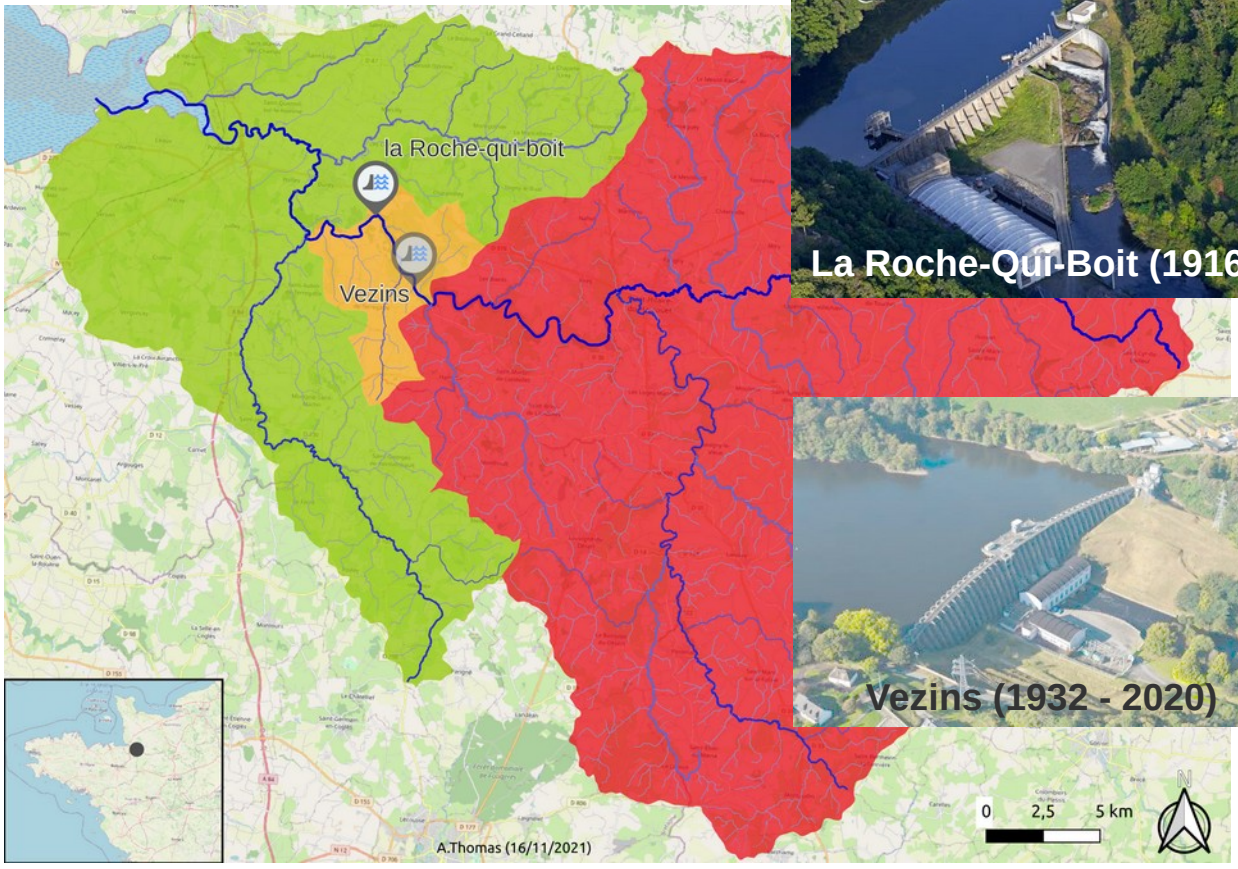
SISelune : où en est-on ?

Alban THOMAS 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 (« boite à outils »)



Acteurs / Publics

- **Programme scientifique**
  - Scientifiques (!)
  - Produisent des données
  - Doivent les diffuser

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

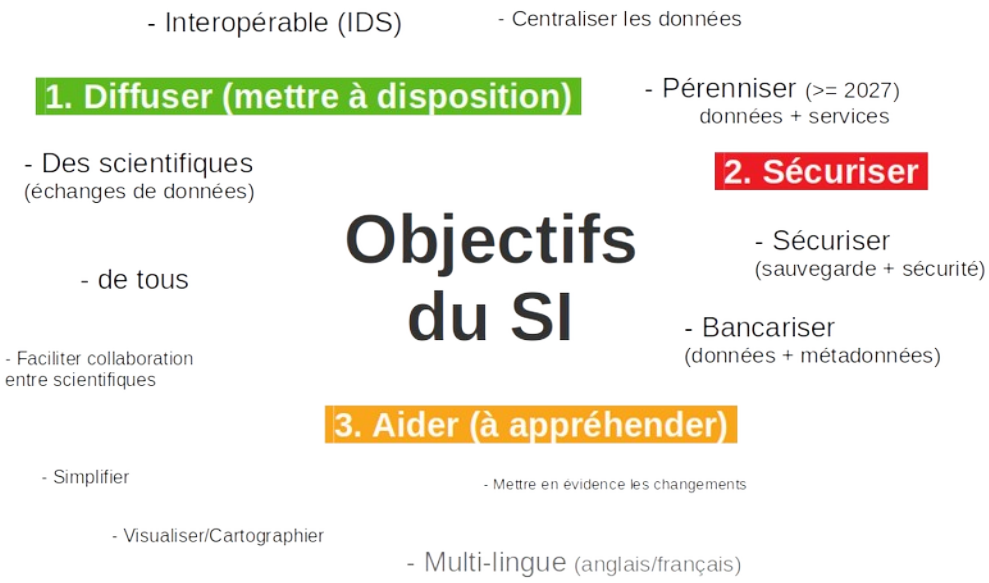
Observatoire Photographique du Paysage de la Sélune (SMBS & Univ. Paris Nanterre)

- **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

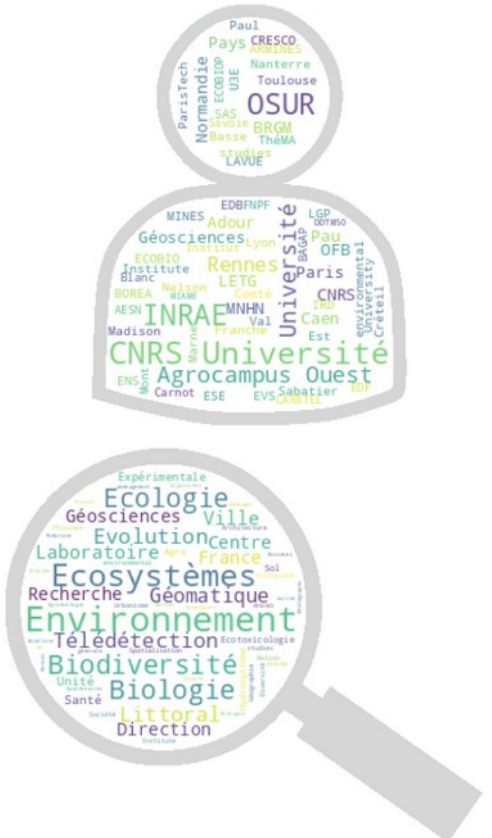


- Début du projet : 02/2020

→ Retour d'expérience (à une étape charnière)

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)



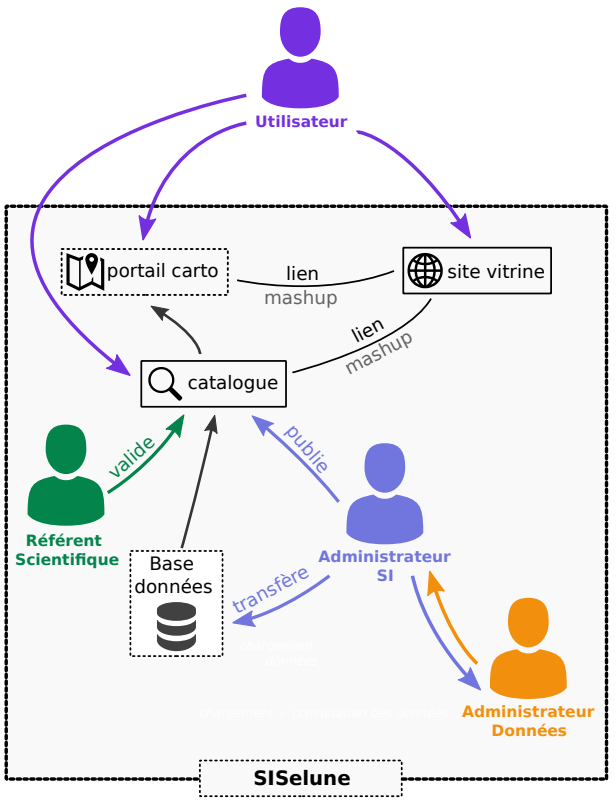
Rôles dans SISelune

- **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)

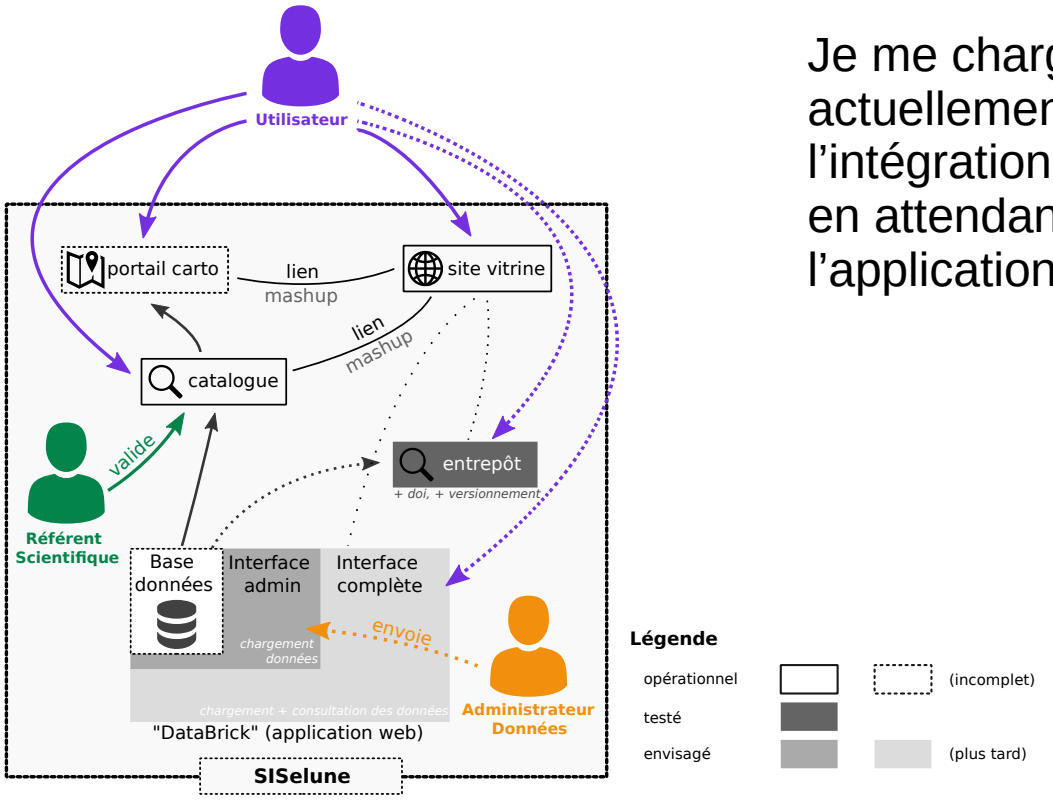
- **Utilisateurs** : acteur, avec un compte. Accès à toutes les données
- **Visiteurs** : consulte librement le SI

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

Architecture de SISelune



Actuellement

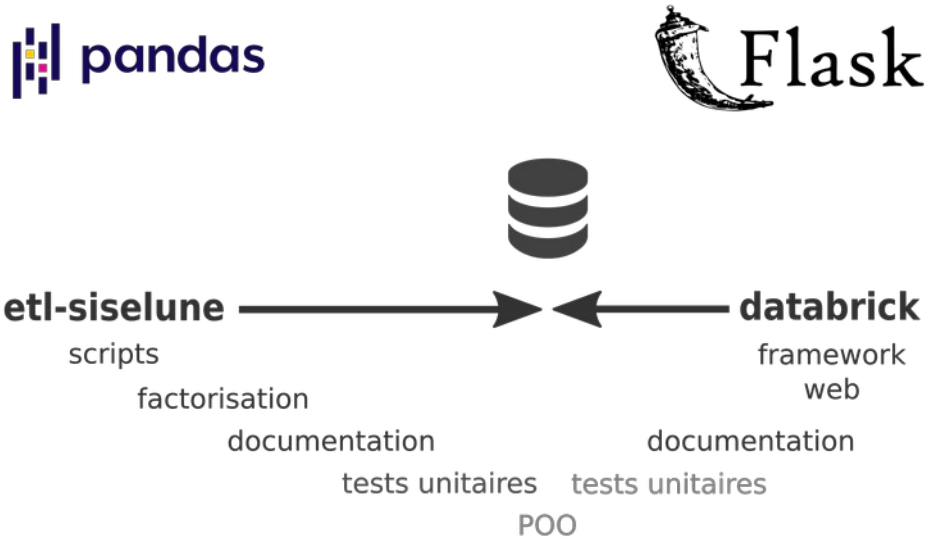


A terme

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

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

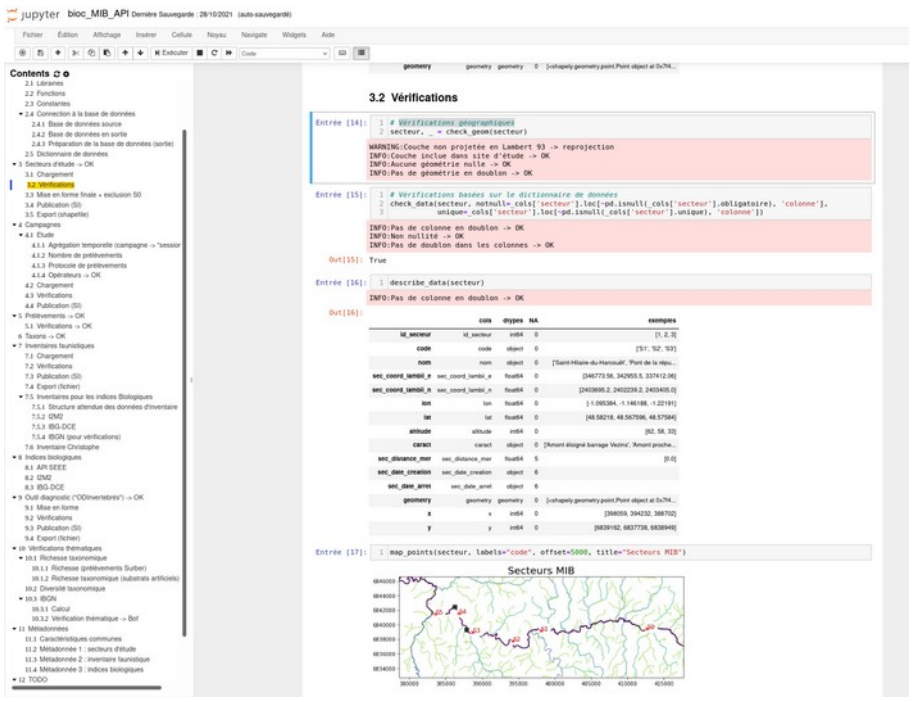


~2300 Sloc (Python) ~280

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

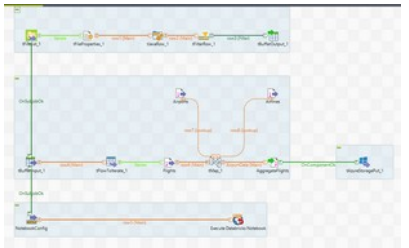


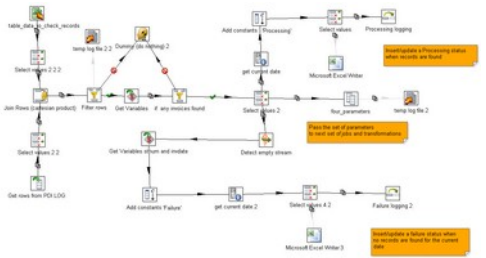
- Code + sortie du code + texte (+ images)
- Support d’échange si :
  - Code reste « discret »
  - Structure claire
- Intérêts :
  - Vérifications (dont comparaison avec précédents résultats)
  - Tester et soumettre des visualisations
  - Export d’un « rapport d’intégration »
- Limites :
  - Contenu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

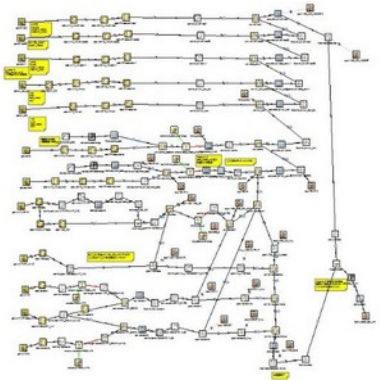
Intégration : coder « à la main » ?

- 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)







✓

?

X

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

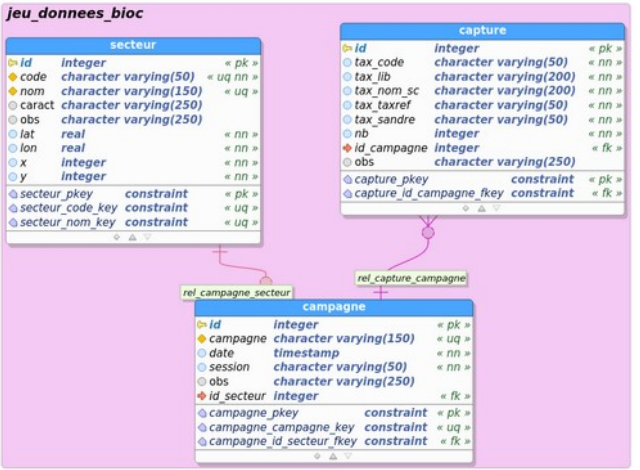
- Données **centralisées** mais **indépendantes**
- Schémas (Gestion des droits)
  - **1 schéma = 1 jeu de données** (isolement)
  - Public : **diffusion** (vues)
  - app : logique applicative

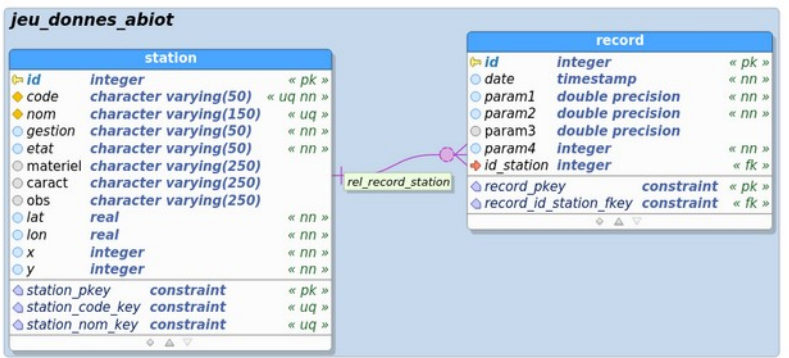
Base de données : Un lac de données ?

- 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



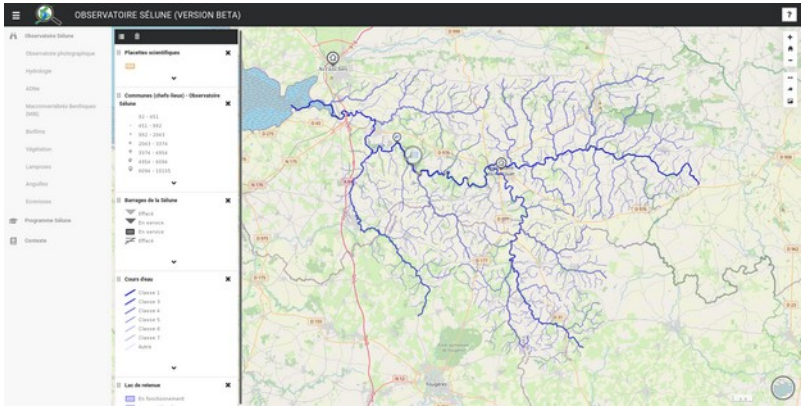


Spatialisation

- Forte demande de tous les acteurs
  - Scientifiques ou non (chantier)

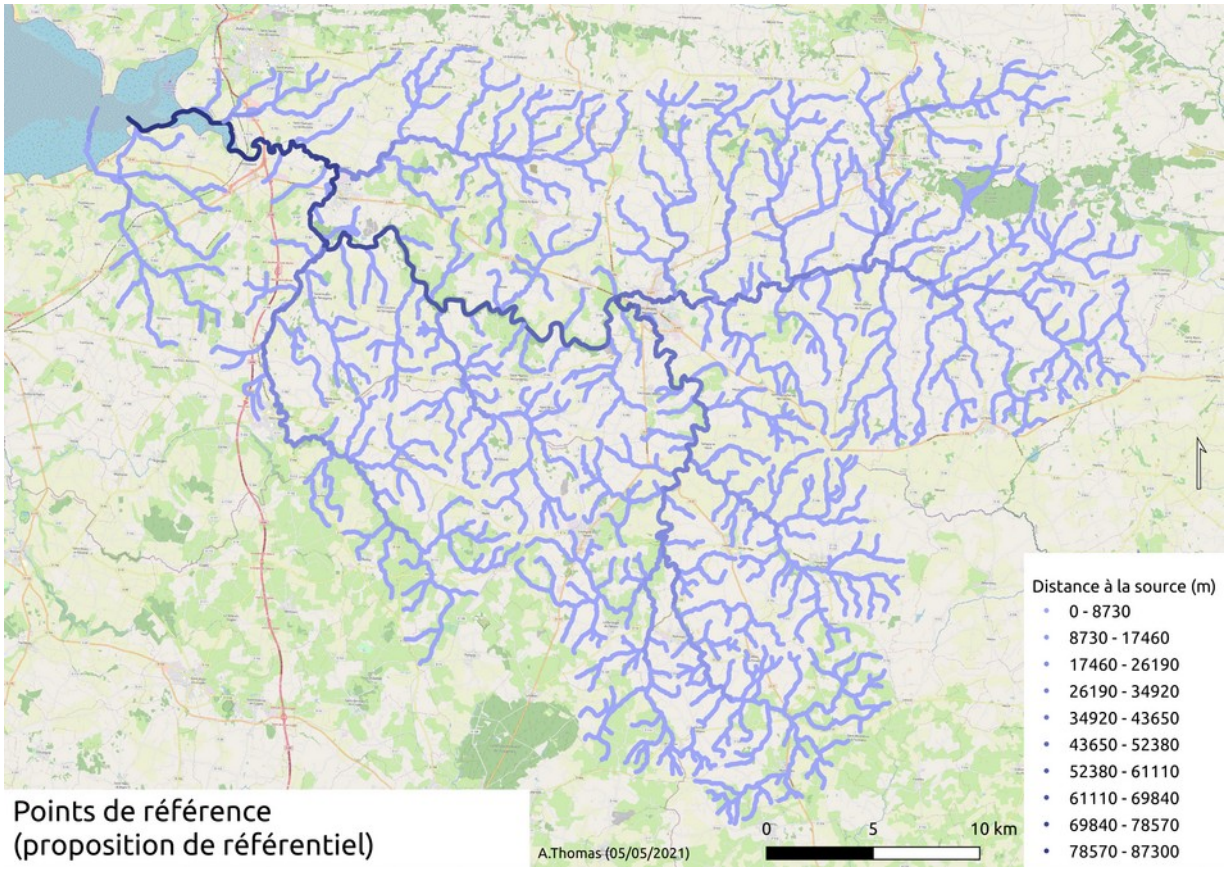
  → Besoin de recenser et localiser :
    - les études passées et actuelles

  → 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



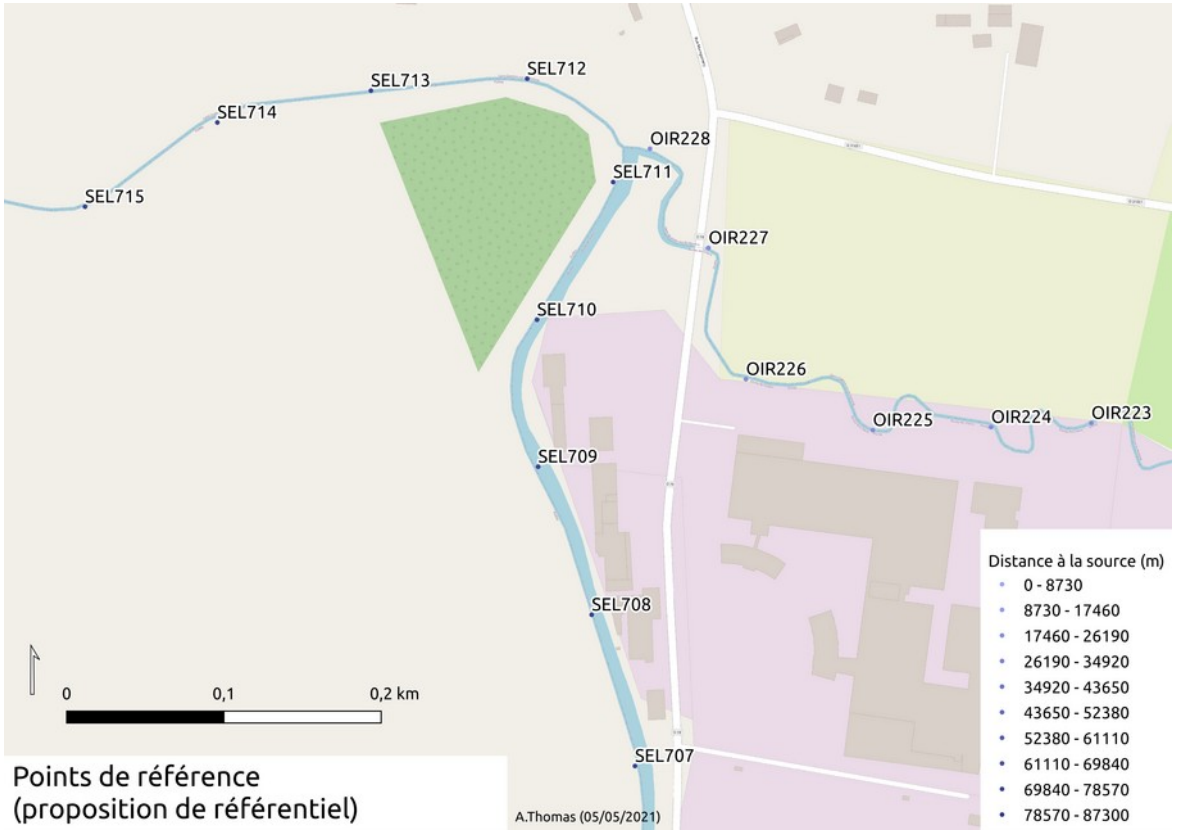
Points de référence
(proposition de référentiel)

- 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



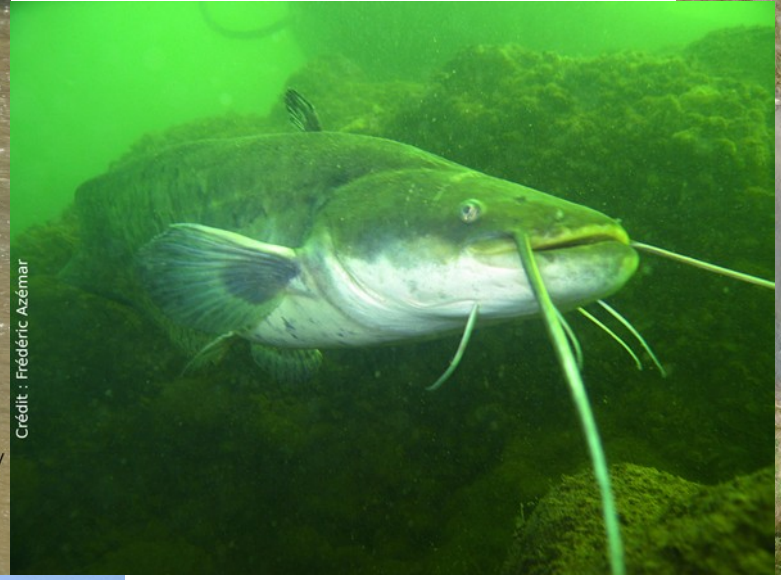
Pour conclure (ou presque)

- **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**