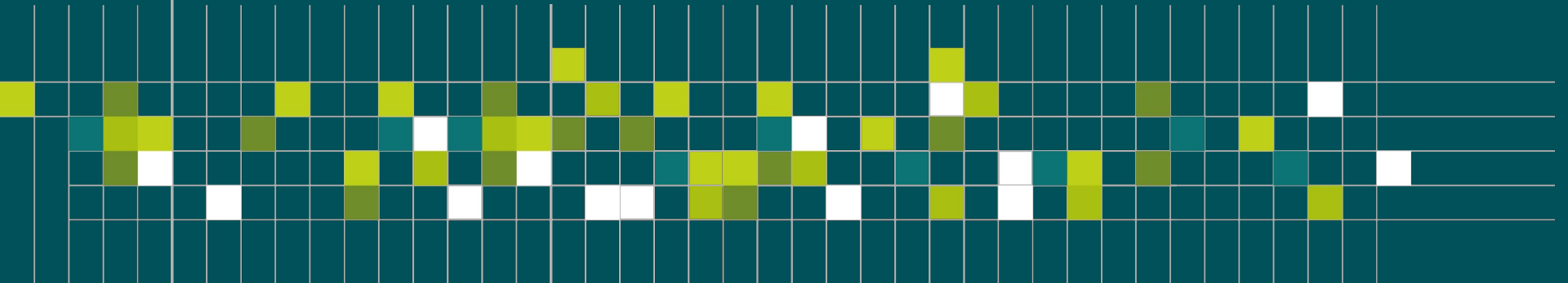
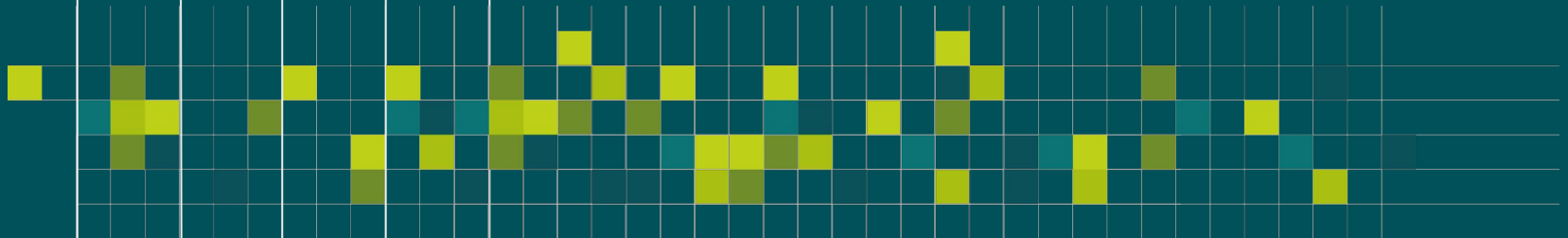
Etalab talk - 26 janvier 2023

Partage des données “d'intérêt général”



Sommaire



Introduction

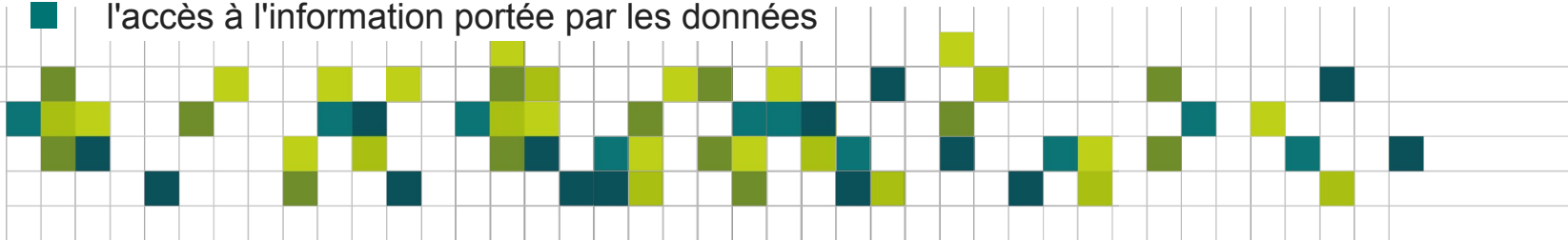
Comment améliorer le partage des données "d'intérêt général" ?

Déroulement

- Questionnement autour de deux exemples de jeux de données data.gouv.fr
  - Jeu de données IRVE : Pourquoi est-ce qu'on trouve des données incohérentes ?
  - Jeu de données qualité de l'air : Pourquoi est-il impossible de trouver des données ciblées ?
- Présentation complétée par des analyses (liens)

Sujets abordés

- les formats d'échange et l'optimisation des données
- la structure des jeux de données et la cohérence des données
- l'accès à l'information portée par les données



Introduction

Qui sommes nous ?

Projet Environmental Sensing

- Améliorer l'interopérabilité des données environnementales
- Analyse / Méthodologie / Propositions / Solutions

Interopérabilité

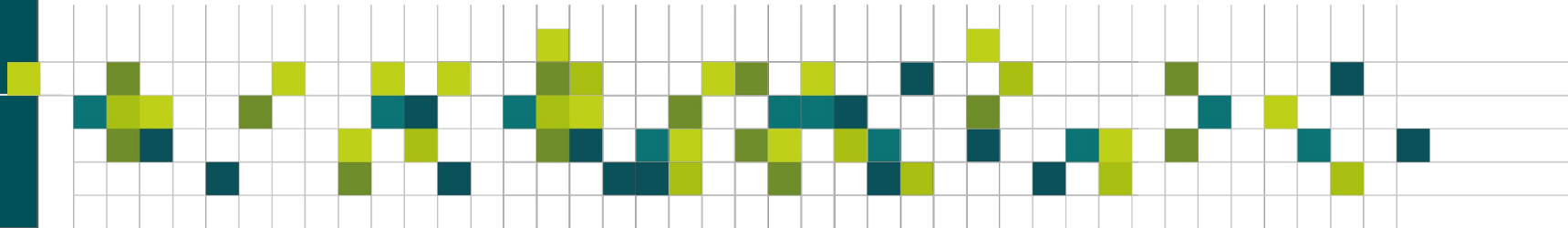
- Format d'échange
- Structuration
- Sémantique
- Accessibilité

Données environnementales

- Données structurées
- Notion d'observation
- Métadonnées
- Multi-dimensions

Exemple de travaux

- Bluetooth / air
- Schéma de données
- Format tabulaire



Exemple IRVE

IRVE : Problème

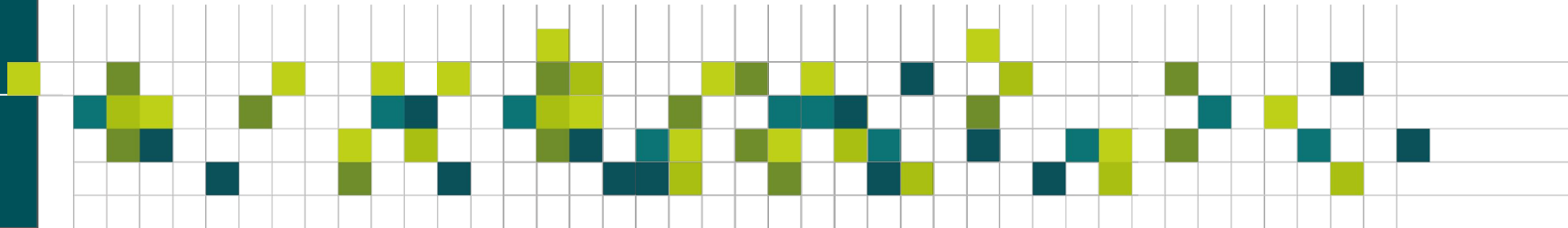
"Je veux exploiter les données opendata pour visualiser sur une carte les bornes existantes mais dans le fichier j'ai des bornes avec plusieurs coordonnées ?!"

Exemple : Camping Arinella : [9.445075, 41.995246] ou bien [9.445073, 41.995246]

Réponse

"C'est normal, il y a une ligne par point de charge donc on duplique les coordonnées avec des risques d'erreur (données)"

"En plus, c'est pas clair, il n'est pas précisé si c'est les coordonnées de la station ou celles du point de charge (schema)"



Exemple IRVE

IRVE : Problème (niveau 2)

"Pourquoi est-ce qu'il n'est pas explicitement précisé que les coordonnées sont celles de la station ?"

Réponse

"Si, c'est indiqué ! (schema)

Dans la description du champ 'coordonnéesXY' du schéma de données : *"La longitude suivie de la latitude en degrés décimaux (point comme séparateur décimal) de la localisation de la station exprimée dans le système de coordonnées WGS84 au format [lon,lat]. "*

Le problème, c'est qu'il n'y a pas de moyen pour identifier, décrire et contrôler cette information (guide schéma) "

Exemple IRVE

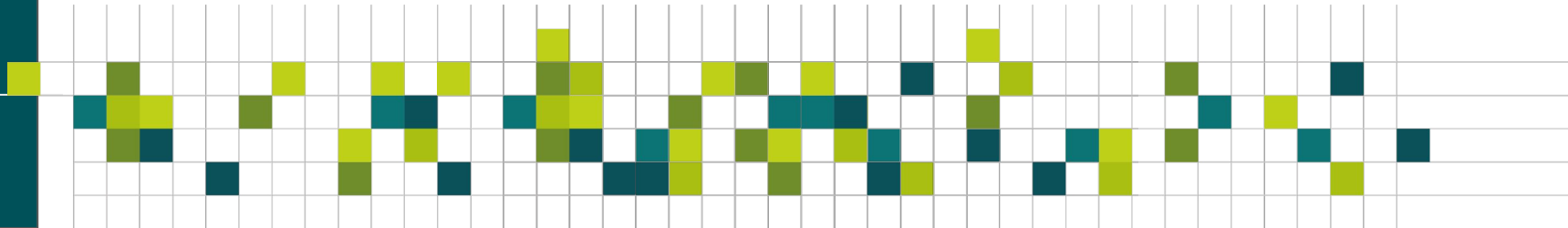
Comment résoudre le problème ?

Problème : Gestion des dépendances entre champs

- Aspect méthodologique (Modèle de données <-> Schéma de données) -> voir guide ouverture des données
- Documentation dans les schémas de données (inexistant)
- Outils de contrôle (pas d'outils)

Processus de résolution

- Etude de solutions
- Discussion d'une proposition (via TeamOpendata)
  - Interoperabilité et standards, Evolution TableSchema(top10), Methodologie



Exemple IRVE

Gestion des dépendances entre champs

Méthodologie (exemple méthodologie)

Documentation des schémas

- Propriété “relationship” -> extension TableSchema

Outils de contrôle

- Vérification (exemple)
- Identification des erreurs (exemple)

Exemple IRVE

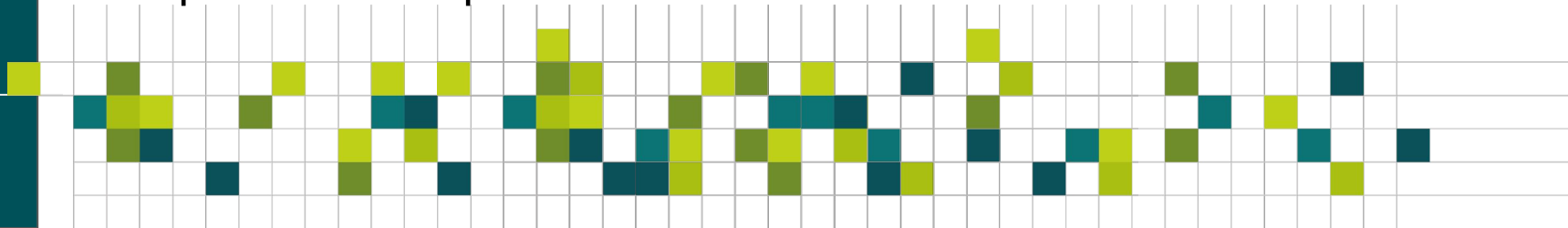
Synthèse - Mise en oeuvre

Les dépendances entre champs sont identifiées dans le MCD et explicitées dans le schéma de données.

Le contrôle (avec identification des erreurs) peut être effectué simplement dès la collecte et sur les données consolidées.

L'utilisation d'un processus d'agrégation permet de remonter à la source pour traiter le problème.

Des indicateurs simples peuvent être mis en place pour le respect des dépendances définies.



Exemple AIR

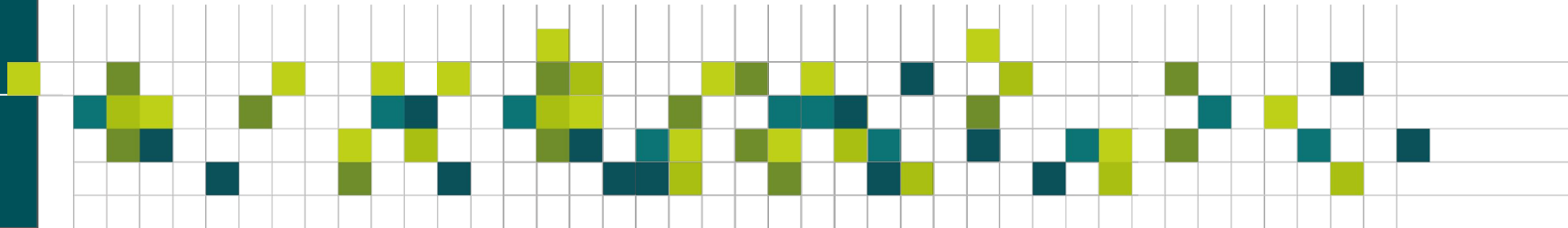
AIR : Problème

"Comment est-ce que je peux accéder facilement à la qualité de l'air concernant une zone donnée et une période donnée ?" (détail)

Exemple : Actuellement je suis obligé de télécharger plusieurs (1 par jour) gros fichiers CSV (10 Mo) et de regrouper les données manuellement ?! (liste des fichiers)

Réponse

"Les mesures sont très nombreuses et contiennent beaucoup d'informations, on est donc obligé de séparer les données en plusieurs fichiers."



Exemple AIR

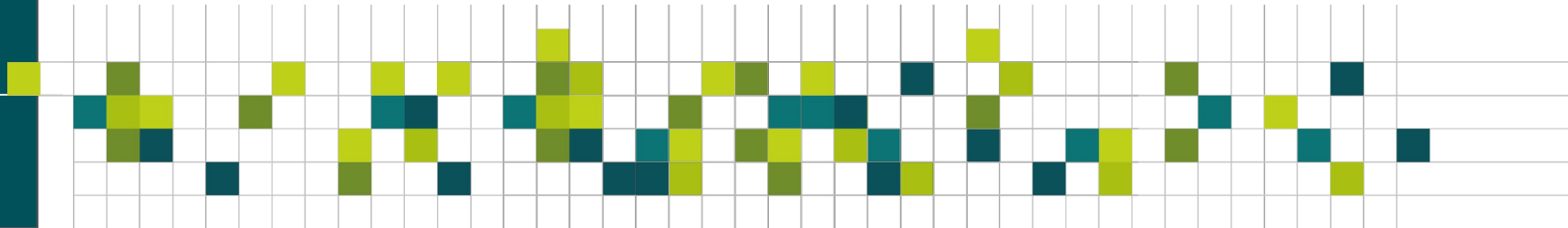
Comment résoudre le problème ?

Problème : Format CSV peu adapté à des données multi-dimensionnelles (taille, accessibilité) (voir indicateur)

- Etude de formats alternatifs sans duplication (avec codage) (voir exemple) (impact volume)
- Outil d'extraction (exemple API)

Processus de résolution

- Discussion d'une proposition (via TeamOpendata)
  - Interoperabilité et standards, Est-ce la fin du format CSV (top10) ?



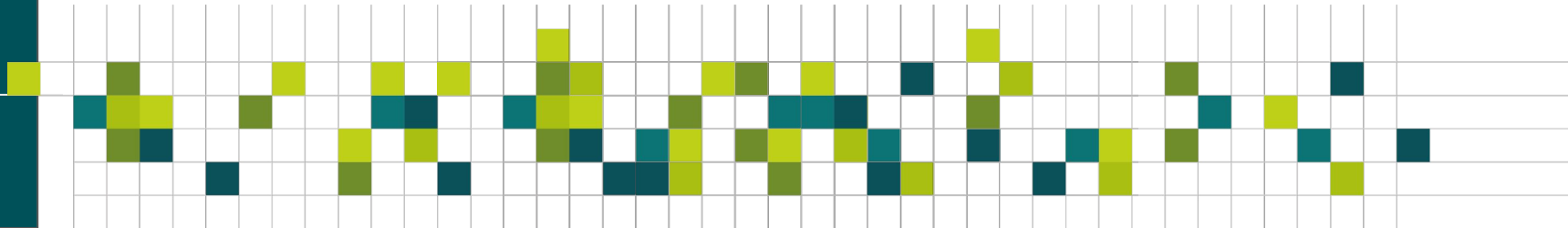
Exemple AIR

Synthèse - Mise en oeuvre

La mise à disposition de fichiers CSV n'est pas adaptée aux données environnementales multi-dimensionnelles.

Un indicateur simple permet d'identifier ces cas d'usage.

Un format alternatif associé à un codage simple et accessible via API est à envisager.



Propositions

Actions à engager

Actions “à gain rapide”

- Actualiser les guide Etalab (ouverture/schéma des données)
- Déployer des outils de contrôle des relations entre champs
- Construire des indicateurs qualité (données tabulaires)

Autres actions

- Etude des alternatives à une mise à disposition par fichier CSV
- Processus d’agrégation et traçabilité des modifications
- Augmentation de la sémantique des données partagées