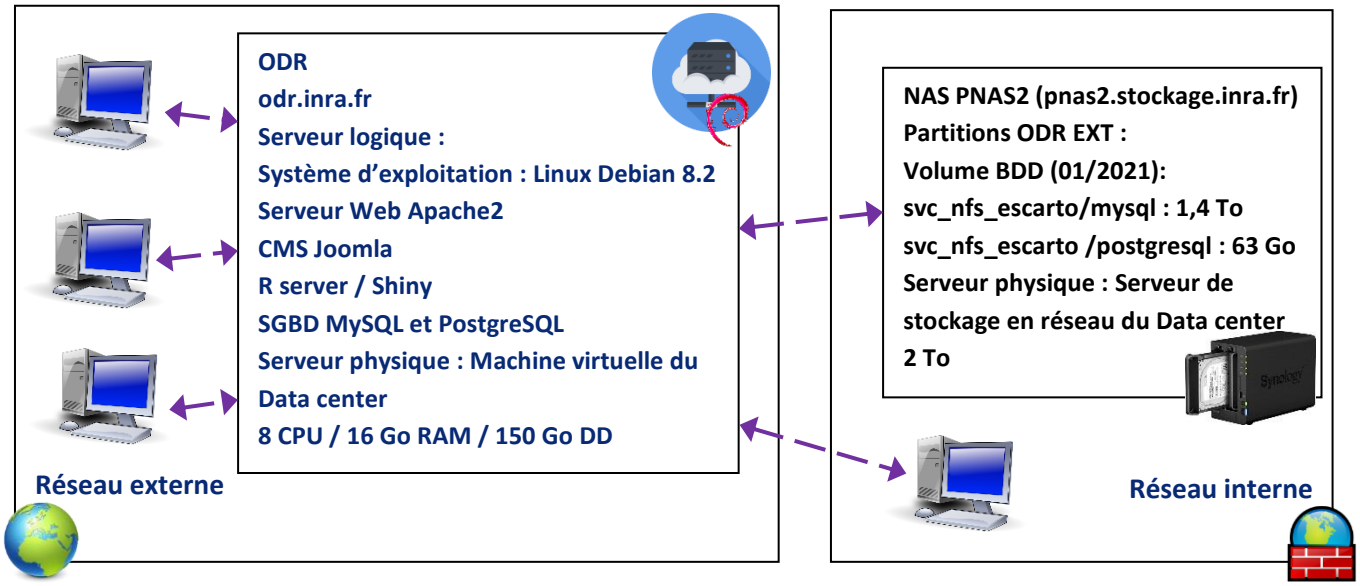
[7] Application informatique permettant de générer et de stocker des données et des traitements cartographiques et statistiques
[8] Données servant à définir ou décrire une autre donnée
[9] https://www.jenkins.io/
[10] https://odr.inra.fr/shiny/stat/ (non fonctionnelle à ce jour)
[11] https://postgis.net/

somme des surfaces de parcelles à l'intérieur d'un département) ou bien à des méthodes d'analyses statistiques plus complexes (exemple : transfert des grilles météorologiques sur des communes).

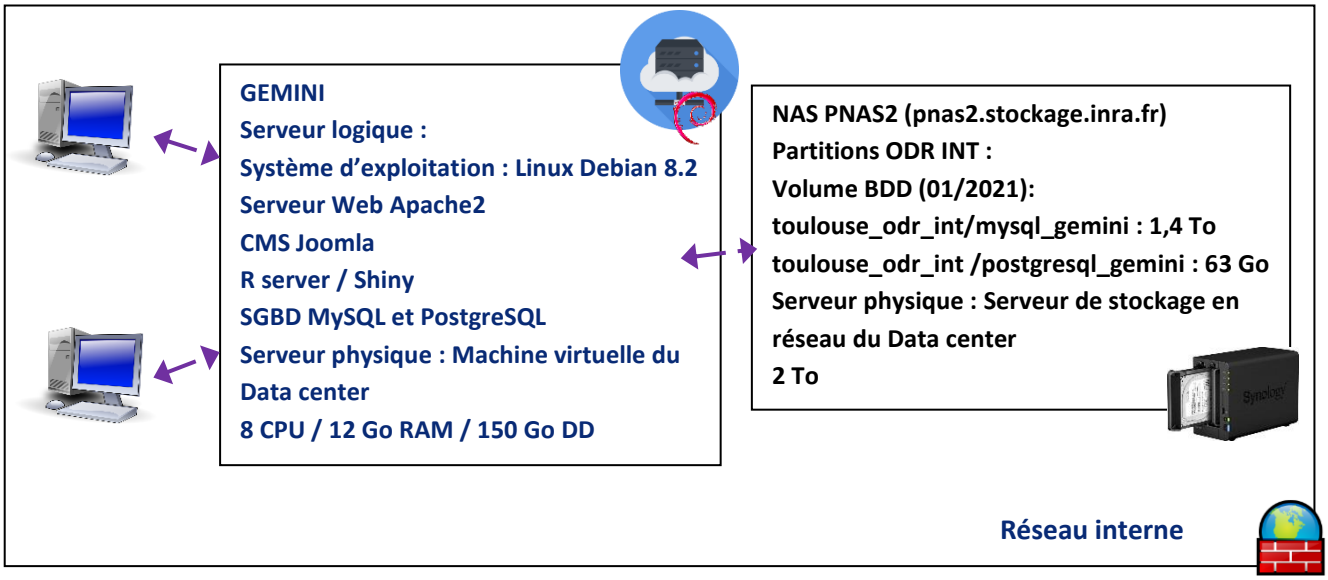
Analyse de l'Architecture réseau actuelle

Les applications du SI de l'ODR sont hébergées sur des architectures client-serveur « pseudo 3-tiers »:

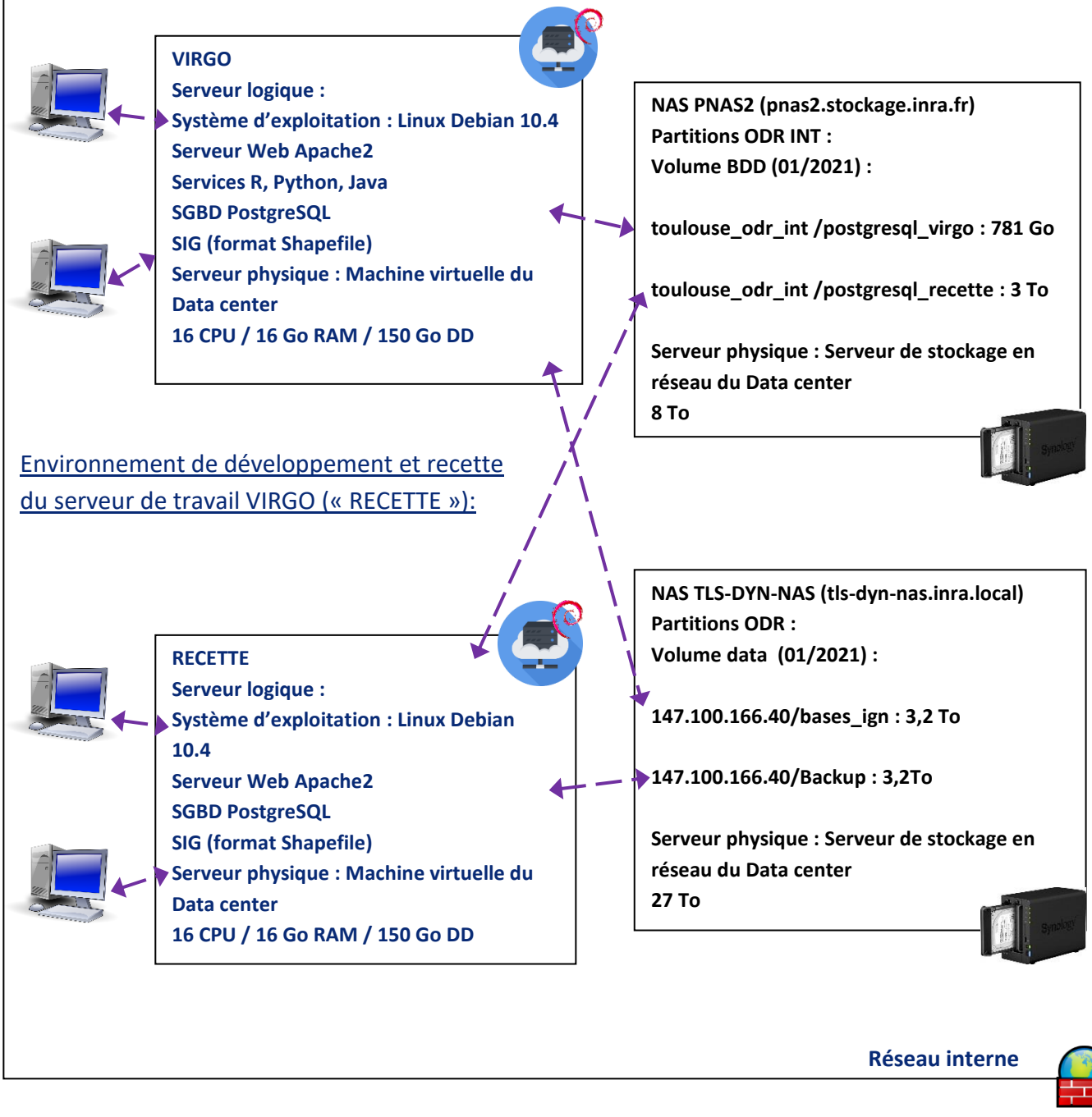
Environnement de production de la Plateforme ODR et *CartoDynamique* :

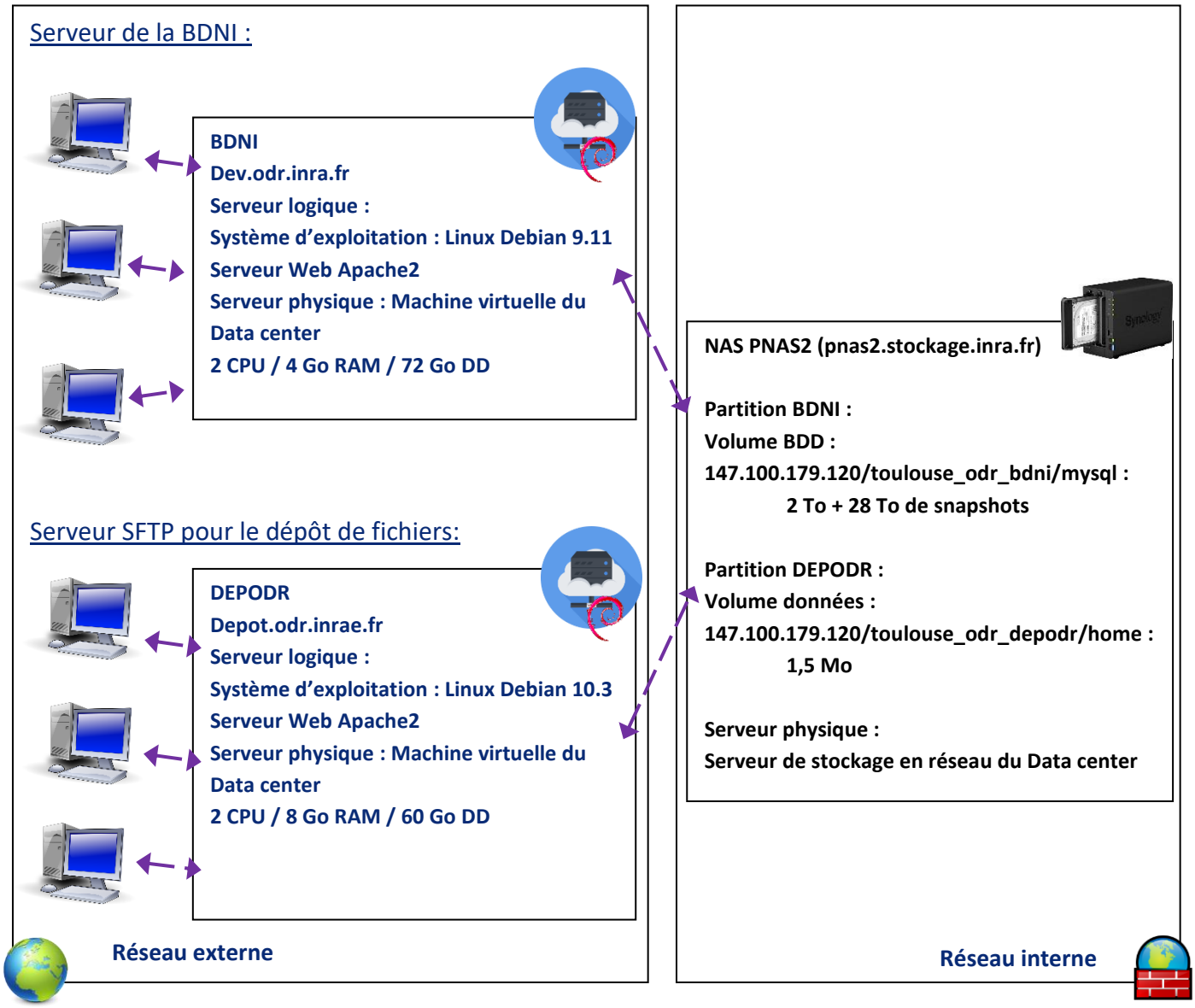


Environnement de développement et recette de la Plateforme ODR et *CartoDynamique* :

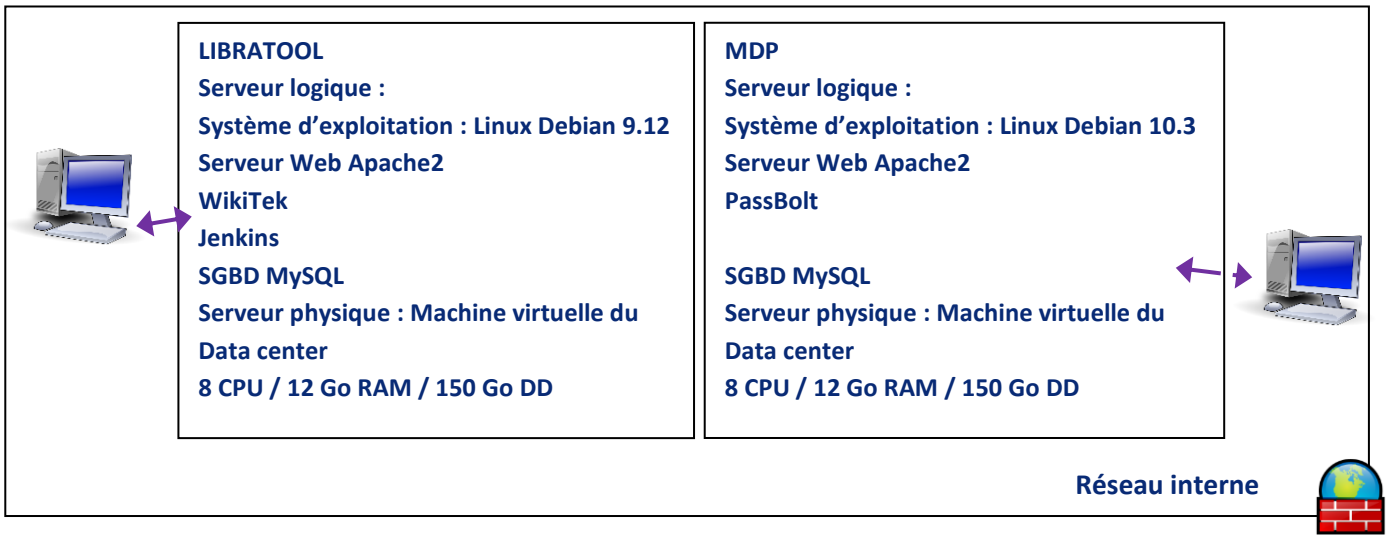


Environnement de production du serveur de travail VIRGO :





Serveurs de support Informatique (architecture client-serveur):



L'espace de stockage utilisé pour les données totalise environ 12 To

Rappel des objectifs stratégiques[12] du projet de l'Unité

Une refonte du Système d'Information est nécessaire pour accompagner les axes de développement stratégique de l'ODR. Le développement du système intégré d'information sur les systèmes et politiques agricoles (SISPA) ambitionne de :

- Valoriser des données en s'appuyant sur les trois piliers du **développement durable** : économique, environnemental et social,
- Fournir une représentation spatialisée **à une échelle fine** des systèmes de cultures (choix des cultures, ordre de successions et itinéraires techniques) et des systèmes de production (structure des exploitations, emploi, diversification, commercialisation),
- Enrichir, stocker, publier le patrimoine informationnel de l'ODR de manière à ce qu'il soit « Facile à trouver, Accessible, Interopérable et Réutilisable » (**FAIR data**).

« Une refonte du Système d'Information est nécessaire pour accompagner les axes de développement stratégique de l'ODR. »

[12] **Projet unité de service Observatoire du Développement Rural (US-ODR) pour 2019-2024**

Démarche qualité informatique

La complexité du Système d'Information, son évolution prévue, le volume de données gérées et leur confidentialité, la prévenance vis-à-vis des partenaires de plus en plus nombreux et les exigences réglementaires en matière de traçabilité nécessitent l'amélioration de bonnes pratiques informatiques en étayant la démarche qualité de l'Unité.

La formalisation des procédures de développements informatiques, de gestion et traçabilité des interventions sur les données et de suivi des travaux effectués doit être renforcée.
La documentation des procédures informatiques est d'autant plus nécessaire que les compétences humaines déployées sont régulièrement renouvelées (personnel non permanent) et doivent pouvoir s'appuyer sur une base de connaissances substantielle et structurante.

« La démarche qualité informatique va être au cœur des préoccupations du Pôle Informatique de l'Unité »

Dans ce contexte la démarche qualité informatique va être au cœur des préoccupations du Pôle Informatique de l'Unité, avec la création d'indicateurs de pilotage structurants:

- **Mise en place d'un gestionnaire de tickets (*GLPI*)** grâce auquel
    - Les anomalies rencontrées et corrigées seront tracées et alimenteront une base de connaissances ;
    - Une Foire Aux Questions (FAQ) sur les aspects informatiques sera mise à disposition de l'US-ODR,
- Mise à jour et enrichissement du **wiki informatique** du Pôle Informatique pour documenter les actions courantes ou à des fins d'inventaire,
- Mise en place d'une **gestion de documents informatiques qualité** codifiés et régulièrement révisés avec notamment la formalisation par écrit des procédures de développements informatiques et de traitement des données collectées. Plusieurs niveaux de documents qualité seront déclinés :
  Procédures générales >> Modes opératoires >> Formulaires (ou Enregistrements)

Mise en œuvre d'une gouvernance

Acteurs et rôles dans le projet

Conception (Maîtrise d'ouvrage)

La phase conceptuelle d'un projet informatique représente les fondations du produit final. Il s'agit de la première étape du cycle de vie d'un développement informatique.
Afin d'organiser efficacement cette phase, une application de gestions de projets va être déployée : ***Redmine***
Cette application permettra de suivre et de prioriser les tâches à réaliser pour chaque projet informatique d'ampleur suffisante (ne sont pas concernés les correctifs de bugs ou améliorations mineures).

> ***« Une application de gestions de projets va être déployée : Redmine. »***

Les utilisateurs de *Redmine* seront les membres de l'équipe ODR et pourront créer un projet collaboratif dans le respect des procédures qualité informatiques.
Ponctuellement, d'autres utilisateurs externes (conventions, chercheurs INRAE) pourront être associés à certains projets, notamment lors de la co-construction de cahiers des charges. Cette participation de « clients », chercheurs et partenaires, permettra de répondre aux besoins que ceux-ci exprimeront en lien avec les objectifs de l'unité.

L'utilisation du gestionnaire de projets ne dérogera pas à la formalisation par écrit des phases de projets dans la documentation qualité (Note de cadrage, Cahier des charges, Dossier de Spécifications Fonctionnelles, etc.), ces documents pouvant eux-mêmes être « attachés » dans l'application *Redmine*.
À chaque utilisateur sera associé un (ou plusieurs) rôle(s) par projet, les rôles définissant les permissions accordées aux utilisateurs dans un projet. Ces rôles seront cohérents avec ceux définis dans la Note de Cadrage.

La traçabilité des tâches effectuées sur les projets informatiques sera ainsi assurée puisque *Redmine* indexe toutes les mises à jour, la personne qui les a faites, ainsi que les dates.

Réalisation (Maîtrise d'œuvre)

La réalisation des projets informatiques sera conduite par les administrateurs et développeurs du pôle informatique, en collaboration avec d'autres membres de l'équipe US-ODR identifiés dans la note de cadrage et/ou dans les rôles attribués dans *Redmine*. Dans les situations où un membre de l'équipe ODR, hors pôle informatique, serait amené à participer à la maîtrise d'œuvre, une **machine virtuelle de développement** personnalisée sera mise à disposition sur son poste de travail.

Les projets informatiques (améliorations des applications, nouvelles applications) vont suivre un cycle de développement mis en place par le pôle informatique, dans le respect de bonnes pratiques (qualité) et de contraintes liées à la sécurité.
Des initiatives prises unilatéralement par des membres de l'équipe hors pôle informatique pour modifier le SI partagé sont à éviter car il serait difficile d'en apprécier la portée (effets de bords potentiels que le pôle informatique devra assumer) et d'éviter de possibles développements redondants ou contradictoires.

Le pilotage de la maitrise d'œuvre par le pôle informatique devra permettre :

- De maîtriser **globalement**, et non plus en organisation en silo, les actions appliquées aux données du SI partagé. Cela n'inclut pas les traitements des données qui seront réalisées sur un espace personnel de travail
- L'exploitation des ressources informatiques **en fonction de chaque étape identifiée** : (1) serveur de **développement** pour le codage et les tests unitaires, (2) serveur de tests fonctionnels pour la **recette**, (3) serveur de **production** pour le déploiement.
- La mise en place de **bonnes pratiques** de développement informatique (codage, conventions de nommages, etc.), de gestion des bases de données (identification, clés étrangères, etc.) et d'organisation des fichiers de scripts.
- Le **contrôle de version** des scripts via un **gestionnaire de versions** (***Git***) pour le partage et le suivi du code informatique.
- L'amélioration de la **traçabilité des traitements** réalisés sur les données (qui, quand, comment).
- Le renforcement de la **sécurité informatique** pour l'ensemble du système d'information par l'identifications des acteurs impliqués dans son évolution.

Ce pilotage par le pôle informatique implique une grande **réactivité** qui sera améliorée :

- par l'expérience acquise avec le cycle de vie des développements,
- par la mise en place du gestionnaire de tickets GLPI, qui permet de transmettre la globalité des demandes au pôle informatique. Les tickets feront l'objet d'un traitement à court termes pour le maintien en conditions opérationnelles ou d'une planification pour les demandes d'améliorations.

Cette organisation est indispensable pour permettre un suivi complet des modifications effectuées sur les instances partagées du SI. Elle ne concernera pas les travaux de développements (scripts, bases de données) localisés sur les postes personnels de l'équipe ODR.

Maintien en conditions opérationnelles

Le maintien en conditions opérationnelles du SISPA sera assuré par le pôle informatique de l'US-ODR.
Les demandes et relevés d'incidents, communiqués via le gestionnaire de tickets GLPI ou par un autre moyen (notifications directes de la plateforme web par exemple) seront ainsi concentrés et exploités ultérieurement en tant que **base de connaissances** par les administrateurs et développeurs du pôle informatique.

Communication interne

Le pilotage des futures évolutions du SI de l'US-ODR par le pôle Informatique requiert une **communication fluide et permanente avec l'ensemble de l'équipe de l'US-ODR**, au-delà des correspondances documentaires (documents qualités, tickets GLPI).
Cette communication sera maintenue lors des **réunions d'Unité quinzomadaires** et aussi lors des **réunions d'étape/bilan du Groupe de Travail Informatique (GTI)** dont la fréquence devra être augmentée à **1 réunion tous les 3 mois**.

Les comptes-rendus des réunions internes hebdomadaires du pôle Informatique seront aussi mis à disposition chaque semaine à l'équipe de l'US-ODR.

La FAQ du gestionnaire de tickets GLPI sera enrichie pour toutes les informations concernant les solutions de maintenances informatiques utiles à tous.

Vers une organisation fonctionnelle simplifiée

L'évolution de la plateforme ODR sera ambitieuse mais maîtrisée. Elle répondra à 2 grandes catégories de besoins : les **besoins internes** de l'équipe de l'Unité et les **besoins externes** des utilisateurs de la plateforme et partenaires de l'Unité qui souhaitent avoir accès au patrimoine informationnel de l'ODR.

Besoins Internes (Intranet)

La réception et la valorisation de données est une composante stratégique à haute valeur ajoutée du SISPA.
Les données primaires[13] récupérées auprès des partenaires de l'US-ODR (ASP, IGN, INAO, etc.) sont nettoyées, vérifiées, stockées et intégrées dans des bases de données. Ces actions ont vocation à être (semi-)automatisées, traçables et reproductibles.

Un axe majeur du développement du SISPA sera de mettre en place une **chaine de traitement** adaptée aux différents formats de données primaires réceptionnées et de produire de manière **standardisée** des données enrichies et valorisées sous la forme d'indicateurs originaux et de cartes géographiques aux finalités qui seront clairement définies. Cela concerne en premier lieu les productions dites standards[14] de l'unité.
Cette chaine de traitement devra permettre de **pré-calculer** la majeure partie des données spatialisées qui seront proposées aux utilisateurs de la plateforme.
Une première étape sera de développer un outil permettant **une assistance semi-automatisée** au nettoyage et à l'enregistrement des données primaires réceptionnées dans une base de donnée dédiée et préconçue, y compris pour les données administratives, et ce, à chaque actualisation (annuelle ou trimestrielle). Une telle automatisation étant dépendante de la qualité des données des fournisseurs (ASP par exemple), l'outil développé devra être **évolutif**. Toutes les étapes des traitements seront enregistrées pour **traçabilité** dès qu'un traitement sera appliqué aux données primaires.

« Une chaine de traitement devra permettre de pré-calculer la majeure partie des données spatialisées qui seront proposées aux utilisateurs de la plateforme »

[13] Données déposées par les partenaires fondateurs et/ou les tiers agréés – ou en leur nom.
[14] Une production standard est une base de données, et/ou une procédure, une carte… finalisée, avec version, documentée (métadonnées, note méthodologique, codes, datapaper …), associée à des droits de diffusion et à une ou plusieurs finalités identifiées. C'est une production qui peut être répétée (annuellement par exemple) et suffisamment générique pour répondre à des besoins différents, formels ou anticipés (explicites ou implicites). Elle se distingue donc des productions spécifiques , non figée dans le temps , à usage unique, par exemple dans le cadre d'un projet de recherche.

L'exploitation des données par les utilisateurs de la plateforme ODR est actuellement effectuée de manière dynamique au dépend d'une demande importante de ressources informatiques en termes de temps de calcul et de création de tables temporaires anarchiques dans les bases de données.
Ces traitements seront à l'avenir majoritairement effectués **par anticipation** et pour des finalités choisies et orientées (par exemple : nettoyage de données, agrégations, création de cartes), de manière automatisée en évaluant l'ensemble des combinaisons possibles nécessaires pour l'enrichissement des données et leur valorisation. Cette préparation des données permettra d'éviter au maximum les temps de calcul lors de leur visualisation.

La création de tables temporaires dans les bases de données sera donc exceptionnelle au profit de tables de grande capacité contenant les données pouvant être publiées.

Ces objectifs requièrent des compétences dans la conception de bases de données **relationnelles** avec une optimisation des relations entre les données s'appuyant sur la définition de clés étrangères, permettant ainsi d'obtenir un système de gestion de bases de données performant.

Besoins externes (Extranet)

Les besoins externes des utilisateurs partenaires de l'US-ODR, déjà mis en œuvre sur la plateforme actuelle, peuvent être regroupés en 2 grandes catégories : la mise en place d'une **zone de dépôt partagée** et **l'accès aux indicateurs thématiques des « observatoires ».**

Les partenaires fournisseurs de données devront pouvoir continuer à échanger des fichiers volumineux **de manière sécurisée et pratique**, que ce soit pour déposer leurs données primaires ou pour récupérer des résultats produits par l'ODR lors d'études ou projets de recherche.

Pour le dépôt de données par les partenaires, le service de transfert sécurisé RENATER ***FileSender***[15] sera à privilégier pour les personnes qui en ont l'autorisation et/ou qui y sont invitées. Le cas échéant ou en cas de défaillance du service *FileSender*, un service ***SFTP*** leur sera proposé.

L'organisation fonctionnelle thématique (cf. Contexte) de la plateforme ODR peut être conservée. Chacun des « observatoires » constituera un **module** du système d'information indépendant qui partagera des fonctionnalités communes avec d'autres modules, tout en visant l'interopérabilité entre observatoires afin de faciliter les usages croisés. Cette organisation sera enrichie par les nouvelles opportunités de valorisation de données dont :

- La valorisation de données de séquences temporelles de culture calculées à partir du Registre Parcellaire Graphique (RPG) grâce à l'outil gratuit *RPG Explorer* (collaboration avec l'UMR SADAPT), sous la forme de fichiers de données récupérables par les utilisateurs dans un premier temps puis sous la forme de modules d'exploitation (requêtes) de ces données sur la plateforme selon des paramètres agronomiques. L'objectif initial est de mettre à disposition les séquences de cultures en décembre 2021.

- La valorisation de données géo-référencées d'usage des produits phytosanitaires à partir de la base de données BNV-d dont l'enjeu principal est de permettre une connaissance exhaustive des usages des produits phytosanitaires pour mesurer les inégalités territoriales dans l'utilisation et donc l'exposition aux produits phyto.

[15] **https://www.renater.fr/fr/filesender**

La plateforme ODR, du fait de sa complexité, demande un apprentissage qui est un obstacle à sa convivialité et popularité[16]. Il est donc prévu une simplification de l'ensemble de la plateforme. Un effort d'accessibilité doit être fait sur les fonctionnalités proposées, en proposant un design plus « allégé » et une « aide » en ligne réécrite.

« La plateforme ODR, du fait de sa complexité, demande un apprentissage qui est un obstacle à sa convivialité et popularité. Il est donc prévu une simplification »

Comme c'est le cas sur la plateforme ODR actuelle, les informations disponibles pourront être mise à disposition dans un espace de travail réservé pour un utilisateur. Cet espace « Projet » restera la « propriété » de l'utilisateur et devra pouvoir être **facilement transférable** à un autre utilisateur selon des conditions à définir.
L'administration des accès à ces espaces devra être réajustée et simplifiée pour **remplacer la superposition actuelle des droits** (administrateurs habilités, membres, modérateurs des demandes de titularisation, modérateurs des demandes d'inscription ou d'accès à un programme, etc.).
Enfin, et de manière transversale à l'organisation thématique, le **catalogue de métadonnées** existant (***Agrilogue***[17]) sera mis à jour et intégré dans la nouvelle plateforme.

Base documentaire de la définition des besoins

Que ce soit pour les besoins intranet ou extranet, les choix, nouveautés ou évolutions applicatives devront être formalisés **dans plusieurs cahiers des charges**, conformément à la procédure qualité. Ces documents devront décrire les besoins de la manière la plus exhaustive possible avec, si possible, la définition des :

- axes d'amélioration,
- fonctionnalités existantes devenues obsolètes,
- algorithmes utilisés et/ou qui nécessitent une révision (filtres de secrets statistiques, nettoyages de données,etc.)

Le nombre de cahier des charges pourra correspondre au nombre de thématiques de la plateforme. Dans chaque cahier des charges seront déclinés les besoins en termes de zone de dépôt, de traitement de données et de présentation des indicateurs et fichiers valorisés. **Chaque chargé de mission participera à la rédaction d'un cahier des charges pour son domaine d'expertise** (cf. organisation fonctionnelle des données).

Un **cahier des charges « transversal »** devra prévoir les besoins en termes de niveaux d'accès à la plateforme, de gestion des utilisateurs, de publication d'informations générales (présentation de l'ODR, équipe, etc.)
Ces premiers documents seront par la suite complétés par un **Dossier de Spécifications Fonctionnelles** unique qui deviendra en quelque sorte un catalogue global des fonctionnalités attendues, certaines pouvant se révéler **transversales** entre les différentes thématiques, et pour lesquelles des solutions seront formalisées.

[16] **Rapport établi par la commission d'évaluation Evaluation Unité de service INRA US-ODR, le 21 mars 2011**
[17] **https://odr.inra.fr/agrilogue/**

Vers une nouvelle organisation des bases de données

Contraintes liées à la sensibilité des données :

Dans tout SI, selon la nature des données, leur mise à disposition peut être obligatoirement libre (données publiques), interdite (données sensibles), ou soumise à conditions (données confidentielles).
Les informations hébergées au sein du SI de l'US-ODR présentent des niveaux de confidentialité qui peuvent être classés par ordre croissant :

- Les **données primaires publiques** (-),
- Les **données secondaires traitées**, vérifiées et travaillées par l'équipe ODR et pseudonymisées sauf dans le cadre de conventions spécifiques permettant de traiter directement les données non anonymes (++)
- Les **données primaires confidentielles** (ASP, MSA, INAO, etc.) (+++). Ces données restent la propriété des fournisseurs qui en sont juridiquement propriétaires.

Il en résulte la nécessité de prévoir 3 catégories d'organisation de bases de données sécurisées et **indépendantes** :

- Les bases de données contenant les données primaires publiques.
- Les bases de données de travail à partir des données primaires propriétaires, conçues pour les interfaces de traitements (statistiques, agrégations, etc.) et vouées à être purgées pour être archivées de manière sécurisée post-traitements.
- Les bases hébergeant les données secondaires traitées (pseudonymisation et secret statistique) et valorisées.

Inventaire des principales données confidentielles fournies par les partenaires :

Les données confidentielles exploitées sont :

- De la part de la **Caisse Centrale de la Mutualité Sociale Agricole**:
    - Fichiers de cotisants non-salariés de la MSA (COTNS), annuellement depuis 2002.
    - Données des contrats salariés de la MSA (SISAL), annuellement depuis 2002.
- De la part de **l'Agence de Services et de Paiement**:
    - Les données d'engagements et de paiements se rapportant aux PDR des programmations 2000-2006, 2007-2013, 2014-2020 et suivante
    - Un ensemble de données individuelles communes aux bénéficiaires de tous les dispositifs, mises à jour tous les 3 mois
    - Des données spécifiques sur la mise en œuvre de chaque type de dispositif des Règlements de Développement Rural (RDR)
    - Le RPG dit version 1 et de niveau 4 (anonyme), pour la période 2006-2014
- De la part de **l'Institut National de l'Origine et de la qualité** :

- Données des opérateurs habilités SIQO, annuellement depuis 2011
- Données certification en Agriculture Biologique
- Données économiques sur les SIQO

- De la part du **Ministère de la Transition Ecologique**: couches d'informations géographiques
- De la part du **Ministère de l'Agriculture et de l'Alimentation** : le RPG dit version 2 et de niveau 2 depuis 2015, fourni en tant que prestataire du MTE dans le cadre du projet de spatialisation de la BNVD (Ecophyto)
- De la part de l'**Institut national de l'environnement industriel et des risques (INERIS):** la Banque Nationale des Ventes distributeurs (BNV-d) de produits phytosanitaires dans le cadre du projet **Ecophyto**.

De nouveaux modèles de données[18] :

Base de données contenant les données primaires[19] (intranet) :

Le modèle de données des bases de données hébergeant les données primaires publiques ou propriétaires **sera inspiré du formalisme inclus dans les fichiers fournis** par les partenaires, avec la souplesse requise par les évolutions prévisibles. **Ces données seront archivées et purgées post-traitements.**

Base de données de travail à partir des données primaires (intranet) :

Parmi les fonctionnalités de l'Intranet de l'ODR, la pseudonymisation et la « secretisation[20]» des données primaires confidentielles permet de rendre les données traitées moins sensibles. Pour faciliter la pseuonymisation, il sera proposé de **regrouper toutes les données personnelles dans un nombre limité de tables** dès l'enregistrement des données primaires, en attribuant un identifiant numérique unique (clé primaire) dans les relations aux autres tables (pseudonymisation).

Ces bases seront archivées et purgées post-publication sur l'extranet.

« Le modèle de données devra d'avantage intégrer la composante relationnelle des SGBDR »

Le modèle de données conçu pour ces nouvelles bases de données qui enregistreront les données traitées devra intégrer davantage la composante **relationnelle des SGBDR**.

Cela implique la définition de clés primaires et de clés étrangères au travers d'**objets**[21] dont le choix et la pertinence détermineront l'interopérabilité des données pour créer un système d'information plus global et performant.

La conception d'un modèle de données relationnel repose sur la définition des objets, chaque objet correspondant à une table de base de données, et est caractérisé par une liste **d'attributs** et des **relations** définies par les clés étrangères.

Les attributs s'inspireront d'un **catalogue d'indicateurs**. Ce catalogue pourra être amélioré à partir des informations fournies par la plateforme ODR actuelle au niveau du répertoire des dossiers thématiques, du répertoire des données primaires et du répertoire des données géographiques (couches géographiques géocodées + couches d'habillage).

Techniquement les clés étrangères devront toujours être **un identifiant informatique** (jamais un attribut métier).

Enfin, pour anticiper les éventuelles évolutions à l'international (Europe), les modèles de données devront prévoir, pour les objets concernés, un attribut correspondant au pays, renseigné de préférence en anglais.

[18] Modèle qui décrit la manière dont sont représentées les données dans un système d'information

[19] Données déposées par les partenaires fondateurs et/ou les tiers agréés.

[20] La notion de secret statistique est définie en référence à l'article 11.4 de la Charte de qualité des enquêtes de branche dans l'industrie approuvée par le comité du label du *Conseil national de l'information statistique* (CNIS), configuration « Entreprises », lors de sa séance du 19 mars 2001. Il est exclu de publier des informations concernant des communes ou des groupes d'individus dès lors que le seuil du secret statistique (3 individus) n'est pas atteint.

[21] Conteneur symbolique et autonome qui contient des informations concernant un sujet manipulé dans un programme.

Bases de données hébergeant les données secondaires traitées et valorisées (extranet) :

Le modèle de données conçu pour les données secondaires traitées et valorisées sur l'extranet sera **équivalent** à celui récoltant les résultats des traitements sur l'intranet, **à l'exception des données personnelles**. Le contenu sera quant à lui plus important puisqu'enrichi après chaque campagne de traitement des données. Le volume de ces bases de données contenant les données calculées pourra atteindre les volumes actuels occupés par les bases de données « cache ».

En raison de finalités différentes, le modèle de données de l'extranet devra comporter 2 partie distinctes :

- **Le cœur**, composé des objets critiques pour le bon fonctionnement de la nouvelle plateforme : tables utilisateurs, tables de configurations, etc.
- **Les objets métier**, sur lesquelles la gestion des clés étrangères sera la plus importante.

Vers une nouvelle organisation technique

Complémentaires à la nouvelle organisation fonctionnelle, les solutions informatiques proposées doivent correspondre aux usages communément employés pour les applications web, et aux compétences acquises ou en cours d'acquisition en informatique, en programmation web, visualisation de couches géographiques et gestion de base de données. Toutes ces solutions seront gratuites et open source.

Les solutions décrites ici concernent les ressources partagées. Elles ne concernent pas les ressources individuelles que chacun est libre d'utiliser sur son poste de travail. Des choix de langages informatiques sont donc proposés pour les ressources partagées qui n'excluent donc pas l'utilisation d'autres langages sur les postes de travail individuels.

Ressources logicielles pour l'Intranet

Techniquement, le SI « Interne » sera indépendant et sera :

- **Déployé sur des serveurs dédiés sécurisés et isolés.**
- **Développé soit avec les technologies *PHP/Symfony*[22], soit avec les technologies *R/Shiny*** [23]:
  Le choix du **langage *R*** devra être privilégié **si les traitements statistiques partagés deviennent complexes** tels que les analyses de variables (distribution, discrétisation), les intervalles de confiance, les tests d'hypothèses, etc. Le langage *R* peut être privilégié sur les ressources partagées à condition de prévoir une couche applicative supplémentaire pour palier aux imperfections des packages d'authentification-utilisateur via *R* ou *Shiny* (sauf dans le cas de l'utilisation d'une version payante *RStudioConnect*). Ce choix n'est pas nécessaire si la pratique de statistiques complexes est essentiellement faite **sur des postes informatiques individuels**. Le langage *R* peut être écarté pour des traitements de statistiques descriptives possibles avec le langage *PHP*.
  Le **langage *PHP*** peut être privilégié sur les ressources partagées en raison de sa souplesse et la richesse de ses bibliothèques qui vont au-delà de traitements statistiques : gestion de la sécurité, des utilisateurs, des connectivités (LDAP), des notifications mail, etc. La **version 7** de PHP est à ce jour un bon compromis stabilité/pérennité. Enfin, l'utilisation d'un **Framework** comme ***Symfony***, outre les nombreux avantages en terme d'efficience pour les développeurs, présente l'avantage d'inciter au regroupement des fichiers constituant une application web par type: images, modèles de documents, etc...
  Quel que soit le choix, ces langages sont communément associés au langage de requêtes ***SQL***.
  Il conviendra **d'éviter l'accès à des ressources distantes** (javascript, feuilles de styles) pour une maitrise complète des contenus et pour améliorer l'empreinte carbone du SI.

[22] **https://symfony.com/**
[23] **https://shiny.rstudio.com/**

Ressources logicielles pour l'Extranet

Techniquement, le nouveau SI « Externe » sera indépendant et sera :

- **Développé avec les technologies *PHP* version 7/*Symfony* version 4**,
- **Déployé sur deux serveurs dédiés** : l'un présentant une interface **de dépôt de données primaires**[24], l'autre contenant le **portail de présentation des indicateurs** de l'US-ODR.

La **visualisation** et la diffusion des contenus (tableaux, cartes, fichiers) pourrait se faire :

- Soit à travers le maintien du ***CMS JOOMLA***,
- Soit sans ce *CMS* à condition de consacrer du développement à l'ergonomie via des « templates[25] » correspondants au vues gérées par le Framework ***Symfony***, en les combinant avec l'héritage multiple, la surcharge d'opérateur et la programmation orientée objet. Ces atouts sont possibles avec le moteur de templates ***twig*** fourni avec *Symfony*, qui interagit avec les bibliothèque *Javascript* ***Jquery*** ou Bootstrap permettant de créer des pages Web évoluées affichant des graphiques interactifs
- Soit avec une nouvelle solution de *CMS* : ***Drupal***[26], dont les dernières versions 8 et 9 intègrent des composants *Symfony* et le moteur de templates *twig*, ce qui peut être un bon compromis. De plus ***Drupal*** facilite le développement pour les sites sur terminaux mobiles. A noter que le dictionnaire actuel de métadonnées « *Agrilogue* » a été conçu avec *Drupal*.

L'affichage d'indicateurs spatialisés à travers des cartes interactives (*WebMapping*) pourra être envisagé avec plusieurs solutions qu'il conviendra de tester et comparer:

- Soit la mise à jour de ***MapServer*** dont la dernière version à ce jour est la 7.6.[27] Cette solution ne présente pas de compatibilités reconnues avec le framework *Symfony*.
- Soit avec la bibliothèque javascript ***Leaflet***[28] et le projet ***OpenStreetMap***[29]. Cette solution a déjà été intégrée dans des applications développées avec Symfony. La bibliothèque javascript *Leaflet* présente l'avantage de pouvoir être intégrée à la fois dans une application ***PHP/symfony*** et avec le package ***R Shiny***.
- Soit avec la bibliothèque javascript ***OpenLayers***[30] . Cette solution, compatible avec le langage PHP, a été intégrée avec le FrameWork *Symfony* dans un bundle intéressant : ***Mapbender***. Cependant, la dernière version de ce bundle fait appel à *Symfony* version 2, obsolète.

[24] Données déposées par les partenaires fondateurs et/ou les tiers agréés.
[25] Thème ou encore modèle graphique permettant de séparer le contenu rédactionnel (contenu textuel) de la forme (la manière dont il est présenté)
[26] https://www.drupal.org/
[27] https://mapserver.org/
[28] https://leafletjs.com/
[29] https://www.openstreetmap.org/
[30] https://openlayers.org/

Systèmes de Gestion de Bases de Données (SGBD)

Pour des raisons de reprise au moins partielle de l'existant, le choix peut être porté sur l'un des deux SGBD libres actuellement utilisés: ***MySQL*** et ***PostgreSQL***.

PostgreSQL est le SGBD pouvant être géré par les logiciels SIG ***QGIS*** et ***PostGIS***. Il demeure donc un choix incontournable dans le SI de l'US-ODR.

Les ressources logicielles choisies pour le nouvel Extranet vont déterminer l'utilisation de MySQL conjointement à PostgreSQL:
- si l'aspect visuel de la plateforme est mis en page avec le ***CMS Joomla***, le SGBD ***MySQL*** sera requis car il est préconisé avec Joomla (même si Joomla est compatible avec PostgreSQL, cette configuration est très peu documentée).
- si l'ergonomie est conçue avec le Framework ***Symfony*** et/ou avec ***Drupal***, le SGBD ***PostgreSQL*** seul **peut suffire**.

Dans le cas où le SGBD *MySQL* serait maintenu il conviendra de le configurer de sorte à ce que le moteur de stockage soit ***InnoDB*** (et non le moteur par défaut *MyISAM*). En effet seul le moteur de stockage *InnoDB* gère les **clés étrangères** et les contraintes d'intégrité tel que souhaité pour la conception d'un nouveau modèle de données **relationnel**.
Le principal inconvénient du moteur *InnoDB* étant d'avoir la réputation de requérir davantage de ressources comparé au moteur *MyISAM* lorsque l'ensemble des bases et tables est géré dans 1 seul fichier (configuration par défaut du moteur *InnoDB*), il conviendra d'**améliorer les performances en paramétrant 1 fichier par base ou table prévue**

Architecture réseau:

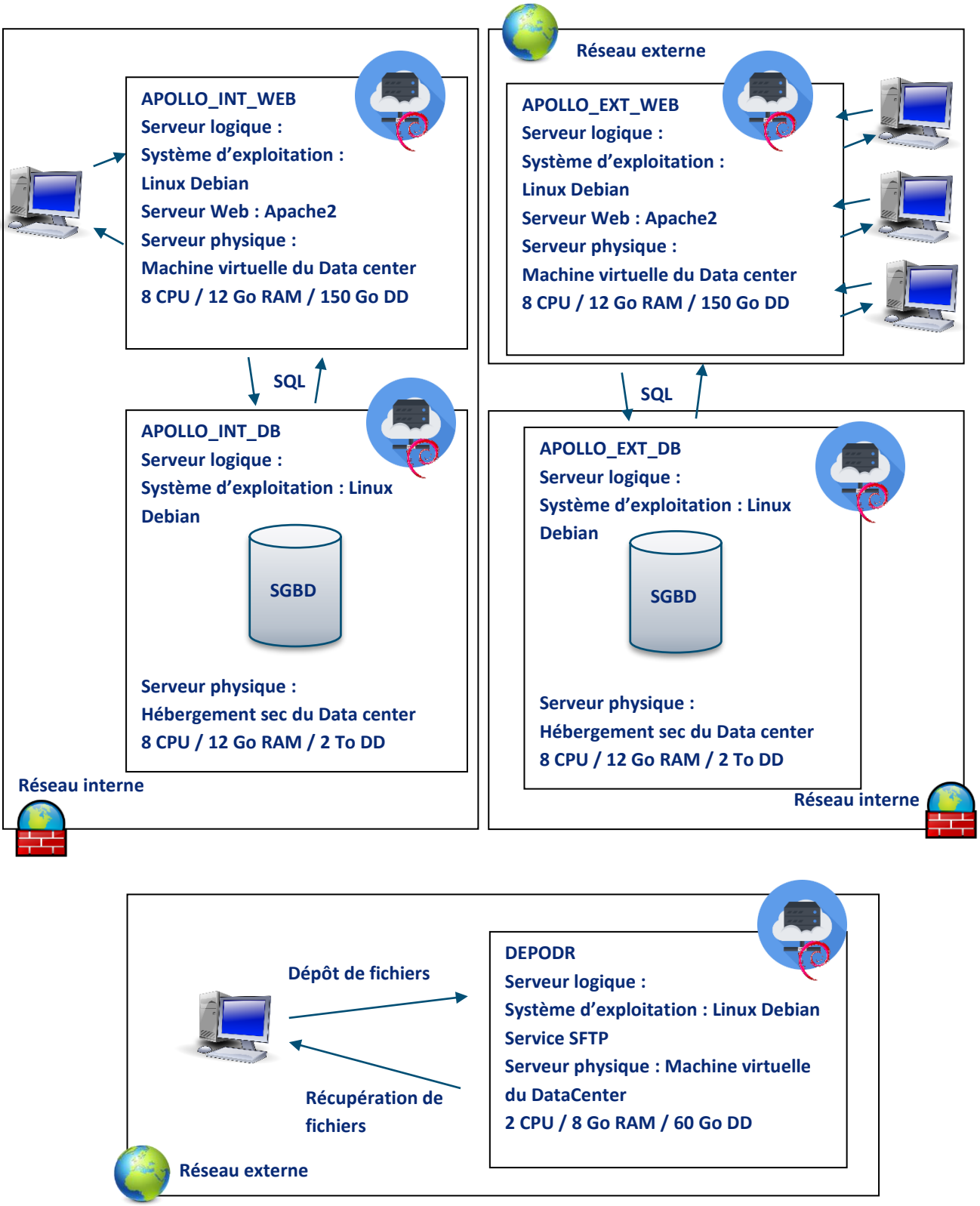
Concepts de l'organisation des serveurs :

- Un **serveur de développements** est le serveur de travail pour les développeurs sur lequel est « rédigé » ou modifié le **code** d'une application. Il héberge aussi les **bases de données** en cours de conception. Ce serveur sert également aux **tests unitaires** effectués en cours de développements. Il n'a pas vocation à être allumé en permanence (par ex. nuit et jours, week-end et jours fériés), ce qui permet de **réduire l'empreinte carbone du SI**. Il peut prendre la forme d'une **machine virtuelle** utilisée localement.
- Un **serveur de tests ou recette** est une copie du serveur utilisé en production sur lequel vont être testés (**tests fonctionnels**) et validés les développements faits sur le serveur de développements. Ce serveur n'a pas vocation à être allumé en permanence ce qui permet de **réduire l'empreinte carbone du SI**. Il peut aussi prendre la forme d'une **machine virtuelle** utilisée localement à condition qu'elle puisse être accessible en réseau.
- Un **serveur de production** est le serveur de **déploiement** des développements validés et qui constituent les applications accessibles aux utilisateurs finaux. Il héberge les bases de données dont le modèle de données a été validé et contenant l'ensemble des données destinées aux utilisateurs. **Ce serveur a vocation à être allumé sans interruption de service** (ou interruption exceptionnelle lors de maintenances). Il doit prendre la forme d'une **machine virtuelle hébergée dans un centre de ressources informatiques continues ou sur un serveur ondulé.**

Ressources matérielles pour les environnements de production (architecture 3-tiers) :

Un serveur est le maillon faible du réseau client-serveur. Il conviendra de maintenir des solutions hébergées par le Data Center qui présentent de grandes tolérances aux pannes (système RAID) et des moyens déployés suffisants pour le sécuriser (local sécurisé, onduleur, etc.).
Pour respecter les préconisations des solutions techniques, l'organisation des serveurs continuera à être centralisée et une **architecture 3-tiers** sera mise en place, que ce soit pour l'**intranet** (travaux de traitement de données) qui requiert d'importantes ressources de calculs pour un nombre limité d'utilisateurs, ou pour l'**extranet** (valorisation des données) dont les ressources seront adaptées à un contexte de moindre puissance de calculs mais avec un grand nombre d'utilisateurs.
Cette architecture permet d'améliorer le temps de réponse, d'augmenter la charge, la disponibilité des applications et la robustesse du SI.
En complément, un serveur supplémentaire (architecture 2-tiers) sera prévu pour la couche applicative dédiée au **dépôt des données**.

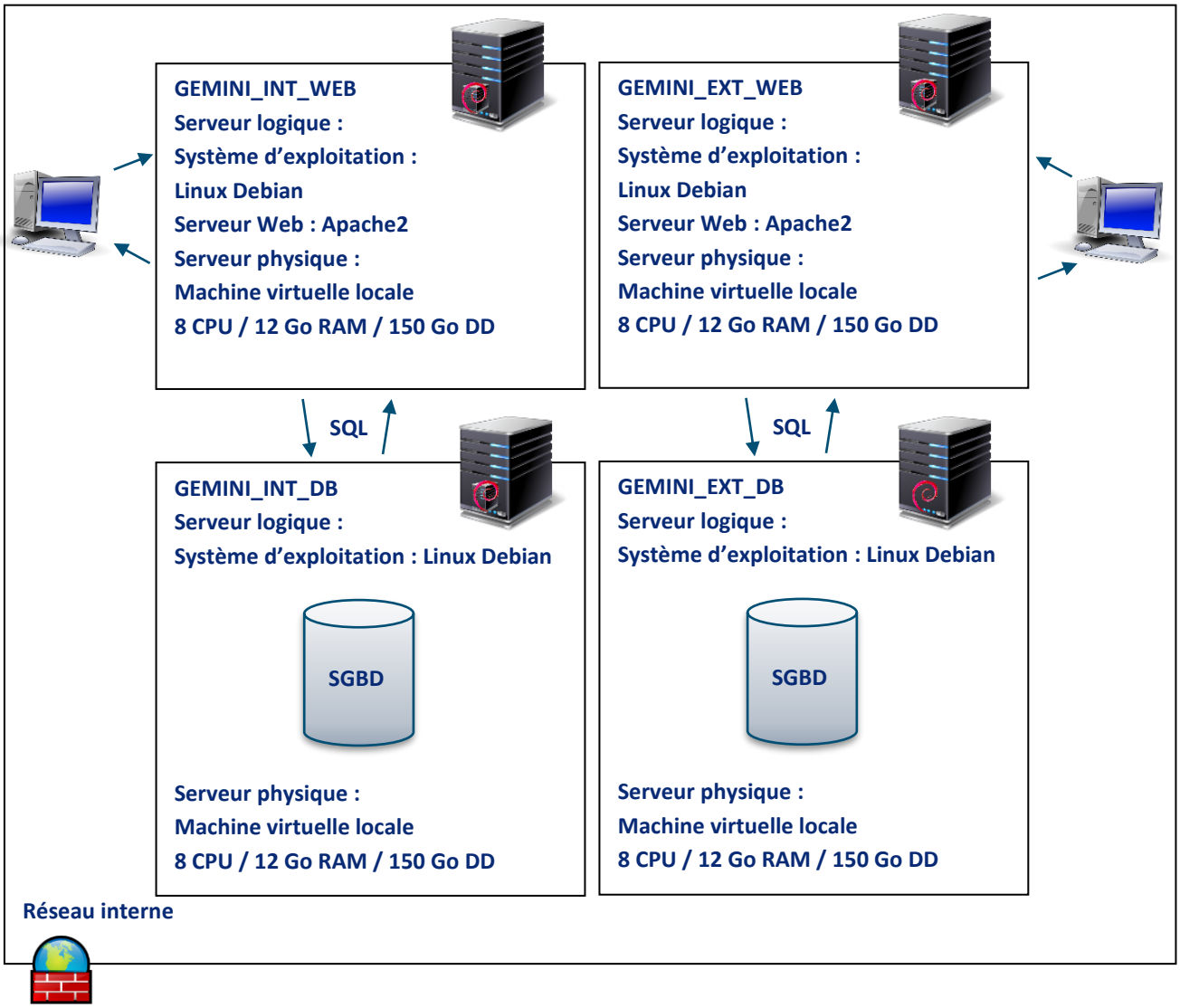
Ces ressources matérielles **représentent 5 serveurs** (hébergements secs ou machines virtuelles) en environnement de production: 2 pour les travaux de traitement de données en intranet (**APOLLO_INT_WEB + APOLLO _INT_DB**), 2 pour la valorisation des données en extranet (**APOLLO_EXT_WEB + APOLLO_EXT_DB**) et 1 pour le dépôt de données (**DEPODR**):



Ressources matérielles pour les environnements de recette (architecture 3-tiers) :

Les environnements de recette doivent être **identiques à ceux de production** afin de valider les évolutions en situations réelles. Seul l'espace de stockage des données peut être réduit car les tests peuvent être faits avec un **échantillon de données**.

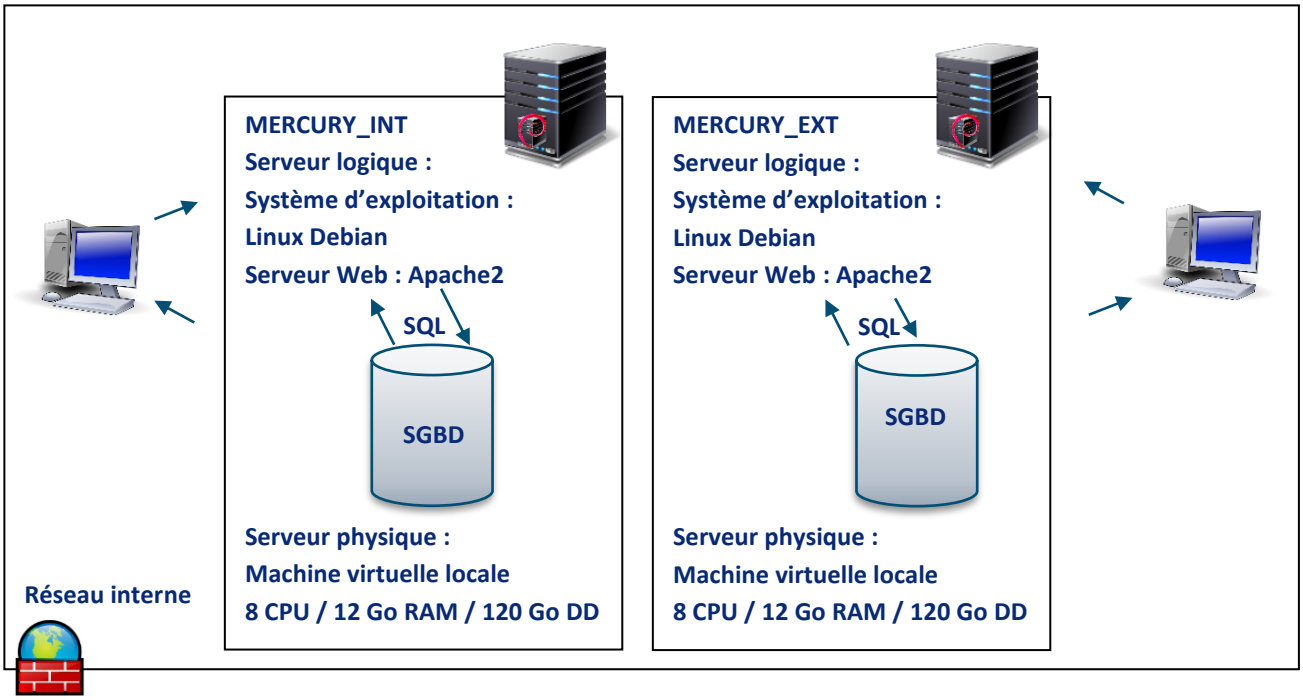
Les ressources matérielles **représentent 4 serveurs** (machines virtuelles) en environnement de tests 3-tiers: 2 pour les évolutions de l'intranet (**GEMINI_INT_WEB + GEMINI_INT_DB**) et 2 pour les évolutions de l'extranet (**GEMINI_EXT_WEB + GEMINI_EXT_DB**):



Ressources matérielles pour les environnements de développements (architecture client-serveur) :

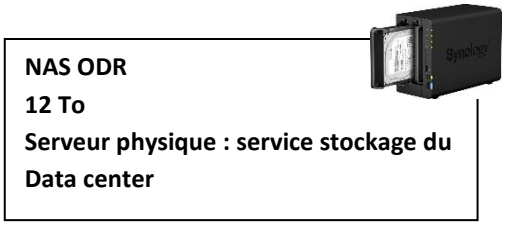
Une **architecture 2-tiers** est suffisantes dans un contexte de développements pour l'intranet et l'extranet, hébergeant la couche applicative pour le traitement-valorisation des données **ET** le SGBD utilisé. Cette architecture est ainsi aisément manipulable par le biais d'une machine virtuelle unique qui peut être mises à disposition par le pôle informatique sous forme de ***VirtualBox***[31] locales pour les membres de l'équipe ODR souhaitant participer ponctuellement à des développements.

Les ressources matérielles **représentent 2 serveurs** (machines virtuelles) en environnement de développements: l'un pour les évolutions de l'intranet (**MERCURY_INT**) et l'autre pour les évolutions de l'extranet (**MERCURY_EXT**):



Ressources matérielles pour l'environnement de sauvegardes :

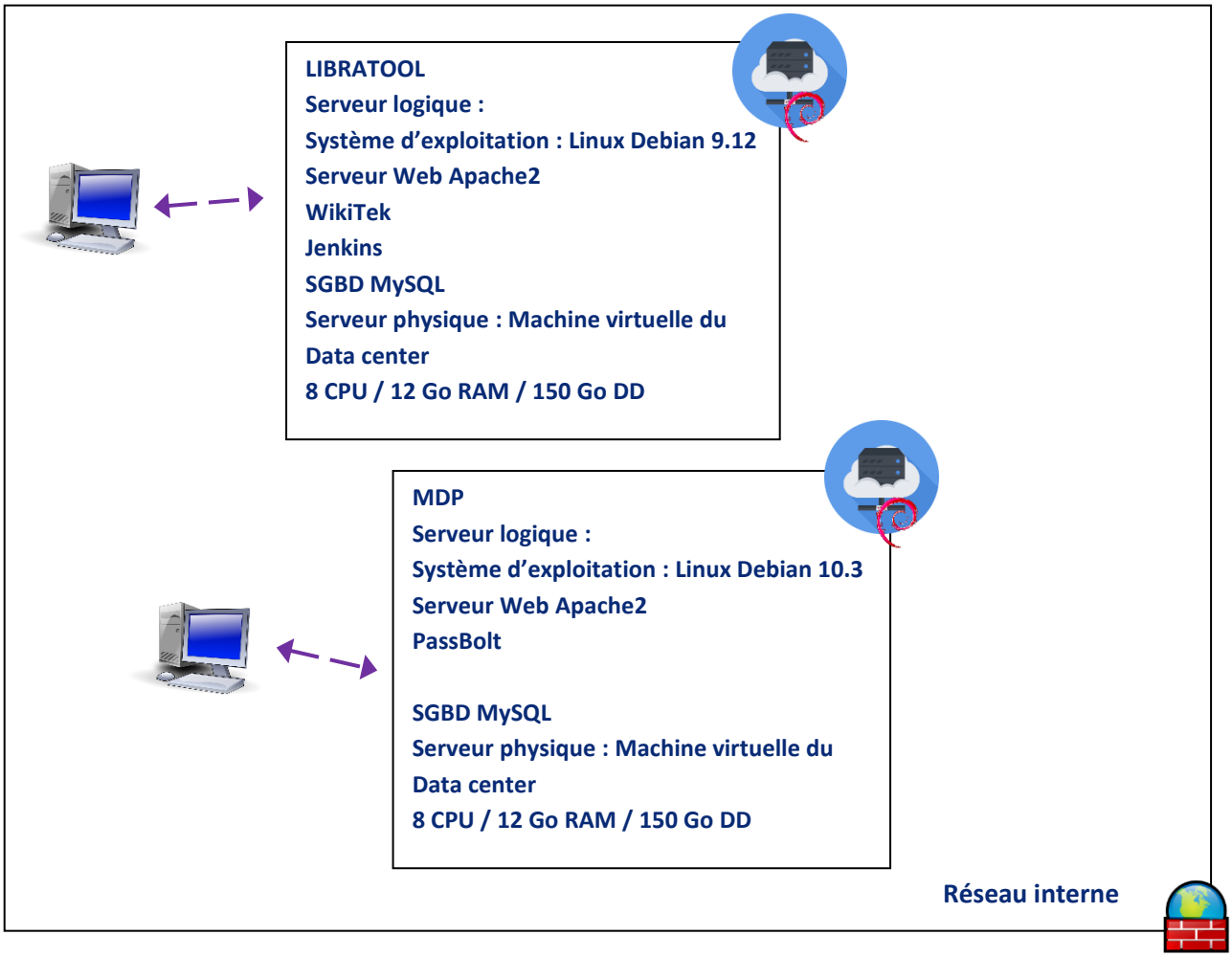
Les applications originales développées pour l'unité et les données devront être sauvegardées sur un **serveur de stockage en réseau (NAS) de capacité de 12 To.**



[31] **https://www.virtualbox.org/**

Ressources matérielles pour le support Informatique (architecture client-serveur):

L'architecture actuelle sera conservée :



Sécurité

La gestion de la sécurité informatique[32] sera précisée dans une PSSI « **Politique de sécurité des systèmes d'information** » rédigée par le pôle informatique. Les principes généraux sont exposés dans ce schéma directeur, pour la gestion des accès intranet et extranet, la gestion des sauvegardes et archivages**.**

Gestion des accès à l'intranet

Selon la Direction Général de la Sécurité Intérieure, la moitié des pertes d'informations dans un SI est due à des négligences ou des malveillances internes[33]. Selon IBM[34], 27 % des violations de données en 2018 sont dues à des erreurs humaines et selon Verizon[35], 28 % des piratages commis en 2018 impliquaient des personnes internes.

« Statistiquement la moitié des pertes d'informations dans un SI est due à des négligences ou des malveillances internes »

Parce que le SI de l'US-ODR regroupe des données confidentielles, une meilleure gestion des mots de passe est nécessaires.
Cette amélioration de la gestion de mots de passes se fera en 6 étapes principales :

- La mise en place d'un **gestionnaire de mots de passe sécurisé**, afin de bannir l'inscription de mots de passe sur des supports non sécurisés (registres papier ou fichier informatique non crypté).
- La création de **comptes nominatifs** pour chacun des besoins, y compris pour les comptes administrateur. Les autorisations de connexion seront toujours données nominativement.
- L'identification des comptes génériques obsolètes et/ou inutiles et leur **désactivation**.
- La **modification des mots de passe actuels pour les comptes génériques** qui doivent être maintenus.
- La **redéfinition des droits d'accès sur les serveurs linux** selon les environnements (développement/recette/production), en prenant soin de limiter les points d'entrée :
    - Environnements de développements : accès *sudo* pour les comptes nominatifs.
    - Environnements de tests (recette) : accès avec les comptes nominatifs sans *sudo*.
    - Environnements de production : accès limité aux membres du pôle informatique (*sudo*).
- La **définition de droits d'accès nominatifs à la plateforme web intranet**.

[32] **https://intranet.inrae.fr/cybersecurite/**
[33] **Bilan de la réunion avec la Direction Général de la Sécurité Intérieure le 04/03/2020**
[34] **https://www.ibm.com/downloads/cas/861MNWN2**
[35] **https://enterprise.verizon.com/resources/reports/DBIR_2018_Report.pdf**

Gestion des accès à l'extranet :

Trois niveaux d'accès à la plateforme extranet peuvent être maintenus :

- Un niveau d'accès **public sans création de compte**,
- Un niveau d'accès **utilisateur qui requiert la création préalable d'un compte**, ce compte devant être paramétrable en fonction des droits d'accès (thématiques, avec agrément ou non, etc.)
- Un niveau d'accès **administrateur**.

Le niveau d'accès utilisateur avec compte permettra, en fonction des droits :

- Soit une « **co-administration** » limitée dans le cadre d'accords passés avec les partenaires, comme c'est le cas actuellement. Les comptes de « co-administration » auront accès à la zone de **dépôt de données**.
- Soit un accès à des programmes/projets/dossiers spécifiés par les « co-administrateurs » partenaires aux utilisateurs de ces mêmes partenariats.

Afin de garantir la maitrise des comptes d'utilisateurs et pour sécuriser les échanges et la protection des informations, **le niveau d'accès administrateur sera limité à des membres titulaires de l'équipe ODR et au pôle informatique.**
Les comptes ayant un niveaux d'accès permettant la « co-administration » seront toujours gérés par les comptes à niveau d'accès administrateur. Les droits des comptes à niveau d'accès administrateur hériterons des droits de tous les comptes.

Sauvegardes

La gestion des sauvegardes passera par l'établissement d'un **plan de sauvegardes** (document qualité) qui comprendra la méthode de sauvegarde, le lieu de sauvegarde ainsi que la fréquence des sauvegardes.

Il existe 3 **méthodes de sauvegarde**:
Sauvegarde complète ou totale : sauvegarde totale de tous les fichiers sans distinction de date ou d'évolution.
Sauvegarde incrémentielle ou incrémentale : sauvegarde de tous les fichiers modifiés ou créés depuis la dernière sauvegarde complète ou incrémentielle.
Cette sauvegarde se réfère à la première sauvegarde totale puis aux sauvegardes incrémentielles successives.
Points de restaurations : sauvegarde de l'état entier du serveur au moment où il est déclenché, le point de restauration comprendra le contenu de la mémoire du serveur ainsi que ses paramètres et l'état des disques durs. C'est le cas par exemple des ***Snapshots***[36] de stockage des machines virtuelles effectués par les gestionnaires du DataCenter.

Préconisations pour la **fréquence des sauvegardes** :

Une sauvegarde des données situées sur les **serveurs de production et le serveur de support informatique** doit être réalisée toutes les 24h avec une rétention d'une copie toutes les semaines sur 3 semaines, ce qui fait 4 lots de sauvegardes au total.
Un **journal de sauvegarde** doit être édité, celui-ci indique la durée de la sauvegarde, sa taille ainsi que les erreurs éventuelles.
Un courrier électronique sera automatiquement envoyé aux membres du pôle informatique après l'action de sauvegarde. L'absence de réception de ce courrier ou la présence d'une erreur dans ce courrier implique un traitement par le pôle informatique.

La planification des sauvegardes des données sur les **serveurs de tests** sera identique à celle des serveurs de production mais ne sera effective que lorsque ces serveurs seront utilisés, lors des phases de tests fonctionnels.

Les **serveurs de développements** ne seront pas automatiquement sauvegardés. Il est préconisé aux utilisateurs de ces serveurs de développements de conserver leurs travaux (scripts, échantillons de données) dans le répertoire « Mes documents » qui est synchronisé avec un NAS et dont le contenu est sauvegardé durant 1 mois.

Restauration des sauvegardes :

Lorsque des données sont manquantes et qu'une anomalie est détectée, une demande de restauration devra être effectuée si les supports de sauvegarde ou d'archivage ne sont pas de la responsabilité de la personne constatant l'anomalie.
Toute demande de restauration de données ou d'un logiciel (réinstallation) se fera par le biais du gestionnaire de tickets GLPI.
Cette intervention peut être effectuée par le pôle informatique ou un gestionnaire du DataCenter.

Une application de gestion de sauvegardes comme ***Bareos***[37] devra être installée pour l'administration des sauvegardes et des restaurations de données pour l'ensemble des serveurs concernés vers un NAS.

Archivages des données

La décision d'archivage peut intervenir lorsque la présence des données n'est plus obligatoire sur les serveurs de l'Unité et que l'encombrement de ces données oblige à les transposer sur un autre support afin de permettre l'accès à des données plus récentes. L'archivage est aussi recommandé pour accroitre les performance d'un SGBD dont les bases de données seront allégées d'autant de données archivées.

Les supports d'archivage peuvent être **déportés** (NAS, technologie cloud sécurisée) **ou conservés localement**. La deuxième solution présente l'avantage d'être **moins énergivore et plus respectueuse de l'environnement**. Dans ce cas les supports d'archivage sont usuellement des supports électroniques amovibles tels que disques optiques, clés USB, disques durs externes, qui peuvent être stockés dans un **compartiment sécurisé** lorsqu'ils contiennent des données confidentielles (armoire forte, coffre).
Un support d'archivage amovible est identifié par marquage physique : étiquetage, écriture à l'encre indélébile, etc.
Le choix se fait selon sa capacité, sa vitesse, sa fiabilité, la simplicité d'utilisation, la facilité de restauration et le coût.
Les supports d'archivage ont des durées de vie limitées dans le temps (de quelques mois pour des disques optiques en usage intensif à plusieurs dizaines d'années pour les disques SSD bien conservés). Le système d'archivage doit donc prendre en compte leurs paramètres physiques pour garantir la conservation des données archivées pendant toute la durée prévue d'archivage.

L'archivage de données confidentielles (données primaires par exemple) sur des supports amovibles devra être crypté dans une archive *VeraCrypt*[38] par exemple ou dans un conteneur *ZED !*, outil préconisé[39] par l'Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI).

[36] « Instantané » (anglais snapshot) : sauvegarde de l'état d'un système à un instant donné.
[37] https://www.bareos.org/en/
[38] https://www.veracrypt.fr
[39] https://www.ssi.gouv.fr/administration/qualification/zed/

RGPD

En référence avec la politique de l'Institut, pour les données de l'ODR, la loi européenne en matière de **Règlement Général sur la Protection des Données** (RGPD) s'applique **pour les données personnelles individuelles**. Elle s'applique également pour les exploitations agricoles « personnes morales », souvent identifiées par leur numéro de Siret, lorsque, derrière la personne morale il y a une seule personne physique (l'exploitant ou un ménage). **Les exploitations agricoles unipersonnelles doivent donc être protégées par le RGPD[40].**

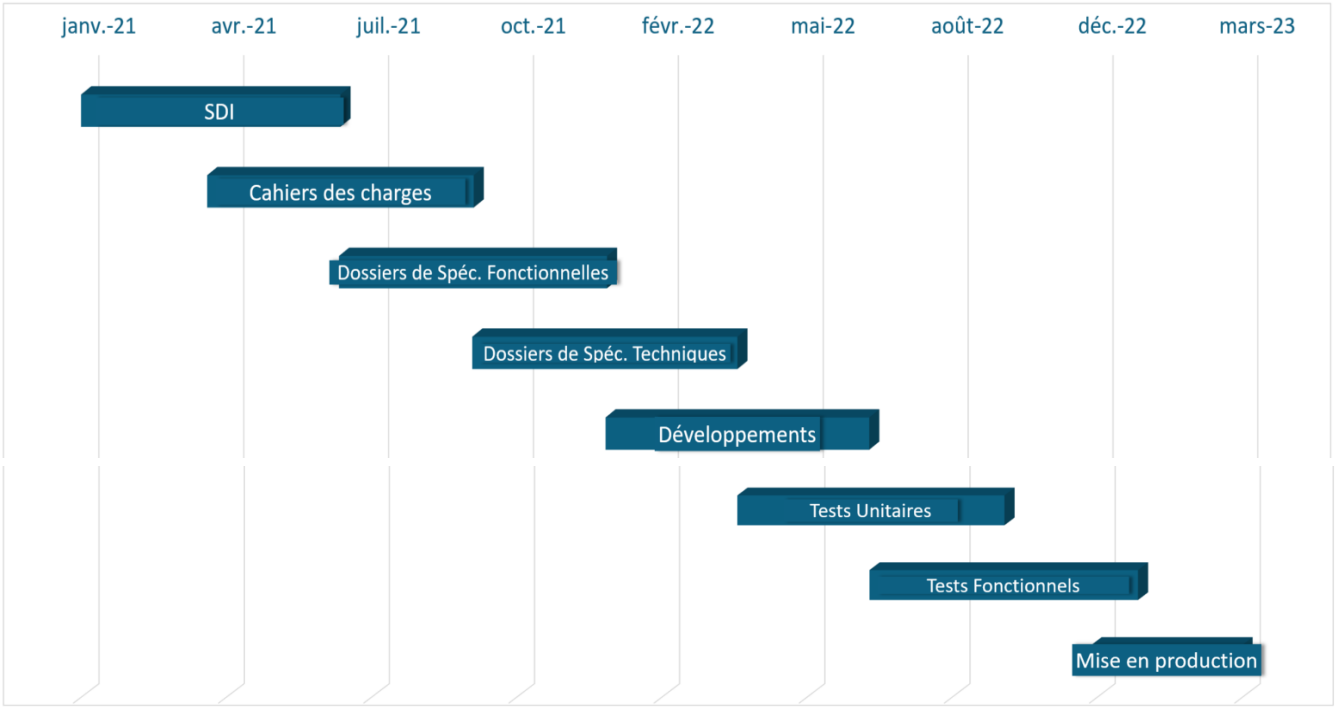
Des points de vigilance RGPD vont être appliqués pour la nouvelle plateforme du SISPA, que ce soit pour les développements informatiques[41] ou pour le stockage des données:

- Intégration de la protection de la vie privée, y compris les exigences de sécurité des données, **dès la conception des applications**.
- Conception d'un système **initialement simple** permettant d'en comprendre précisément les rouages et d'identifier ses points de fragilité, **puis en augmenter la complexité petit à petit**, tout en continuant de sécuriser les nouveautés qui s'ajoutent.
- Utilisation de normes de programmation prenant en compte la sécurité (bonnes pratiques ou guides de codage). Les données ne doivent pas être l'objet de fuites (**principe d'intégrité et de confidentialité**).
- Développement en évitant de coder dans un langage tout juste appris et pas encore maîtrisé.
- Une analyse d'impact à l'aide de l'outil **Privacy Impact Assesment (PIA[42])** de la Commission nationale de l'informatique et des libertés (CNIL) sera effectuée lors de la phase de spécifications fonctionnelles pour la nouvelle plateforme, en s'inspirant du travail d'analyse d'impact sur la vie privée déjà effectué sur la plateforme actuelle.
- **L'ajout d'informations légales** : il s'agira de mettre à jour les mentions légales et les conditions générales d'utilisation (ou CGU) pour un site extranet avec comptes utilisateur. Tout utilisateur inscrit doit être au courant de l'utilisation de ses données (**principe de transparence**). Les objectifs d'utilisation de ses données devront aussi être affichés (**principe de limitation des finalités**).
- La vérification (**principe d'exactitude**) puis l'**anonymisation** ou la « **pseudonymisation »** des données valorisées par l'Unité et exposées sur l'**extranet**. Les processus informatiques nécessaires seront effectués sur le serveur de travail du SI interne sécurisé. Une **minimisation** des informations devra aussi être prévue (remplacer jour de naissance par mois de naissance par exemple).
- La création d'une procédure d'**archivage** puis de **destruction** des données d'un projet à l'issue d'un délai[43], qui peut être estimé à la durée du projet additionnée de 5 à 10 ans ou plus si besoin, **la durée de conservation devant avant tout être formalisée par écrit**.

Ces différents points devront être intégrés dans des Plan de gestion de données (PGD) qui seront associés à chaque cahier des charges thématiques, permettant le lien entre la composante informatique du traitement des données et le cycle de vie complet de ces dernières.

[40] **https://cil.toulouse.inrae.fr/page/1**
[41] **https://www.cnil.fr/fr/guide-rgpd-du-developpeur**
[42] **https://www.cnil.fr/fr/outil-pia-telechargez-et-installez-le-logiciel-de-la-cnil**
[43] **https://cil.toulouse.inrae.fr/page/21**

Planification sur la période du SDI



Ressources humaines

Pilotage

La présence d'un titulaire chef de projet informatique permet de mettre un place un pilotage global des travaux informatique et leur arbitrage, en relation directe avec le chef d'unité.
Ce pilotage garantit la **continuité des travaux** et la **répartition des compétences partagées chacune par au moins 2 personnels du pôle informatique**. Cette organisation sera formalisée sous la forme d'une **grille de compétences du pôle informatique** (document qualité).

La gouvernance de projets pourra être **partagée** par l'ensemble des personnels titulaires (Maitrise d'Ouvrage), via l'application de gestion de projets *RedMine*.

Equivalents temps plein (ETP)

Le pôle informatique devra être idéalement composé de **4 ETP** sur la durée de ce schéma directeur pour permettre :

- De **maintenir le SI existant** en conditions opérationnelles et veiller à sa sécurité.
- De **répondre aux demandes** réglementaires et opportunités de valorisation de données **à courts termes.**
- Et de **moderniser le SI**, objet principal de ce schéma directeur.

En conclusion

Le schéma directeur informatique 2021-2024 s'inscrit dans la lignée du projet d'Unité qui pose les objectifs du Système d'Information Intégré sur les Systèmes et Politiques Agricoles (SISPA). Il reprend en grande partie les objectifs d'intégration technologique et méthodologique développé dans ce projet.
La refonte du Système d'Information orienté services actuel reflète une volonté : amplifier les expertises et analyses de l'équipe ODR, dans le respect des orientations de l'INRAE, afin de permettre aux partenaire de mieux comprendre les changements globaux, d'accélérer la transition agro-écologique des systèmes alimentaires, d'encourager une économie basée sur la gestion circulaire et durable des ressources agricoles.

Le « mot du DU » :

Cette ambition en cache une autre, celle de la valorisation et production d'une base de connaissances exhaustive et complexe pour comprendre les systèmes agricoles et leurs évolutions, en lien en particulier avec l'impact des politiques agricoles. L'ODR a su historiquement répondre à ce besoin et doit évoluer pour continuer à le faire dans un contexte qui a évolué. Ce contexte est modelé par des nouvelles opportunités (multiplication des données mobilisables, puissance de stockage et de calcul, développement de nouveaux logiciels, etc.) mais aussi de nouvelles exigences (science ouverte, RGPD, FAIRisation des données et reproductibilité). Tout ceci conduit à ce schéma directeur ambitieux, qui propose un socle informatique solide pour développer l'ingénierie de la donnée nécessaire au SISPA et aux objectifs qu'il porte en termes de production et diffusion de connaissances. Il facilitera le travail de traitement de la donnée, à presque tous les niveaux du cycle de vie, et accroitra les possibilités de valorisation internes et externes de ces données.

Historique des versions

Indice de révision	Date	Historique des modifications
0.1	29/01/21	Première version
0.2	03/03/21	- Précisions sur la révision du SDI tous les 6 mois (p.2) - Précisions sur l'intégration de R/Shiny dans la plateforme (p. 6) - Référence au travail d'analyse d'impact relative à la protection des données déjà effectué sur la plateforme actuelle (p. 34 et 40) - Ajout d'un tableau d'historique des versions (p. 38) - Corrections diverses.
0.3	30/03/21	- Ajout d'un sommaire (p.2) - Précisions sur la présentation des activités à l'ODR (p. 5) - Ajout de la plateforme de collaboration et de documentation WikiMedia : WikiODR (p. 7) - Précisions sur l'utilisation de FileSender (p. 17) - Remplacement des termes « anonymes » par « pseudonymisés » (p. 19) - Précision sur les possibilités de créations de graphiques interactifs pour l'extranet (p. 24)

Abréviations et acronymes

AB Agriculture Biologique
ANR Agence Nationale de la Recherche
ANSSI Agence Nationale de la Sécurité des Systèmes d'Information
AOC Appellation d'Origine Contrôlée
AOP Appellation d'Origine Protégée
ASP Agence de services et paiement (organisme de paiements des aides PAC, 1er pilier et RDR)
BDNI Base de Données nationale d'identification (suivi individuel de tous les animaux de races bovines)
COTNS cotisants non-salariés de la MSA
CMS Content Management System
CNIL Commission nationale de l'informatique et des libertés
CSV Comma-separated values. Format de fichier utilisé lors de l'import de données sur Carto Dynamique.
DATAR Délégation interministérielle à l'Aménagement du Territoire et à l'Attractivité Régionale
DCE Directive Cadre Eau
DGPAAT Direction Générale des Politiques Agricole, Agroalimentaire et des Territoires du ministère de l'agriculture français.
ENVT Ecole Nationale Vétérinaire de Toulouse
ETP Equivalents temps plein
FEADER Fonds Européen Agricole pour le Développement Rural / European Agricultural Fund for Rural Development (EAFRD)
IGN Institut Géographique National
IGP Indication Géographique Protégée
INAO Institut National des Appellations d'Origine (INAO), désormais appelé Institut national de l'origine et de la qualité. Il accompagne les producteurs dans leurs démarches pour l'obtention d'un signe officiel de l'origine et de la qualité
INERIS Institut national de l'environnement industriel et des risques
INRAE Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement
MAE Mesure Agro-environnementale
MAA Ministère de l'agriculture
MEED Ministère de l'environnement, désormais dénommé Ministère de l'Ecologie, du Développement Durable, des Transports et du Logement
NAS Network Attached Storage (Serveur de stockage en réseau)
MNHN Muséum Nationale d'Histoire Naturelle. Gestionnaire de la base de données Suivi Temporel des Oiseaux Communs (STOC) et d'une base d'information sur les sites NATURA2000.
MSA Mutualité Sociale Agricole
ODR Observatoire du Développement Rural
OT-SIQO Observatoire territorial des Signes d'Identification de la Qualité et de l'Origine
PDR Programmes de Développement Rural

PGD Plan de Gestion de Données
PIA Privacy Impact Assessment
RDR Règlement de Développement Rural
RGPD Règlement Général sur la Protection des Données
RPG Registre Parcellaire Graphique
SAE2 Département d'économie de l'INRAE (Sciences Sociales, Agriculture et Alimentation, Espace et Environnement)
SDI Schéma Directeur Informatique
SGBD Systèmes de Gestion de Bases de Données
SIQO Signes d'Identification de la Qualité et de l'Origine
SISPA Système intégré d'Information sur les Systèmes et Politiques Agricoles
STG Spécialité Traditionnelle Garantie

Références

- Projet unité de service Observatoire du Développement Rural (US-ODR) pour 2019-2024
- Diagnostique premier – Février 2019 (Philippe Serard)
- Dossiers d'évaluation de l'US-ODR
- Guide de l'utilisation avancée du site de l'ODR et de l'outil Carto Dynamique
- Plan de gestion de données de l'ODR (Elise Maigné)
- Analyse d'impact relative à la protection des données (mars 2020 - Sogeti)
- Bilan de la réunion avec la Direction Général de la Sécurité Intérieure du 04/03/2020 (Cédric Gendre)
- Notice accès aux zones de dépôt partagées de l'ODR – Avril 2020
- Présentation de l'accès aux données publiques MSA via l'ODR – Mars 2018 (Elise Maigné)
- Guide de l'utilisateur de l'Observatoire territorial des SIQO – Avril 2020
- Présentation « La spatialisation et les changements d'échelles. » - 2020 (Pierre Cantelaube)