L'outil SPIDER

« Système Partagé d'Interopérabilité des Données d'Expérimentation et de Recherche »

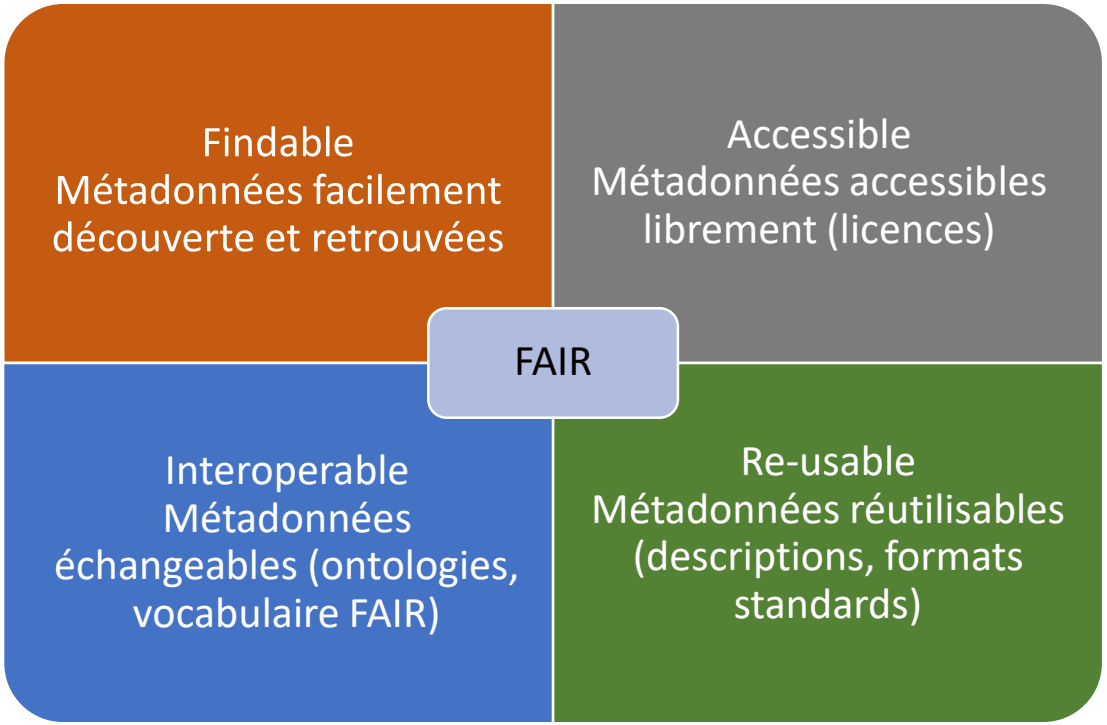
GenPhySE
Thierry Heirman & Laurence Drouilhet

Systèmes d'Informations et Calcul pour le Phénotypage Animal

Génétique Physiologie et Systèmes d'Elevage

CONTEXTE

Open data repose sur des données FAIR



https://doranum.fr/enjeux-benefices/principes-fair/

Open data repose sur des données FAIR

Importance des métadonnées

Données

Métadonnées

Garantissent la pérennité des données :

- Description des données et du contexte
- Permet leur réutilisation
- Standardisation des formats pour faciliter la recherche

CONCRETEMENT

Comment s’y prendre ?

GenPhySE dans tout ça ?

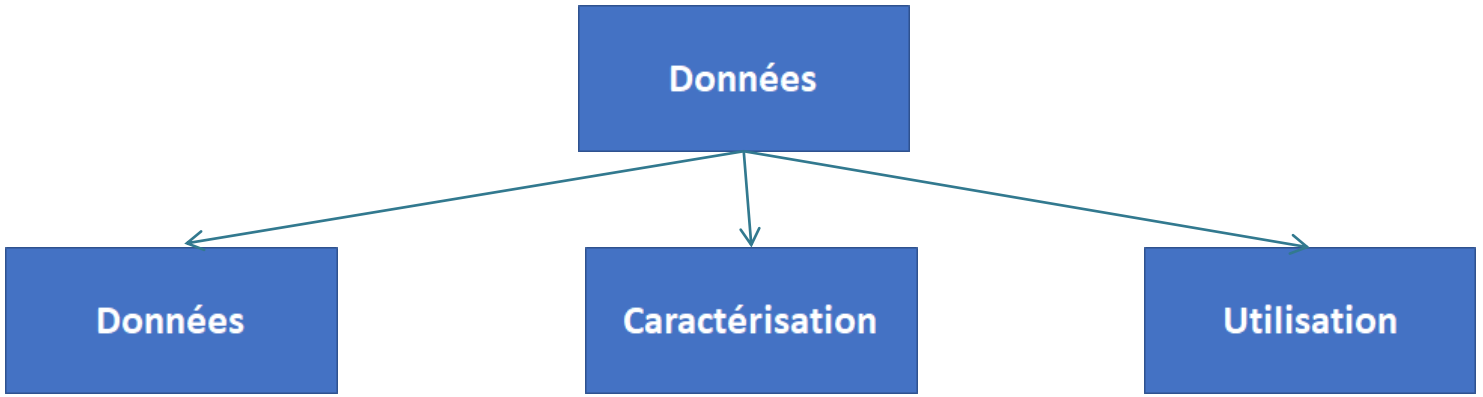
Objectifs

- Organiser le stockage des données produites dans l'unité
- Faciliter l'acquisition des métadonnées associées à nos données
- Faciliter la soumission des données
- Améliorer l'interopérabilité entre les systèmes d'informations (SI)

→ Constitution de groupes par type de données (en 2018)

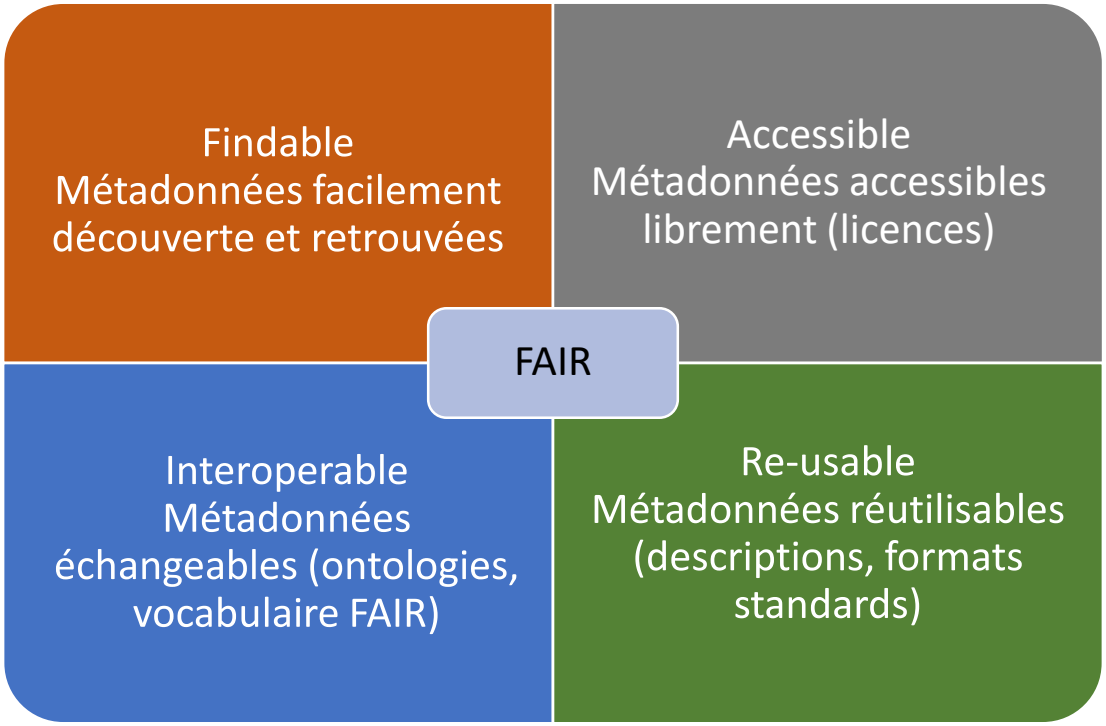
GenPhySE dans tout ça ?

Etape 1 : Inventaire Avant de savoir comment organiser les données, il faut savoir de quelles données on dispose ...



- Connaitre ce qui existe → **Findable**
- Identifier les manques (ex: « espaces de stockage ») → **Accessible**
- Identifier les « liens » permettant de raccorder les SI (ex: « Num animal ») → **Interoperable**
- Définir les souhaits d'accès → **Reusable**

FAIR


Findable
Métadonnées facilement découverte et retrouvées
Accessible
Métadonnées accessibles librement (licences)
FAIR
Interoperable
Métadonnées échangeables (ontologies, vocabulaire FAIR)
Re-usable
Métadonnées réutilisables (descriptions, formats standards)

GenPhySE dans tout ça ?

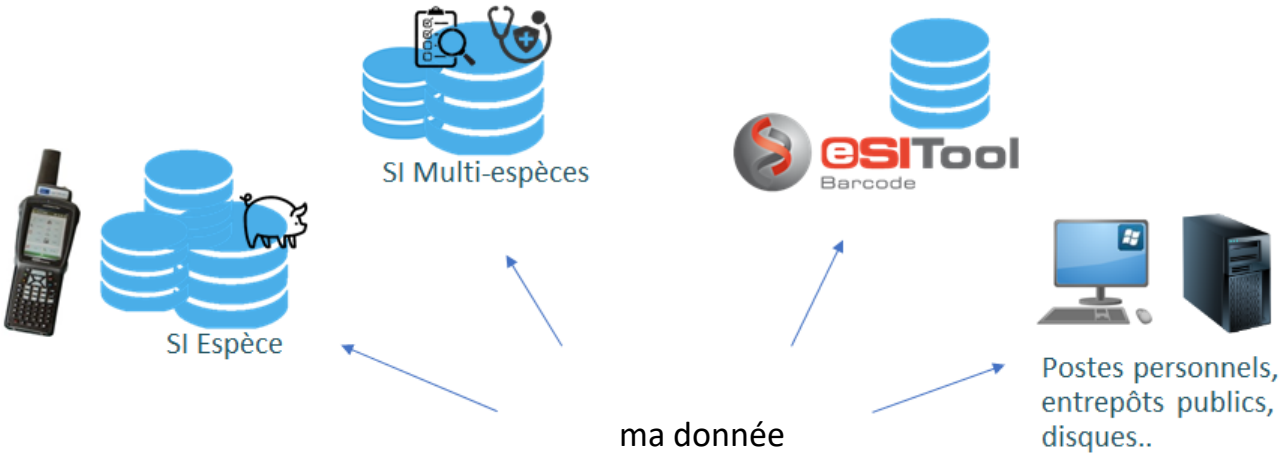
1- étape inventaire : longue et fastidieuse ! (mais c'est fait !)

GenPhySE dans tout ça ?

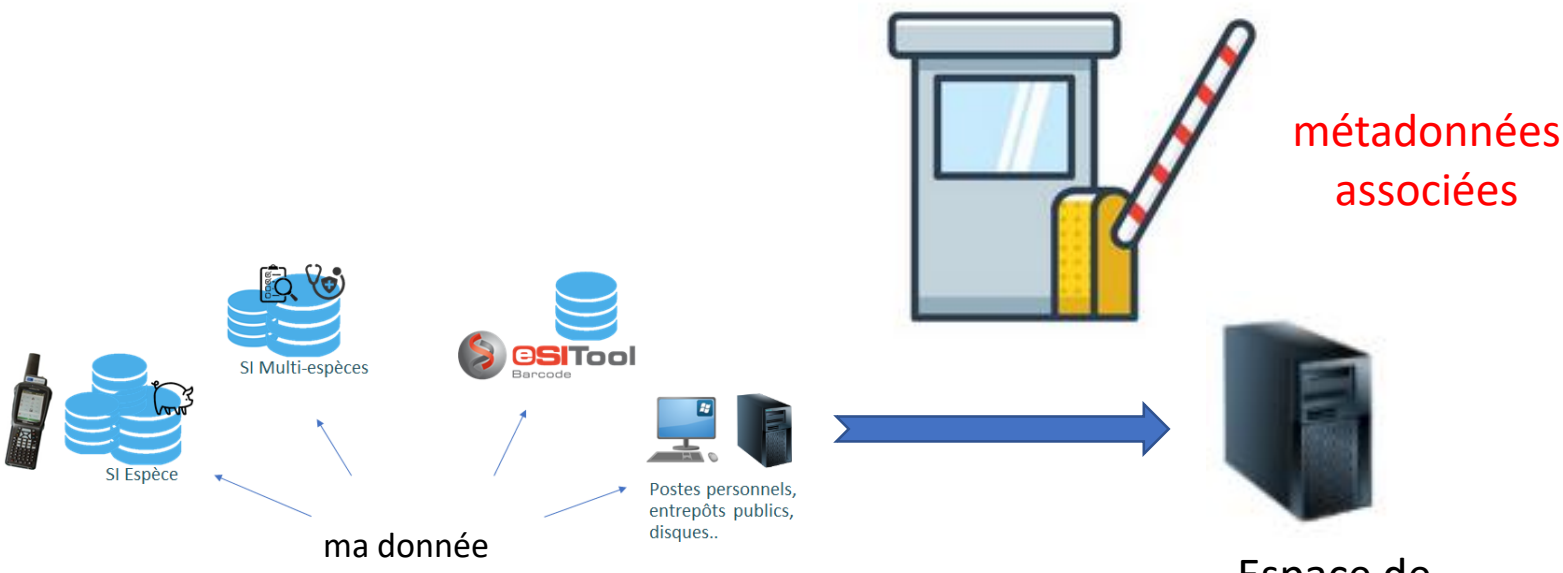
inventaire ...

Type Données -omiques	Fichier Brut	Taille Fichier Brut/échantillon	Taille fichiers bruts pour un projet	Stockage Actuel Fichier Brut	Durée Stockage souhaité Fichier Brut	Informations à stocker pour tracer ce fichier brut?	Fichier pré-analysé	Taille Fichier pré-analysé	Taille fichiers pré-analysés pour un projet	Stockage Actuel Fichier pré_analysé	Durée Stockage souhaité Fichier pré-analysé	Informations à stocker pour tracer ce fichier pré-analysé?	dépôt dans base publique?	Quel type de fichier/information déposé(e) dans base publique?
Données d'ARN circulaires	Fichier idem RNAseq						données d'annotation: Excel	négligeable				version du génome	oui	NCBI
Métagénomique 16s	Fichier fastq.gz	20Mo/échantillon	dépend nombre éch	ng6	jusqu'à publication	Réf. protocole/ phenotype/primer	Fichier .fastq/.tsv/.html	quelques centaines de Ko		ng6/genotoul bioinfo/ Galaxy/PC	jusqu'à publication	Pipeline d'analyse/script	oui (SRA)	Fichier fastq.gz + meata-data
RNAseq, Séquençage	fastq.gz	~80Mo /echant depend de la profondeur de sequencage		save/work	jusqu'à publication		Fichier .bam/bai/ mpileup/.vcf	bam: 40Mo, Bai: 2Mo et mpileup: 22Mo et vcf: 4Mo		save et work	3 ans	genome de reference		
Transcriptome (microarray)	1 fichier .txt/spot + image		66Go/64éch	serveur			fichier excel		403Mo/492fichiers/ 64échantillons	serveur			oui	log2 normalized signal
données qPCR	Données de fluorescence (à demander à la plateforme)		Fluidigm: 26500Ko/puce	serveur + plateforme			fichier excel	quelques centaines de Ko	-	serveur				
Données Métabolome RMN	fichier de Brucker: 1dossier/échantillon		399Mo/493échantillons	serveur		- préparation des éch° - spectro - TopSpin	fichier .txt des buckets transmis par la plateforme	1fichier/spectre=1Mo si bucket	493éch/493dossiers/399 Mo	serveur		Workflow4metabolomics	??	
Données Lipidomique							excel pour							
Données Protéome Shot-Gun	1 Spectre MS/MS par échantillon (à demander à la plateforme?)			Plateforme protéomique INRA			fichier .csv/ .html (séquences des peptides et identification proteines) + rapport doc et		14Mo/70échantillons	Cloud ENSAT (via logiciel SeaFile)			non	
Données Protéome Electrophorèse 2D/analyse d'image	image du gel (fichier .mel)	12Mo/gel (1gel=1échantillon)	dépend nombre éch	DD PC analyse d'image ENSAT + sauvegarde sur DD externe			fichier .samespotsexperiment = Fichier généré par le logiciel d'analyse de gels 2D SameSpots (alignement des		70Mo/300ech 8Mo/12échantillons	DD PC analyse d'image ENSAT + sauvegarde sur DD externe		scan du gel avec identification des spots	non	

GenPhySE dans tout ça ?



GenPhySE dans tout ça ?



GenPhySE dans tout ça ?

1- étape inventaire : longue et fastidieuse ! (mais c'est fait !)

2- métadonnées associées aux données : en cours

GenPhySE dans tout ça ?

1- étape inventaire : longue et fastidieuse ! (mais c'est fait !)

2- métadonnées associées aux données : en cours

3- collaboration CATI SICPA pour création stockage et interface SPIDER

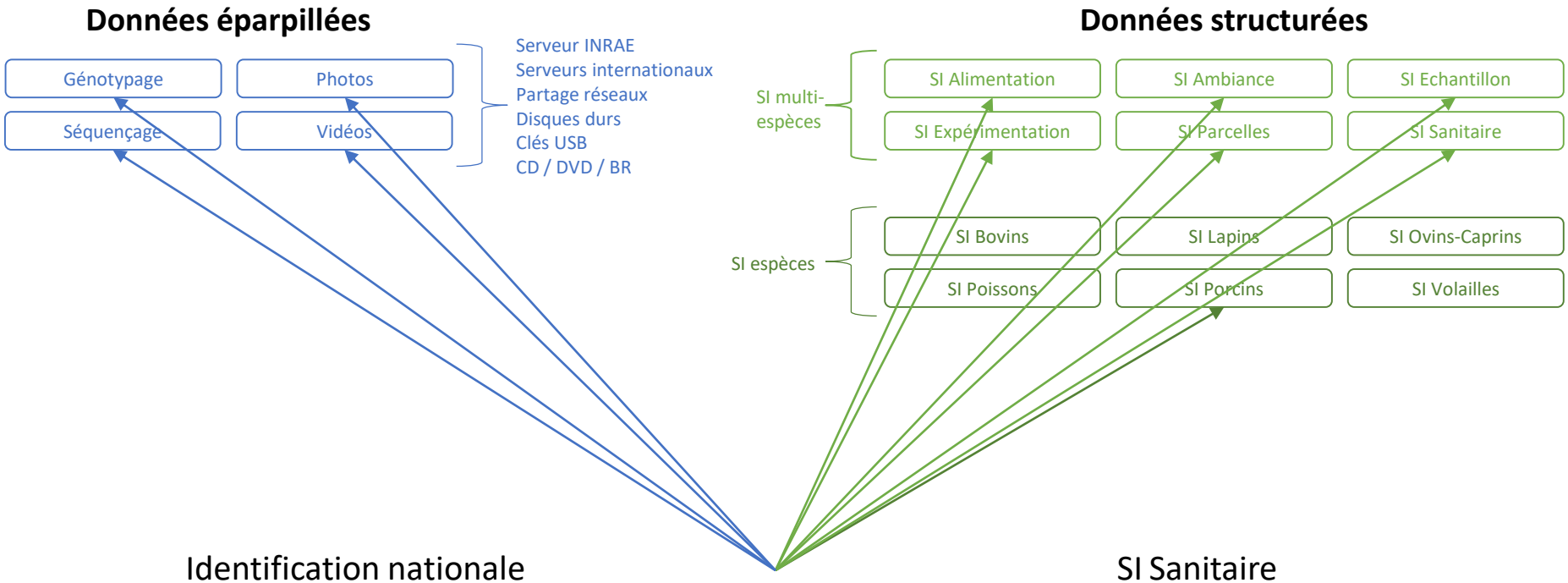
SPIDER

(**S**ystème **P**artagé d'**I**nteropérabilité des **D**onnées d'**E**xpérimentation et de **R**echerche)

Objectif : collecter et interopérer des informations issues de sources hétérogènes afin qu'elles soient centralisées pour faciliter l'accès et la valorisation des données

Cas d'usage exprimés par le groupe de travail « Gestion des Données » :

- CU 1a : à partir d'un numéro d'animal, savoir s'il existe des données …
- CU 1b : … et savoir où elles se trouvent
- CU 2 : récupération des données brutes et/ou pré-traitées



Constats de départ

Données éparpillées

Génotypage | Photos | Séquençage | Vidéos

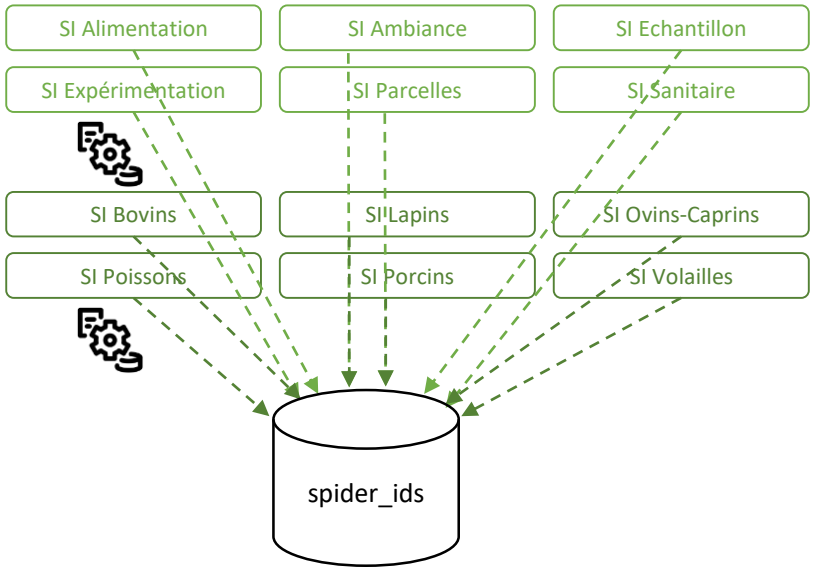
id_animal	margaup_dnais	margaup_nation	margaup_inra	sanit_ani_id	sanit_ani_num	sidex_ani_elev	sidex_ani_ordre	sidex_date	date_reforme
44191	2020-12-26	FR372Q4202041391	041391	NOU;FR372Q4202041391	041391	FR372Q4	041391	2020-01-01	2021-05-19
44190	2020-12-23	FR372Q4202041336	041336	NOU;FR372Q4202041336	041336	FR372Q4	041336	2020-01-01	2021-05-19
44189	2020-12-23	FR372Q4202041329	041329	NOU;FR372Q4202041329	041329	FR372Q4	041329	2020-01-01	2021-05-19
44188	2020-12-23	FR372Q4202041324	041324	NOU;FR372Q4202041324	041324	FR372Q4	041324	2020-01-01	2021-05-19
44187	2020-12-23	FR372Q4202041314	041314	NOU;FR372Q4202041314	041314	FR372Q4	041314	2020-01-01	2021-05-19

id_animal	margaup_dnais	margaup_nation	margaup_inra	sanit_ani_id	sanit_ani_num	sidex_ani_elev	sidex_ani_ordre	sidex_date	date_reforme
41777	2015-09-10	FR17MAG201514313	514313	NULL	NULL	FR17MAG	514313	2015-01-01	2015-10-19
35803	2014-05-19	FR17MAG201412688	412688	NULL	NULL	FR17MAG	412688	2014-01-01	2014-11-17
35802	2014-05-19	FR17MAG201412687	412687	NULL	NULL	FR17MAG	412687	2014-01-01	2014-11-17
35796	2014-05-18	FR17MAG201412680	412680	NULL	NULL	FR17MAG	412680	2014-01-01	2014-11-17
35790	2014-05-16	FR17MAG201412673	412673	NULL	NULL	FR17MAG	412673	2014-01-01	2014-11-17

id_animal	margaup_dnais	margaup_nation	margaup_inra	sanit_ani_id	sanit_ani_num	sidex_ani_elev	sidex_ani_ordre	sidex_date	date_reforme
372548	2021-10-13	FR17MAG202115003	115003	GJS;FR17MAG202115003	115003	NULL	NULL	NULL	2021-12-08
371060	2021-07-22	FR17MAG202113511	113511	GJS;FR17MAG202113511	113511	NULL	NULL	NULL	2021-12-08
370955	2021-07-21	FR17MAG202113406	113406	GJS;FR17MAG202113406	113406	NULL	NULL	NULL	2021-12-08
378603	2021-11-12	FR352WH202163898	163898	SGC;FR352WH202163898	163898	NULL	NULL	NULL	2021-12-07
378602	2021-11-12	FR352WH202163897	163897	SGC;FR352WH202163897	163897	NULL	NULL	NULL	2021-12-07

id_animal	margaup_dnais	margaup_nation	margaup_inra	sanit_ani_id	sanit_ani_num	sidex_ani_elev	sidex_ani_ordre	sidex_date	date_reforme
378752	2021-12-03	FR352WH202164047	164047	NULL	NULL	NULL	NULL	NULL	2021-12-05
378732	2021-12-03	FR352WH202164027	164027	NULL	NULL	NULL	NULL	NULL	2021-12-05
378697	2021-12-03	FR352WH202163992	163992	NULL	NULL	NULL	NULL	NULL	2021-12-04
379578	2021-12-03	FR352WH202167856	167856	NULL	NULL	NULL	NULL	NULL	2021-12-03
379577	2021-12-03	FR352WH202167855	167855	NULL	NULL	NULL	NULL	NULL	2021-12-03

Données structurées



CIBLE : Synchronisation des identifiants

Données éparpillées

Génotypage | Photos

Séquençage | Vidéos

Récupération du fichier

Serveur de fichiers volumineux (200 To)

Dépôt du fichier

Int
de
de

```
margaup:
    date_nais:      "2021-04-16Z"
    n_inra:         "180392"
    n_nation:       "FR86DGH202180392"
    sexe:           "1"
sanit:
  0:
      ani_id:       "GIB;FR86DGH202180392"
      ani_num:      "180392"
      ev_datedeb:   "2021-04-20Z"
      ev_datefin:   "2021-04-20Z"
      nom_ev:       "Carences"
      trait_date:   "2021-04-20Z"
      trait_nom:    "Forceris 250"
      trait_qte:    "2.0 ml"
```

trement
adonnées

Données structurées

SI Alimentation | SI Ambiance | SI Echantillon

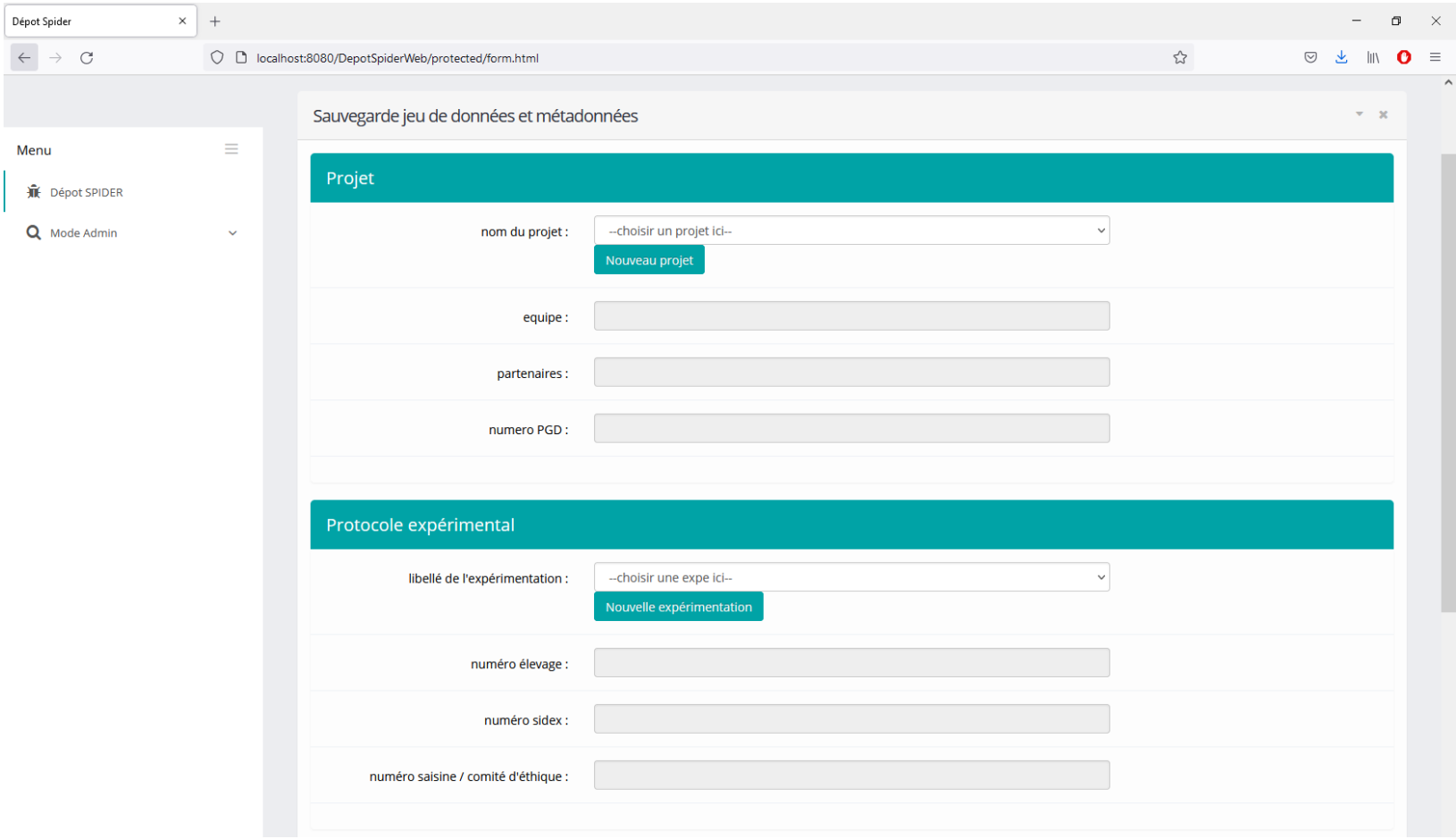
SI Expérimentation | SI Parcelles | SI Sanitaire

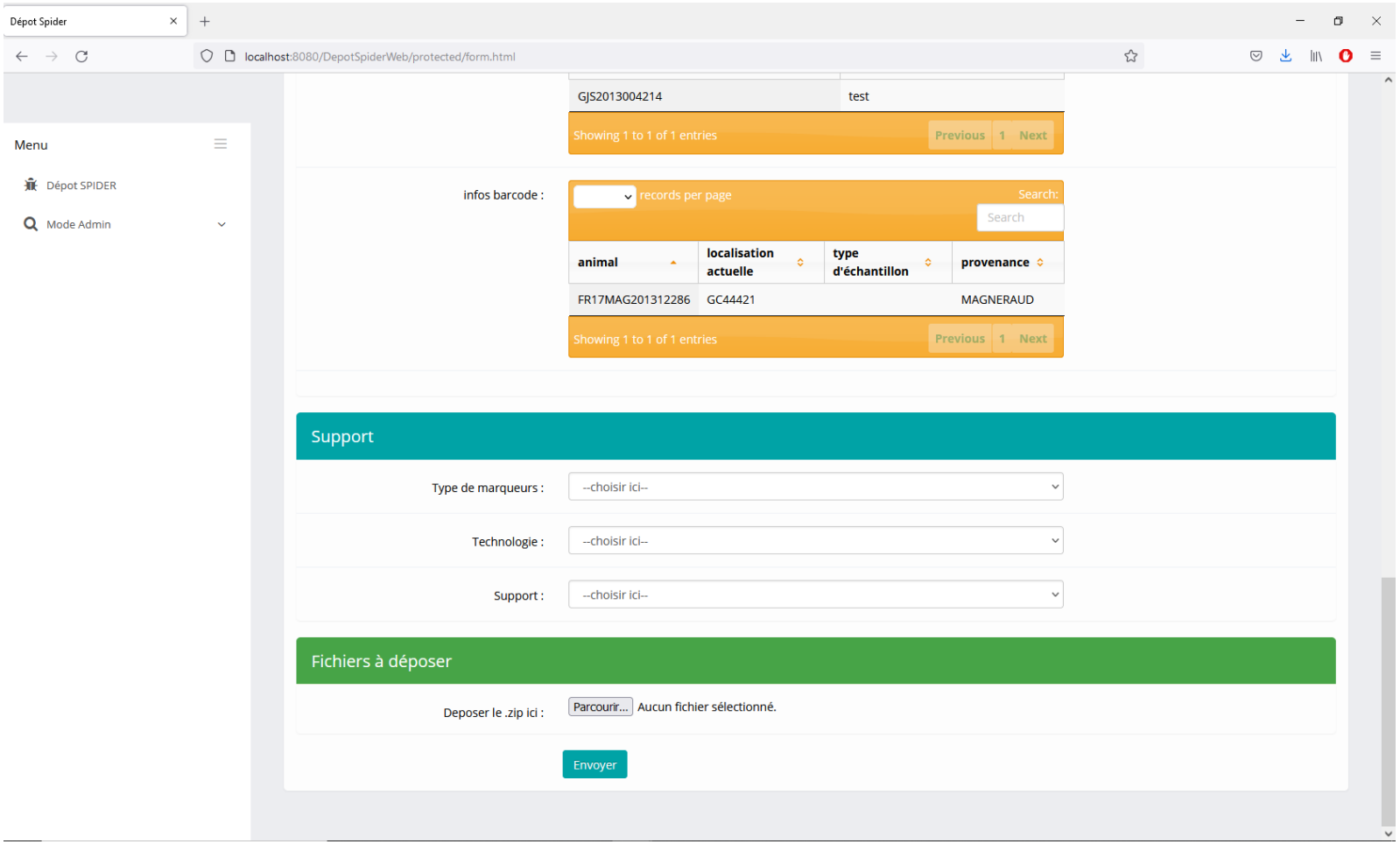
SI Bovins | SI Lapins | SI Ovins-Caprins

SI Poissons | SI Porcins | SI Volailles

Cluster MongoDB

CIBLE : Récupération des données


Dépot Spider
localhost:8080/DepotSpiderWeb/protected/form.html
Menu
Dépot SPIDER
Mode Admin
Sauvegarde jeu de données et métadonnées
Projet
nom du projet :
--choisir un projet ici--
Nouveau projet
equipe :
partenaires :
numero PGD :
Protocole expérimental
libellé de l'expérimentation :
--choisir une expe ici--
Nouvelle expérimentation
numéro élevage :
numéro sidex :
numéro saisine / comité d'éthique :


Dépot Spider
localhost:8080/DepotSpiderWeb/protected/form.html
Menu
Dépot SPIDER
Mode Admin
GJS2013004214
test
Showing 1 to 1 of 1 entries
Previous
1
Next
infos barcode :
records per page
Search:
Search
animal
localisation actuelle
type d'échantillon
provenance
FR17MAG201312286
GC44421
MAGNERAUD
Showing 1 to 1 of 1 entries
Previous
1
Next
Support
Type de marqueurs :
--choisir ici--
Technologie :
--choisir ici--
Support :
--choisir ici--
Fichiers à déposer
Deposer le .zip ici :
Parcourir...
Aucun fichier sélectionné.
Envoyer

Données éparpillées

Génotypage | Photos

Séquençage | Vidéos

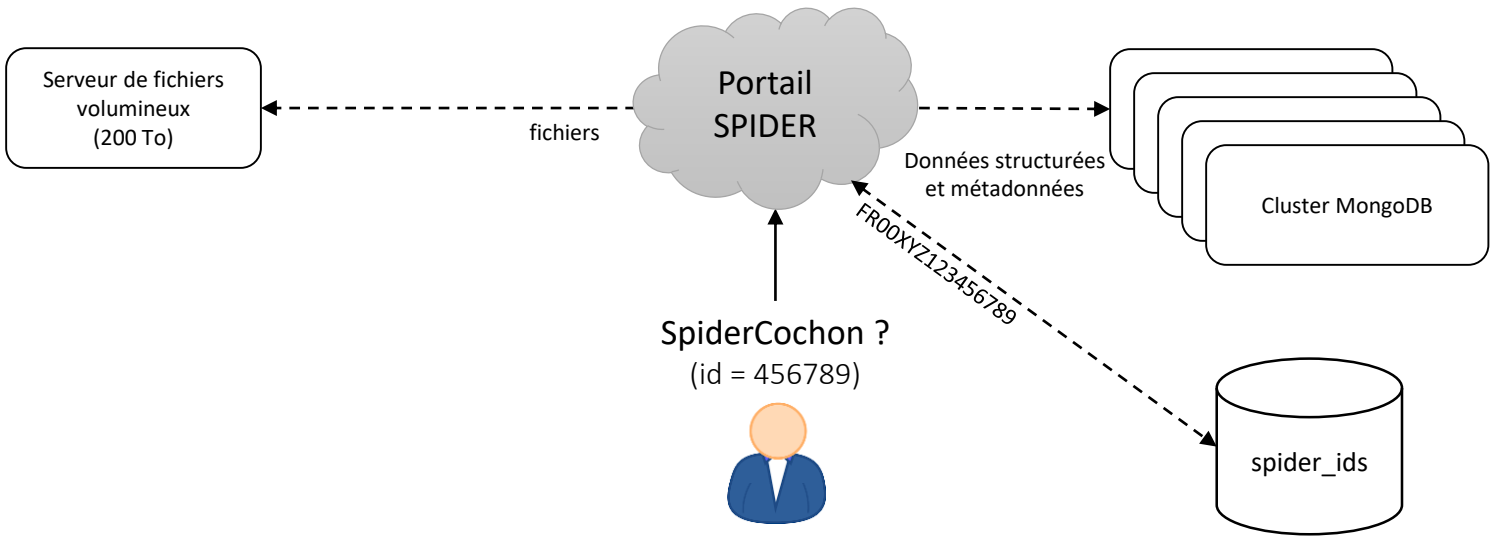
Données structurées

SI Alimentation | SI Ambiance | SI Echantillon

SI Expérimentation | SI Parcelles | SI Sanitaire

SI Bovins | SI Lapins | SI Ovins-Caprins

SI Poissons | SI Porcins | SI Volailles



CIBLE : Interrogation des données

Suivi du projet : Ce qui a déjà été fait

Objectif 1 : Réaliser une preuve de concept du projet SPIDER *(Stage Marie Laur, 5e année INSA)*

- Etudier les différentes technologies rendant possible le projet
- Importer un jeu de test de données
- Réaliser la maquette d'interrogation des données *(par identifiant animal)* et la maquette de dépôts des données *(avec saisie des métadonnées)*
- Valider indépendamment chaque élément du projet

Objectif 2 : Mettre en place l'infrastructure du projet

- Etudier les différentes offres institutionnelles en lien avec la DSI *(stockage, virtualisation, calcul)*
- Choisir les technologies de virtualisation, des stockages capacitifs et distribués
- Mettre en place les différentes briques logicielles
- S'assurer du bon fonctionnement des briques logicielles et de la bonne communication entre briques logicielles

Objectif 3 : Fournir une première version testable au comité des utilisateurs *(CDD Marie Laur)*

- Peupler le stockage distribué avec un jeu de données test (*porcins de GenPhySE)* issues de SI structurés
- Ecrire les interfaces web d'interrogation des données par identifiant animal et de dépôts des fichiers volumineux avec saisie des métadonnées
- Livrer une première version testable du projet
- Traiter les retours utilisateurs, corriger les anomalies techniques et métier
- Ecrire les documentations techniques et utilisateur du projet

Suivi du projet : à venir

Objectif 4 : Passer SPIDER à l'échelle pour répondre à tous les besoins de GenPhySE *(Data Engineer, 8-12 mois)*

- Réaliser une console web d'administration *(Paramétrage des contenus des interfaces web, suivi/curation des données, …)*
- Poursuivre le peuplement du stockage distribué avec les données des autres SI structurés pour les animaux de GenPhySE
- Ajouter d'autres types de données *(autres données -omiques, photos, vidéos, autres, …)*
- Ajouter des fonctionnalités de recherche avancées *(par projet, par lots, par échantillon, …)*

Objectif 5 : Améliorer l'interopérabilité de la solution *(Data Engineer, 4-6 mois)*

- Interopérabilité technique : intégration grâce au projet SICPA_Interop et au recours à des technologies standards
- Interopérabilité syntaxique : utilisation de formats ouverts *(JSON, CSV, TXT, …)*
- Interopérabilité sémantique : mettre en place le lien avec les ontologies *(AOL, ATOL, EOL, …)*

Objectif 6 : Décrire les nouveaux/futurs types de données de manière générique *(Data Engineer, 8-12 mois)*

- Décrire les futurs types de données *(données et métadonnées)*
- Concevoir une méthode générique pour l'ajout de futurs types de données
- Réaliser l'interface « ajout de nouveaux types de données » dans la console web d'administration

Suivi du projet : à venir

Objectif 7 : Tendre de plus en plus vers des données fédérées *(Data Engineer, 4-6 mois)*

- Ajouter les fonctionnalités d'export de datasets vers les dataverses INRAE
- S'appuyer sur les développements SICPA_OpenData

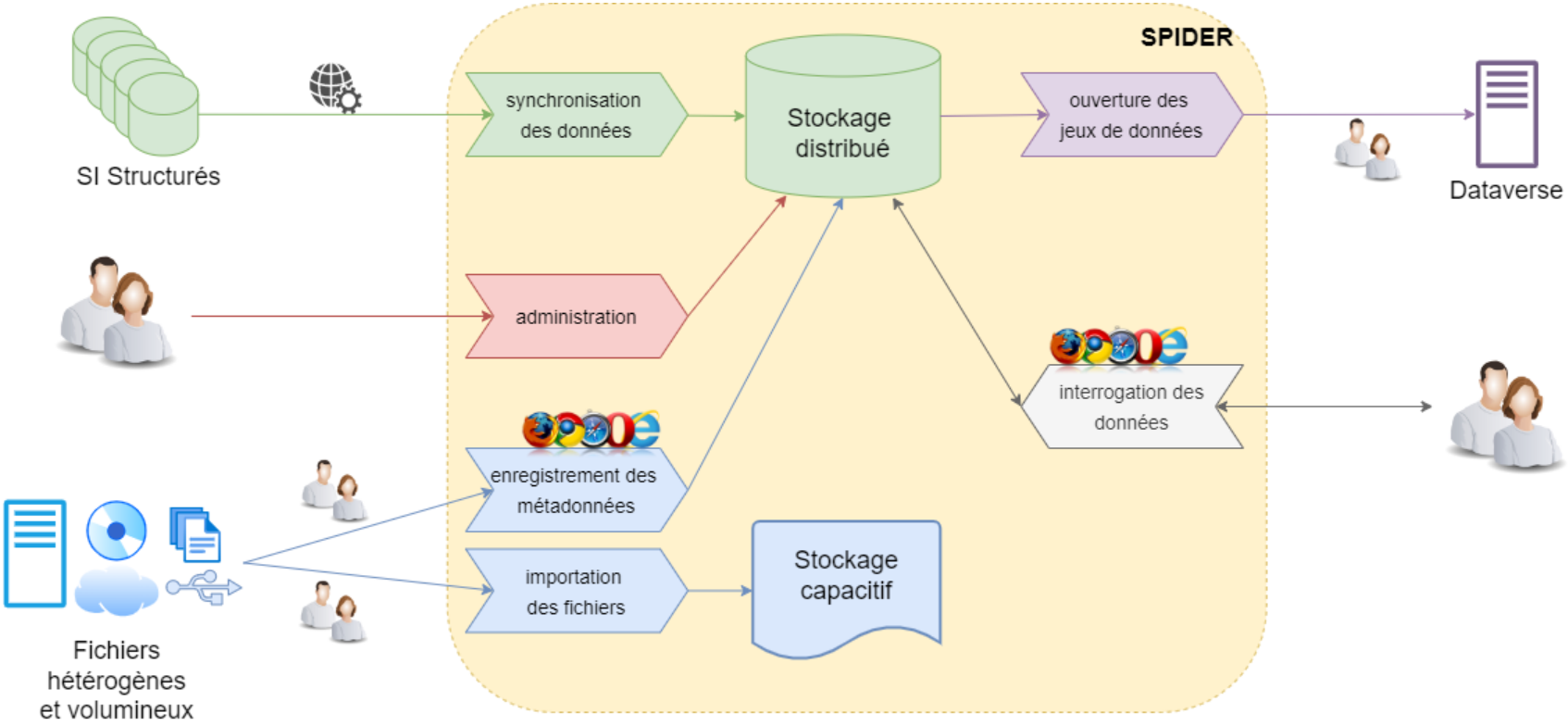
Objectif 8 : Ouvrir la solution SPIDER aux départements Génétique Animale et PHASE *(Data Engineer, 6-8 mois)*

- Etendre la solution à GABI, à PEGASE, à UMRH, …
- Peupler le stockage distribué avec toutes les données de tous les SI du CATI SICPA

Objectif 9 : Industrialiser la solution SPIDER *(Data Engineer, 6-8 mois)*

- Utiliser des technologies d'infrastructure as code *(Ansible ? Docker ? Terraform ?)* pour industrialiser la création du cluster distribué, la création du serveur web, le déploiement des interfaces web, …
- Ouvrir la solution au plus grand nombre

Conclusion : vision fonctionnelle de l'architecture



Ouverture

Objectif 7 : Tendre de plus en plus vers des données fédérées

SICPA_OpenData :

- Objectif : faciliter, de manière transparente, la publication de données issues de nos SI vers le dataverse INRAE *(aide à la création du fichier de métadonnées du dataset, aide à la création distante du dataset, aide à l'upload des fichiers, aide à la publication du dataset)*
- Toutes les bibliothèques s'appuient sur les API natives Dataverse
- Bibliothèques déjà développées :
  - CSharp / .NET *(DLL)*
  - Java *(JAR)*
  - Un PHP *(PHAR)*
- Bibliothèque à venir :
  - Python *(PYC)*
- Bibliothèques disponibles sur la forge DGA, la forge MIA et sur le wiki du PEPI 2G *(lien vers forge la forge MIA)*
- Documentations techniques disponibles sur le serveur germinal de Toulouse

Merci de votre attention

Des questions?

Ouverture

CSharp / .NET :

- Projet : https://forgemia.inra.fr/theirman/sicpaopendata-for-dotnet
- Documentation d'intégration : https://forgemia.inra.fr/theirman/sicpaopendata-for-dotnet/-/blob/master/README.md
- Documentation technique : https://germinal.toulouse.inra.fr/~theirman/SicpaOpenData/DotNET/index.html
- Librairie : https://forgemia.inra.fr/theirman/sicpaopendata-for-dotnet/-/blob/master/SicpaOpenData/bin/Release/SicpaOpenData.dll

Java / JDK :

- Projet : https://forgemia.inra.fr/theirman/sicpaopendata-for-jdk
- Documentation d'intégration : https://forgemia.inra.fr/theirman/sicpaopendata-for-jdk/-/blob/master/README.md
- Documentation technique : https://germinal.toulouse.inra.fr/~theirman/SicpaOpenData/JDK/
- Librairie : https://forgemia.inra.fr/theirman/sicpaopendata-for-jdk/-/blob/master/target/SicpaOpenData.jar

PHP :

- Projet : https://forgemia.inra.fr/theirman/sicpaopendata-for-php
- Documentation d'intégration : https://forgemia.inra.fr/theirman/sicpaopendata-for-php/-/blob/master/README.md
- Documentation technique : https://germinal.toulouse.inra.fr/~theirman/SicpaOpenData/PHP/
- Librairie : https://forgemia.inra.fr/theirman/sicpaopendata-for-php/-/blob/master/target/SicpaOpenData.phar