



# SISelune : où en est-on ?



Alban THOMAS [alban.thomas@agrocampus-ouest.fr](mailto:alban.thomas@agrocampus-ouest.fr)

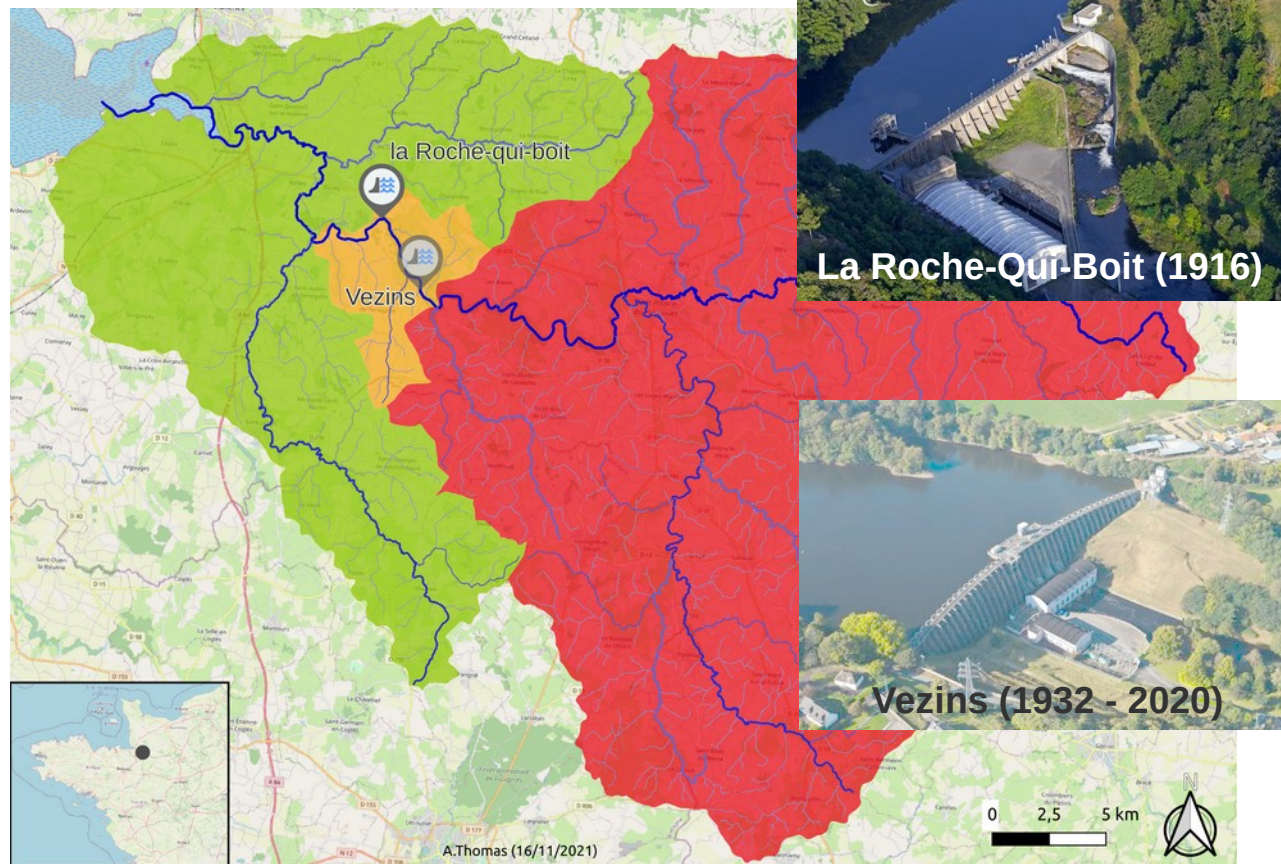
Laura SOISSONS

Jean-Marc ROUSSEL



# Le programme scientifique Sélune

- Un cas d'étude unique en Europe
- Un programme
  - Long-terme (2012 → 2027)
  - Pluridisciplinaire
  - Multi-institutionnel→ vise à servir de référence (« boîte à outils »)



# Acteurs / Publics



© Ophélie Fovet



## • Programme scientifique

- Scientifiques (!)
- Produisent des données
- Doivent les diffuser



## • Chantier

- Maîtres d'ouvrage (EDF, DDTM)
- Problématique = sécurité
- Coordination avec scientifiques
- + données (contextuelles)



Décembre 2019

Observatoire Photographique du Paysage de la Sélune (SMBS & Univ. Paris Nanterre)



© Marie-Anne Germaine

## • Publics

- Scientifiques
- Autres :
  - Curieux
  - Acteurs du chantier
  - Locaux (élus et habitants)

# SISelune

- Système d'information (SI) du programme Sélune

- 3 Objectifs :

**1) Mettre à disposition**

**2) Centraliser et sécuriser**

**3) Aider à appréhender  
l'information**

- Début du projet : 02/2020

→ Retour d'expérience (à une étape charnière)

- Interopérable (IDS)

- Centraliser les données

## 1. Diffuser (mettre à disposition)

- Des scientifiques  
(échanges de données)

- de tous

- Faciliter collaboration  
entre scientifiques

- Simplifier

- Visualiser/Cartographier

- Pérenniser (>= 2027)  
données + services

## 2. Sécuriser

- Sécuriser  
(sauvegarde + sécurité)

- Bancariser  
(données + métadonnées)

## 3. Aider (à appréhender)

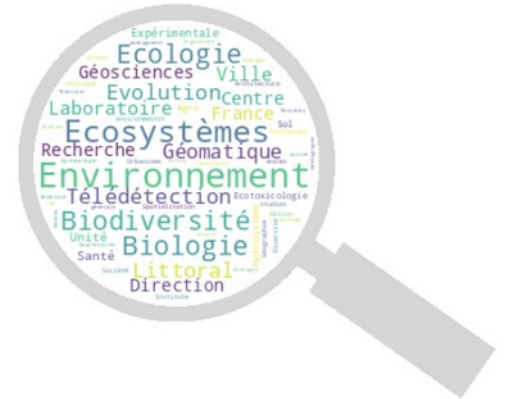
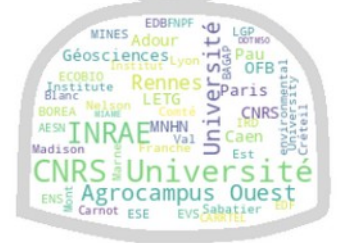
- Mettre en évidence les changements

- Multi-lingue (anglais/français)

## Objectifs du SI

# SISelune : les « challenges »

- Collaboration inter-institutionnelle (~15 unités de recherche, DDTM, EDF, OFB..., AESN)
- Données hétérogènes
  - Thématiques : sciences de la vie, physiques et humaines
  - Nature : tabulaire, raster, vecteur, images, vidéos
- Données volumineuses
- Diffusion à plusieurs niveaux (publiques, sensibles, sous embargo)
- Pérennité (au moins jusque 2027)
- (Coût de maintenance faible)



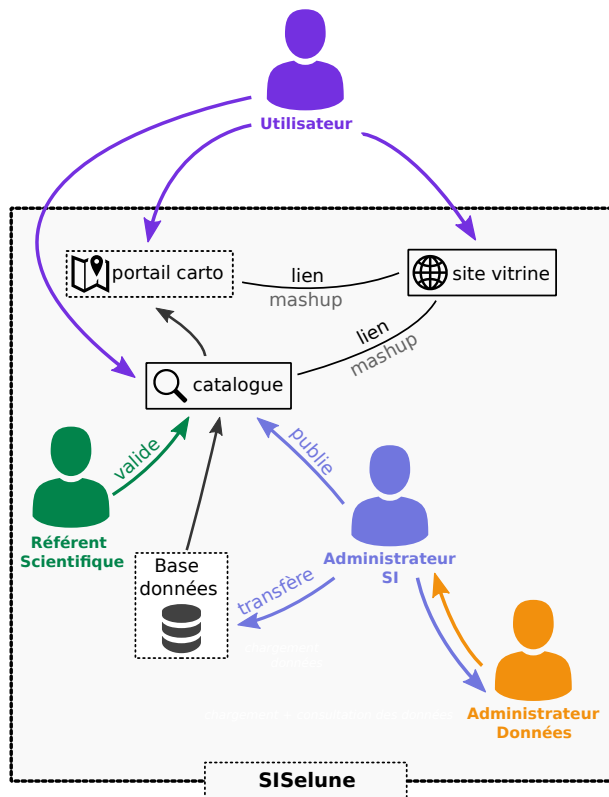
# Rôles dans SISelune

- **Acteurs de SISelune :**
  - Référents scientifiques → a le mot final
  - Administrateurs de données → gère et connaît son jeu de données
  - Administrateur SI → (je suis ici)
    - administre l'application (gestion de comptes, diagnostic défaillances),
    - aide à l'utilisation du SI, ...
  - **Administrateur Système** (ne pas l'oublier)
- **Utilisateurs** : acteur, avec un compte. Accès à toutes les données
- **Visiteurs** : consulte librement le SI

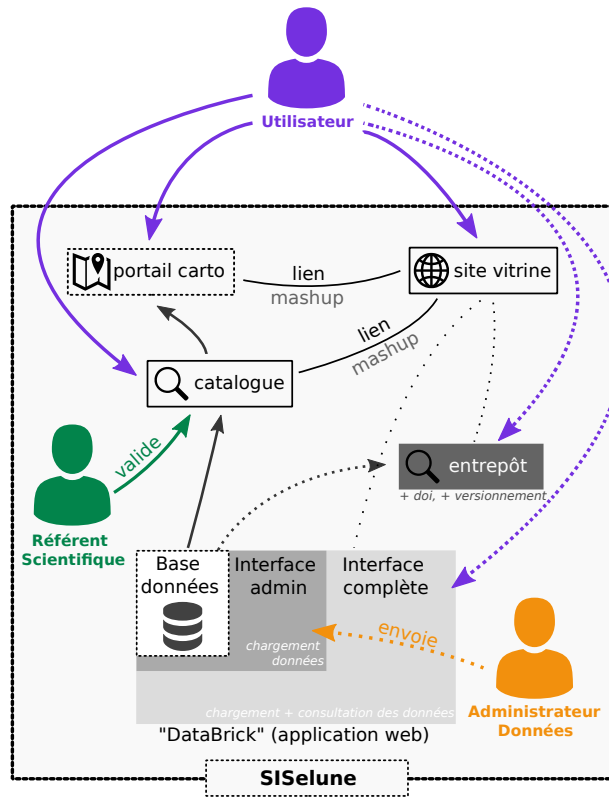


- **Logiques métiers différentes**
- **Pratiques différentes autour données**

# Architecture de SISelune



Actuellement



A terme

Je me charge actuellement l'intégration... en attendant l'application

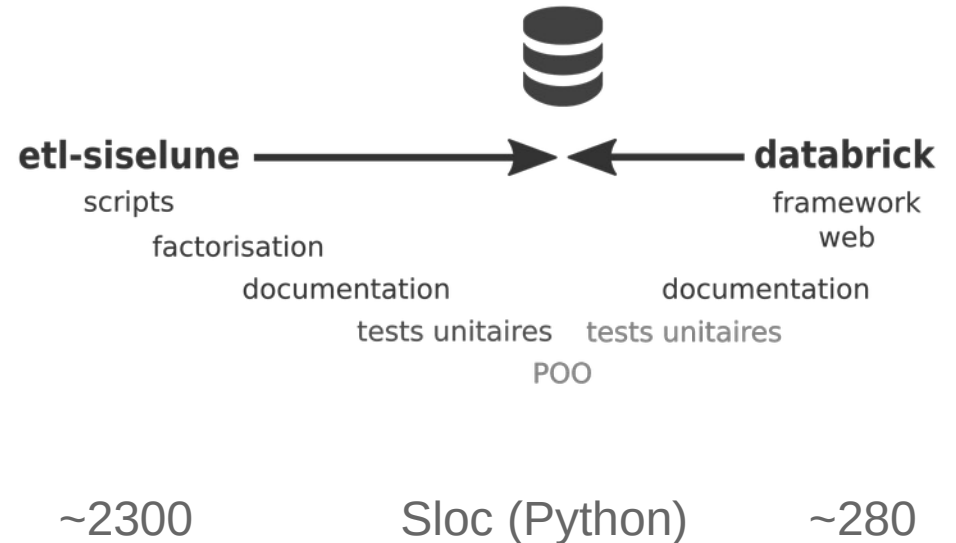
## Légende

opérationnel			(incomplet)
testé			
envisagé			(plus tard)

# Développement



- 2 projets
  - 1) « etl-siselune » : bac à sable pour Extraire Transformer et Charger (ETL) les données dans le SI
  - 2) « databrick » : application web pour charger et les données dans le SI (future prod)
- Bonnes pratiques : versionnement, style (PEP8), documentation (sphinx), tests unitaires





# Intégration : la démarche

- (Intégration : « etl-siselune »)
- **Reproductible** :
  - basée sur du code
  - (sauf exception) sans modifier les données source
- **Ouverte** → le plus « FAIR » possible
  - diffusion publique par défaut (+ métadonnées...)
  - Respect de standards
  - Lien avec bases de données de référence : SANDRE, TAXREF, BD TOPAGE...
- **Echanges** avec administrateurs de données et référents scientifiques
- **Vérifications** : structure, valeurs et géométrie

# Intégration : Notebooks

The screenshot shows a Jupyter Notebook titled 'bioc\_MIB\_API'. The left sidebar contains a table of contents with sections like '1.1 Contexte', '2.2 Fonctions', '3.1 Changement', '3.2 Vérifications', and '12. TOPO'. The main area displays code for '3.2 Vérifications' with the following content:

```
Entrée [14]: 1 # Vérifications géographiques
            2 secteur = check_geom(secteur)
            WARNING:Coche non projetée en Lambert 93 -> reprojection
            INFO:Coche incluse dans site d'étude -> OK
            INFO:Aucune géométrie nulle -> OK
            INFO:Pas de géométrie en double -> OK

Entrée [15]: 1 # Vérifications basées sur le dictionnaire de données
            2 check_data(secteur, notnull_cols['secteur'].loc[~pd.isnull(cols['secteur']).obligatoire], 'colonne'],
            unique_cols['secteur'].loc[~pd.isnull(cols['secteur']).unique()], 'colonne'])
            INFO:Pas de colonne en double -> OK
            INFO:Non nulle -> OK
            INFO:Pas de double dans les colonnes -> OK

Out[15]: True

Entrée [16]: 1 describe_data(secteur)
            INFO:Pas de colonne en double -> OK

Out[16]:
```

	cols	types	NA	exemples
id_secteur	id_secteur	int64	0	[1, 2, 3]
code	code	object	0	[S1, S2, S3]
nom	nom	object	0	[Saint-Hilaire-du-Hautecorail, Point de la République]
sec_coorde_lambert_x	sec_coorde_lambert_x	float64	0	[346773.36, 342055.5, 337412.06]
sec_coorde_lambert_y	sec_coorde_lambert_y	float64	0	[2403069.2, 2402239.2, 2403405.0]
lon	lon	float64	0	[1.1665284, -1.146188, -1.22191]
lat	lat	float64	0	[48.362018, 48.367506, 48.37506]
altitude	altitude	int64	0	[82, 16, 33]
caract	caract	object	0	[Point d'équipement Vesteur, Amont proche...]
sec_distance_max	sec_distance_max	float64	5	[0, 1]
sec_date_creation	sec_date_creation	object	6	
sec_date_mise	sec_date_mise	object	6	

```
Entrée [17]: 1 map_points(secteur, labels='code', offset=5000, title='Secteurs MIB')

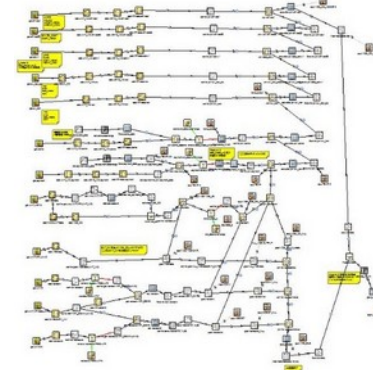
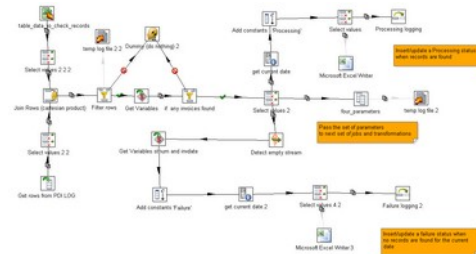
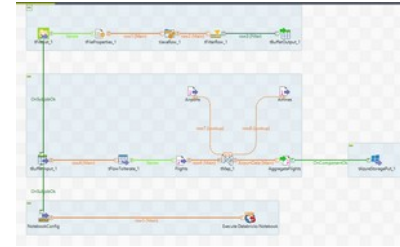
Secteurs MIB
```

The map shows a geographical area with a grid of coordinates. Red and blue markers are placed on the map, corresponding to the data in the table above. The title of the map is 'Secteurs MIB'.

- Code + sortie du code + texte (+ images)
- Support d'échange si :
  - Code reste « discret »
  - Structure claire
- Intérêts :
  - Vérifications (dont comparaison avec précédents résultats)
  - Tester et soumettre des visualisations
  - Export d'un « rapport d'intégration »
- Limites :
  - Contenu (texte et code) doit rester concis et cohérent (à jour)
  - Pas de solution type « template »

# Intégration : coder « à la main » ?

- Logiciels d'ETL existent (Talend open Studio, FME, Lumada/Pentaho)
  - NB : temps d'apprentissage à ne pas sous-estimer
- Un développement (ici sous Python) :
  - Offre plus de liberté/possibilités
  - Peut resservir (partiellement) à la future application Web
- Point faible (?) : pas de graphique du « Pipeline » d'intégration
  - 1 support qui pourrait aider les échanges avec les acteurs ? (oui, dans les cas simples)



# Qualité : dictionnaire de données

nom	description
colonne	Nom de la colonne dans le SI
source	Nom de la colonne dans le jeu de données source
table	Nom de la table dans le SI
publique	Est-ce que la colonne doit être visible (dans le SI) ?
obligatoire	Est-ce que toutes les cellules doivent être renseignées ?
unique	Est-ce que les valeurs doivent être uniques ?
dtype	Type de données dans la colonne
description	Description de la colonne
Commentaires	Lieu d'échange entre les acteurs

- ~ « structure attendue »
- Utilisé :
  - Lors des échanges avec acteurs
  - Pour la programmation (configuration)
- **1 compromis** :
  - peut être complété (liste des valeurs...)
  - ne pourra pas être complet (proposer tout ce que permet une base de données)
- *Remarques sur l'analyse de qualité*
  - *développement pas prévu (au moins pas autant)*
  - *solutions existeraient ([validata](#), [frictionless](#) et pour les données géographiques [quadogeo](#))*
    - à tester/suivre

# Base de données



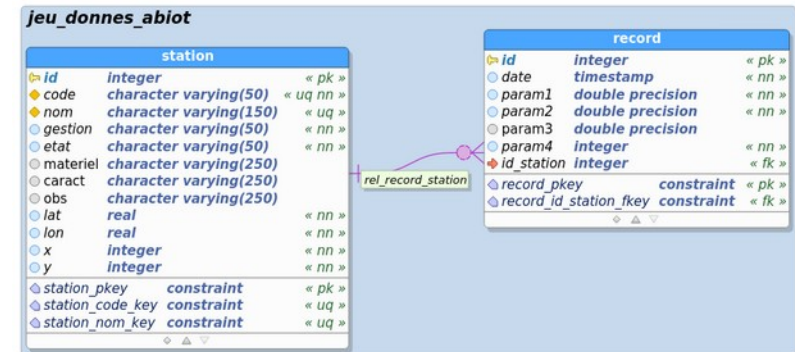
- Données **centralisées** mais **indépendantes**
- Schémas (Gestion des droits)
  - **1 schéma = 1 jeu de données** (isolement)
  - Public : **diffusion** (vues)
  - app : logique applicative

# Base de données : Un lac de données ?

- Un peu :
  - données indépendantes
  - Forme normalisée souvent < 3
- Pas vraiment :
  - Besoin des propriétés ACID (pérennité)
  - ETL et non ELT
  - volonté d'homogénéiser la structure des jeux de données
  - optimisation pour requêtes SQL (index...)

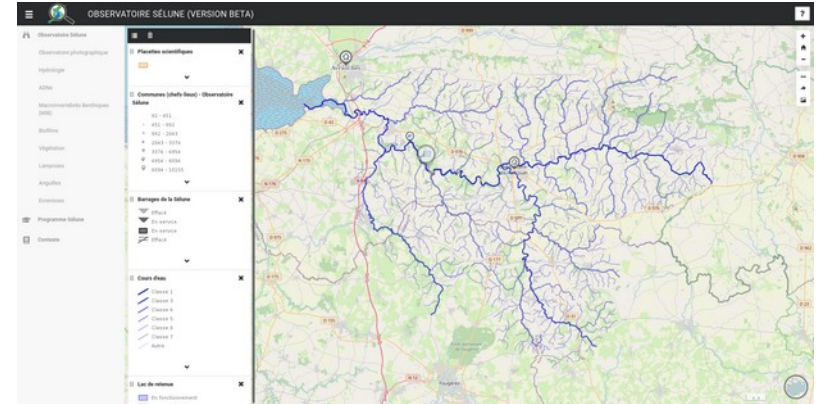
→ Plus proche d'un **data warehouse** (entrepôt de données). *Un petit.*

*NB : Composante spatiale (géométrie) isolée dans une table*



# Spatialisation

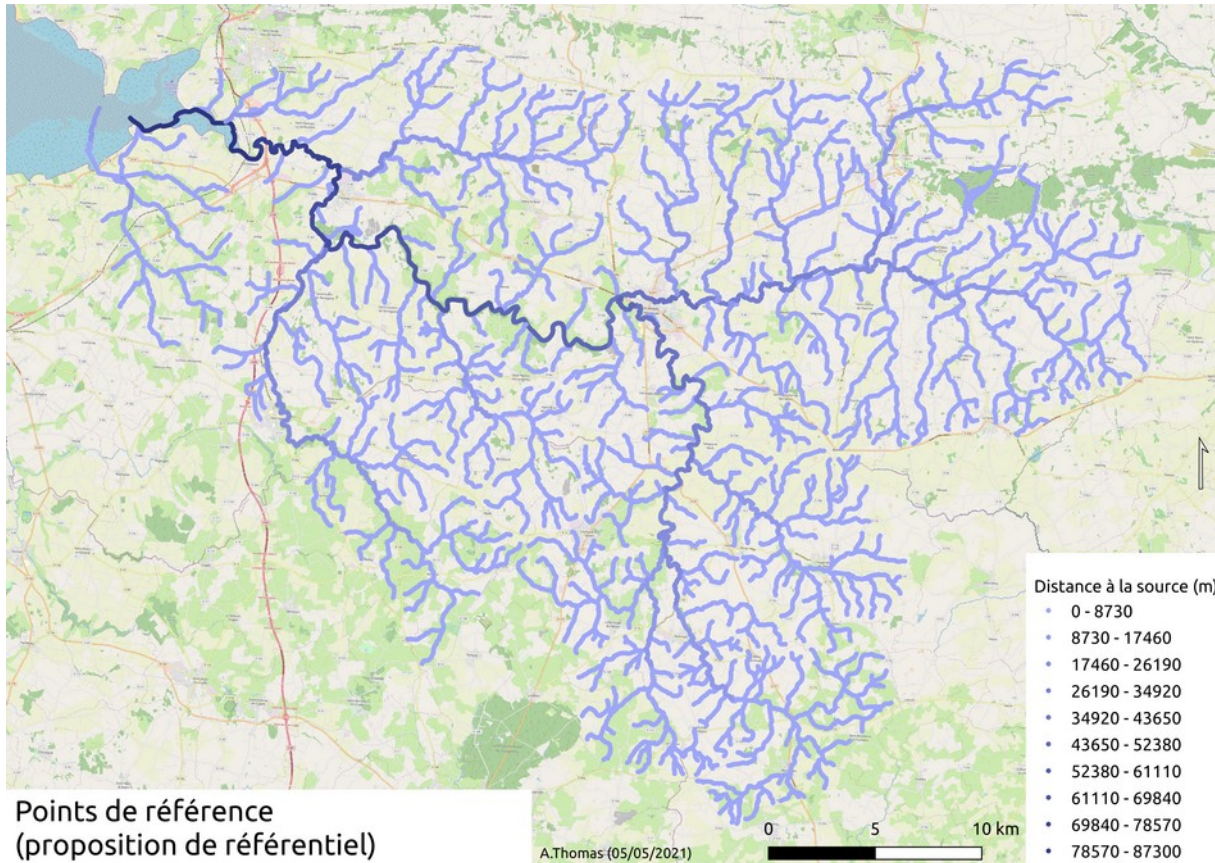
- Forte demande de tous les acteurs
  - Scientifiques ou non (chantier)
  - Besoin de recenser et localiser :
    - les études passées et actuelles
  - Plus de secteurs d'étude publiés (actuellement) que données associées
- Mise en place d'un portail carto ([mviewer](#))
  - Très rapidement utilisé
- Données et métadonnées diffusées par un IDS
  - Mais limite pour les séries temporelles...
    - l'application web devra compléter (API?)
  - Métadonnées méritent d'être stockées en base de données (et idéalement pouvoir créer/mettre à jour l'IDS)



<http://geowww.agrocampus-ouest.fr/selune/>



# Spatialisation : référentiel



- Acteurs se basent sur :
  - Entités hydrologiques
  - Toponymes
  - Noms arbitraires
    - Besoin d'un **langage commun**
- Solutions
  - Centraliser tous les sites dans une table ? (et les vérifier ?)
  - Proposition d'un **référentiel**



# Spatialisation : référentiel

- **Points** de référence
  - basés sur cours d'eau + distance à la source (PK)
  - infos : z, toponyme le + proche
- **Objectif** : un code unique pour tous les secteurs, utilisable par tous



# Pour conclure (ou presque)

- **Programme Sélune riche :**

- De son histoire (2012...) et de ses enjeux
- Par ses thématiques et ses acteurs
  - SISelune devra en être le témoin privilégié
  - SISelune doit occuper maintenant une place centrale



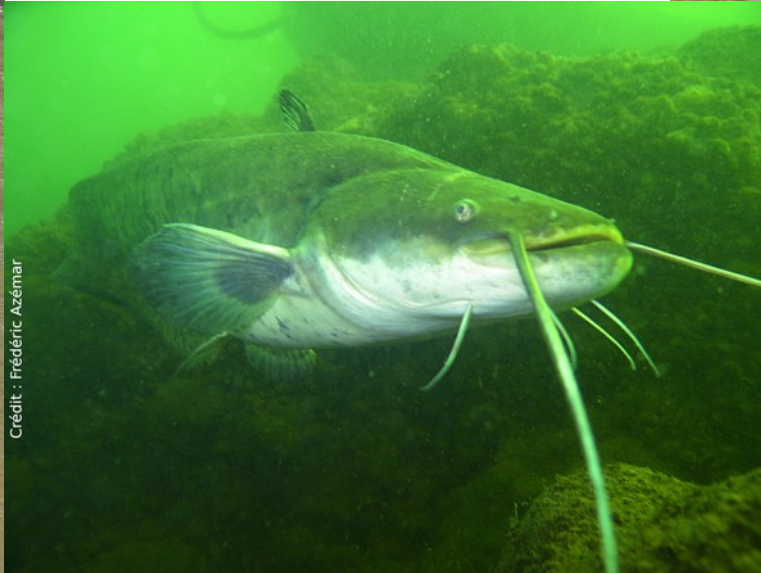
- **Démarche :**

- Qui **s'adapte aux contraintes**
- Avec un **fort accompagnement des acteurs** (coût en temps)
- Qui **part des données** (spécifique → + générique)
- « **Souple** » (sans être tout à fait agile)
- Qui **évolue** (à son rythme) : intégration ralentit développement de l'application web, mais alimente la réflexion
- **Pas forcément sophistiquée**, mais avec un maximum de bonnes pratiques
- Qui a été **acceptée**

# Pour finir de conclure

- **SISelune**

- Très **attendu** (un précédent projet avorté)
- Privilégie des **services existants**  
*(merci à ceux qui les maintiennent)*
- **Composante spatiale** prépondérante
  - seul moyen de communication entre tous les acteurs
- Peu original, mais très **complet**
  - beaucoup voire toutes les **problématiques liées aux données**
- N'est **pas terminé**, vos retours sont les bienvenus



**Merci**



# Un développement agile ?

- Un peu :
  - Relation régulière avec les « clients » (acteurs)
  - Organisation en « sprints » (~ 1 par jeu de données ou fonctionnalités)
- Mais pas vraiment :
  - Pas de contrôle absolu sur le contenu du sprint (disponibilité du/des acteurs)
    - difficulté pour planifier...
  - Tests pas suffisamment pratiqués (mais j'y travaille)
- **Amélioration continue**

