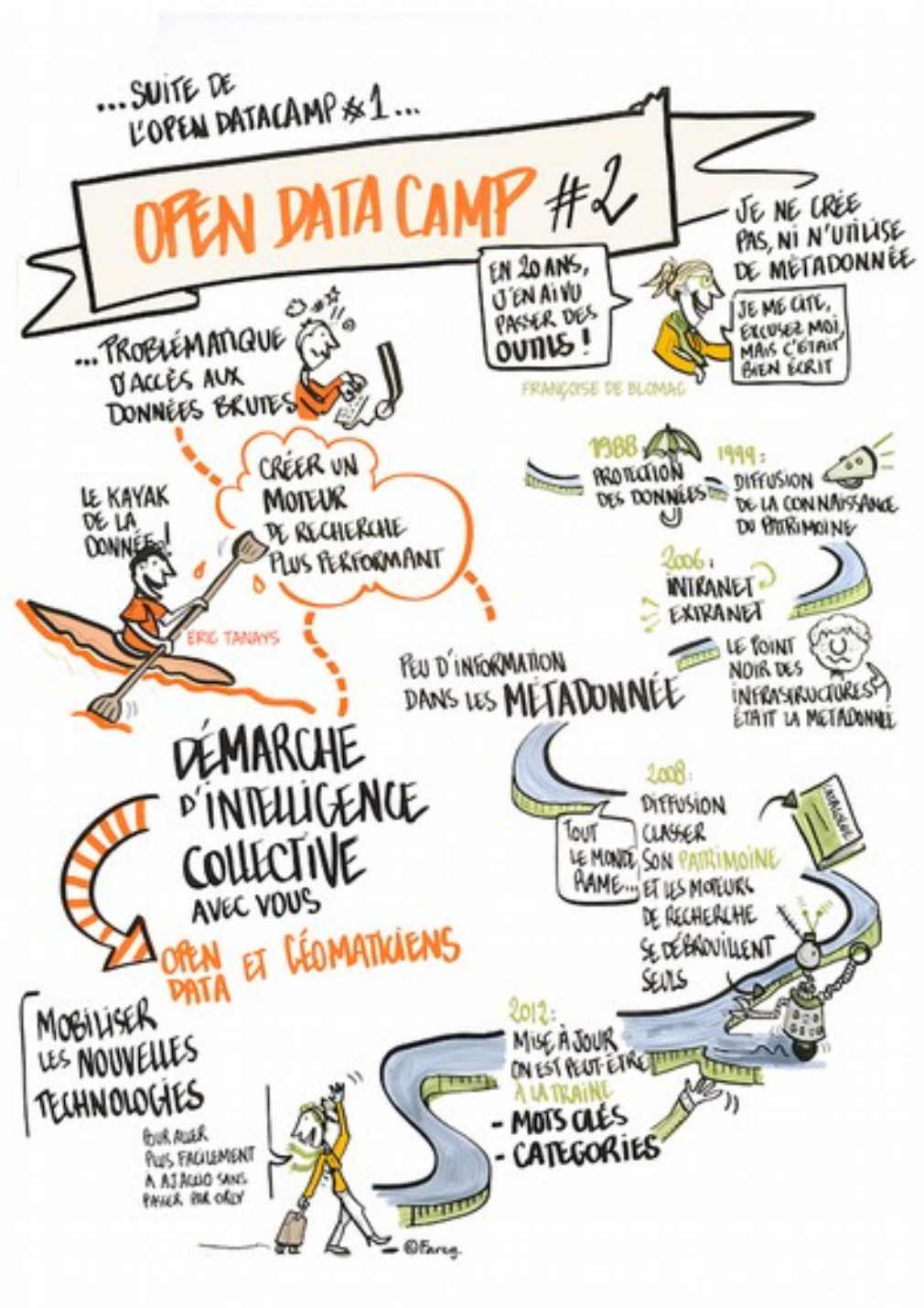
Quel moteur de recherche pour trouver une donnée publiée en OpenData sur Internet?



21 novembre 2019
Synthèse des échanges de la journée

L'objectif de cette journée était de travailler collectivement sur la découvrabilité des données à partir de divers moteurs de recherche. Cette journée a été préparée par les agents du pôle SIG de la DREAL ARA, avec l'appui de Julien Monereau du Dre'Lab et le soutien du CGDD
La journée a été organisé sous la forme d'ateliers en intelligence collective. Les différentes séquences proposées visaient, dans un premier temps, à créer un collectif de travail et à définir le profil technique de ce collectif. Dans un second temps, sur la base de la description des processus à l'œuvre pour rechercher et découvrir la donnée, les participants ont identifiés les problèmes et leviers. Après une phase d'inspiration à partir de travaux de recherche de Monsieur Joliveau de l'Université de Saint-Etienne ou d'outil existant (NeoGeo, Elastic serch), les participants se sont concentrés sur la description des actions à mettre en oeuvre.

Introduction par Eric Tanay, Directeur Délégué de la DREAL

En 2018, la DREAL ARA a organisé avec le soutien du CGDD un "OpenDataCamp" qui a réuni les usagers des données diffusées par la DREAL, principalement des bureaux d'études et des collectivités.
Une cinquantaine de participants avaient répondu présents à l'invitation de la DREAL .
Cette journée a montré que, si les données diffusées par la DREAL sont un atout indéniable pour aider à conduire des études territoriales, il subsiste encore des freins notamment concernant la facilité d'accès aux diverses sources de données, dues à la démultiplication des plateformes de diffusion et des formats.

La problématique de l'accès aux données brutes a été exprimée fortement en 2018 par les participants, malgré les gros efforts faits en matière de diffusion de données par les services de l'Etat et les collectivités.
Historiquement, les DREAL ont été très impliquée dans la diffusion de leurs données dans le cadre de la mise en œuvre de la directive Inspire, et poursuivent naturellement cette démarche au delà des données géographiques dans le cadre de la loi pour une république numérique.

3 pistes de progrès ont été proposées dont une qui consiste à disposer d'un moteur de recherche performant.
Aussi, la DREAL a proposé au CGDD lors de l'appel à projet 2019 sur les "data" de s'investir sur la partie 'moteur de recherche'.

Un outil de recherche performant qui irait retourner des résultats de recherche pertinents sur les diverses plateformes de diffusion de données pourrait répondre aux besoins exprimés par les utilisateurs.
Ce projet a été baptisé « Kayak de la donnée », par analogie avec le moteur de recherche de vols et hôtels « Kayak ». Un outil comme "Kayak" récupère des listes de trajets, avec des dates, heures, coûts, et constitue un graphe pour calculer les meilleurs chemins à moindre coût, en tarifs ou en délais.
Toutefois les données récupérées et agrégées par « Kayak » sont déjà très normalisées.
Ce n'est pas le cas des portails qu'on souhaite interroger. Google s'y essaie avec un succès mitigé avec son projet « DataSet Search ».

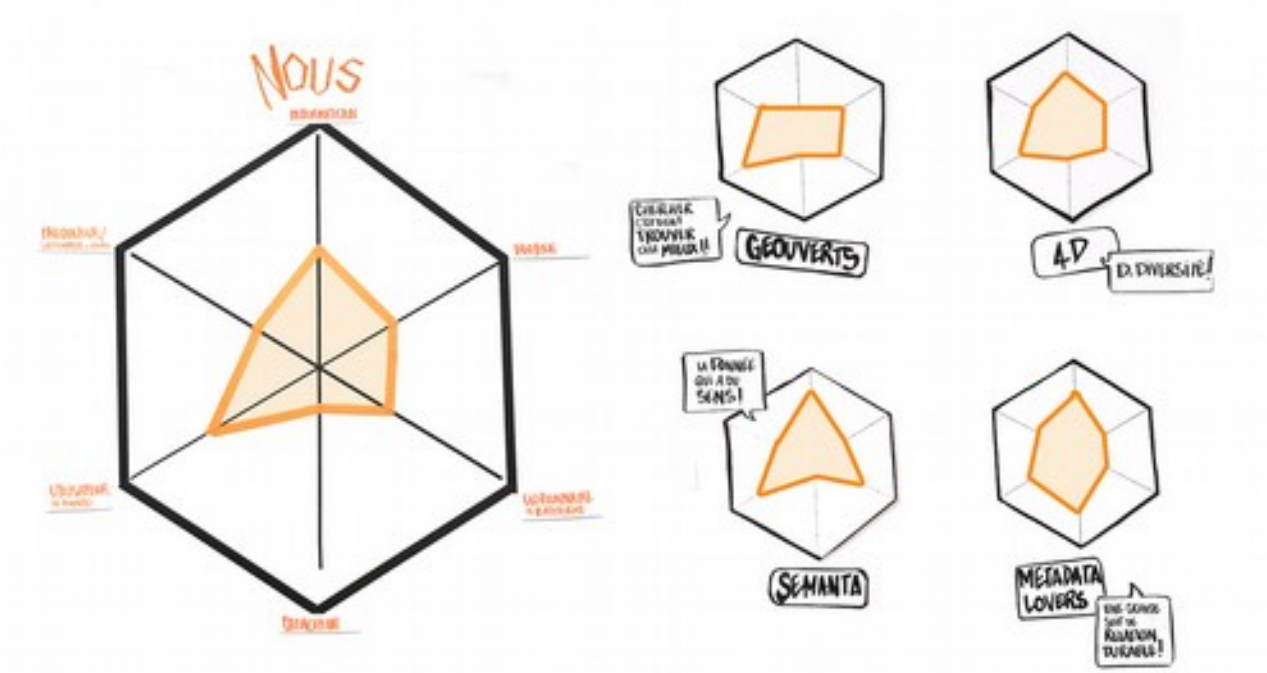
La rencontre d'aujourd'hui vise donc à poursuivre la démarche initiée en 2018 en

impliquant le monde de la recherche et des entreprises, ainsi que des porteurs de plateformes de données.
En effet, il nous faut connaitre les évolutions des solutions existantes et celles qui sont en émergences afin d'identifier les leviers à actionner pour répondre aux besoins de manière efficace.
C'est pourquoi nous vous proposons de travailler ensemble aujourd'hui en s'appuyant sur les méthodes de l'intelligence collective pour avancer sur ce besoin identifié de "super-moteur-de-recherche"
Notre objectif est que les utilisateurs d'une donnée ne passent pas du temps pour la trouver, mais plutôt pour l'exploiter afin d'en tirer une valeur ajoutée.
"Chercher une donnée" est aujourd'hui facile, mais "trouver la donnée qu'on cherche" facilement et rapidement est une autre histoire...

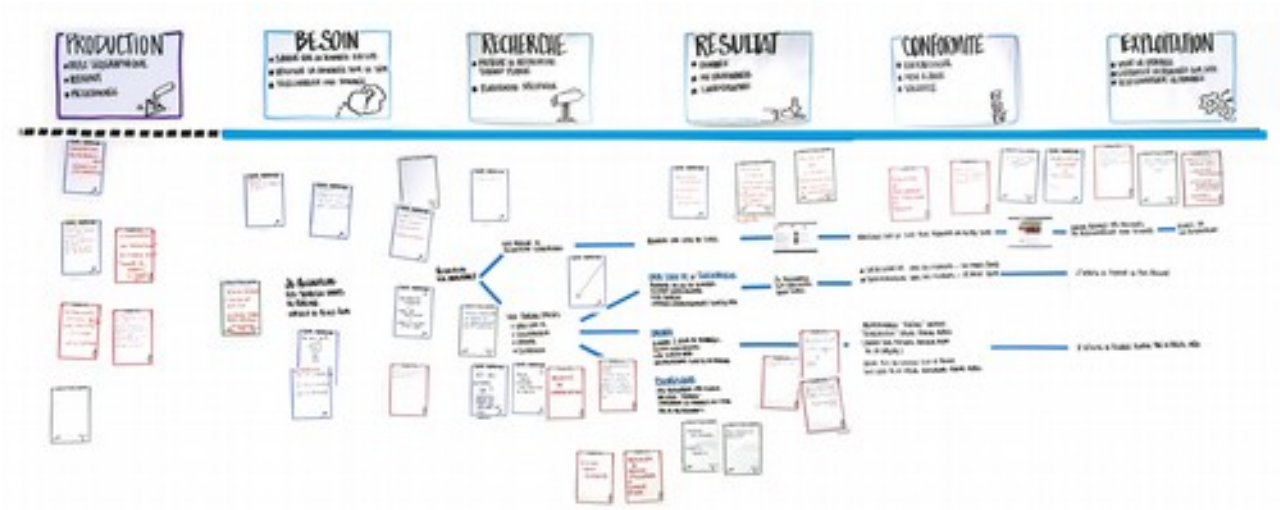
---

20 ans d'observation des métadonnées en 20 minutes par Françoise de Blomac, Decryptageo

Retour sur l'icebreaker et les profils des participants



4 équipes ont travaillé lors de l'atelier du matin à identifier les manques, problèmes et leviers sur "la frise de la découverte des données"
Le profil des participants est principalement "utilisateur", suivi de "informaticien" et "producteur de données"
Les équipes ont travaillé sur les concepts ci dessous...



...pour un résultat final :

Les étapes manquantes sur la frise,

Indexation et Référencement des plateformes
Identifier / référencer toutes les plateformes
Trouver la meilleure porte d'entrée

Caractérisation de la donnée : titre avec contexte, traçabilité, éviter les doublons, politique des gestion des plateformes pérenne
Se remettre dans le contexte : Qu'est ce que je cherche en fonction de qui je suis ?, Comment adapter l'outil à l'utilisateur : profil, contextualisation, pertinence du résultat

Acculturation à la données : pédagogie, besoin, interprétation des résultats
Filtrer le résultat a posteriori, Evaluer le résultat

Etapes de recheche interne, via un tiers de confiance et "appel à un ami"

Ajouter des services associés

Les problèmes identifiés

Gouvernance, qui produit quoi, qui publie quoi ?
Doublons, multiplons

Métadonnées faites par et pour des spécialistes et non des usagers
Vision trop experte

Méconnaissance et non respect des guides et normes
Manque de schémas
Absence de labellisation

Manque d'ergonomie des outils
Manque de capacité à sémantiser et/ou comprendre le langage naturel

Manque de liens entre données
Manque de fédération des résultats de recherche

Replacer le besoin utilisateur à chaque étape
Qualifier la pertincence en fonction du besoin

La donnée de référence devrait être présente sur toute les plateformes au mêrme état
Bien gérer les données sensibles
Gestion de l'obsolescence des données
Manque d'itération entre usager et producteur : comment discuter et améliorer la production ?

Les leviers identifiés

Formation à la saisie des métadonnées
Assistance à la saisie, automatisation
Simplifier langages, outils, modes d'emploi

Mieux mobiliser la carte et les dataviz
Proposer des usages
Transformer la donnée en service

Mieux indexer les données (elastic search)
Rendre visible l'ibndexation automatique : "ce résultat vous est proposé parce que..."

Meilleurs pertinence du résultat
Mettre en place un scoring de la métadonnée

Outil à imaginer pour questionner l'usage du résultat dans une optique d'amélioration continue

Obligation de déclaration des catalogues

Kiosques / présentations

Après le déjeuner, une séance de présentations a permis d'inspirer les participants pour la suite des ateliers

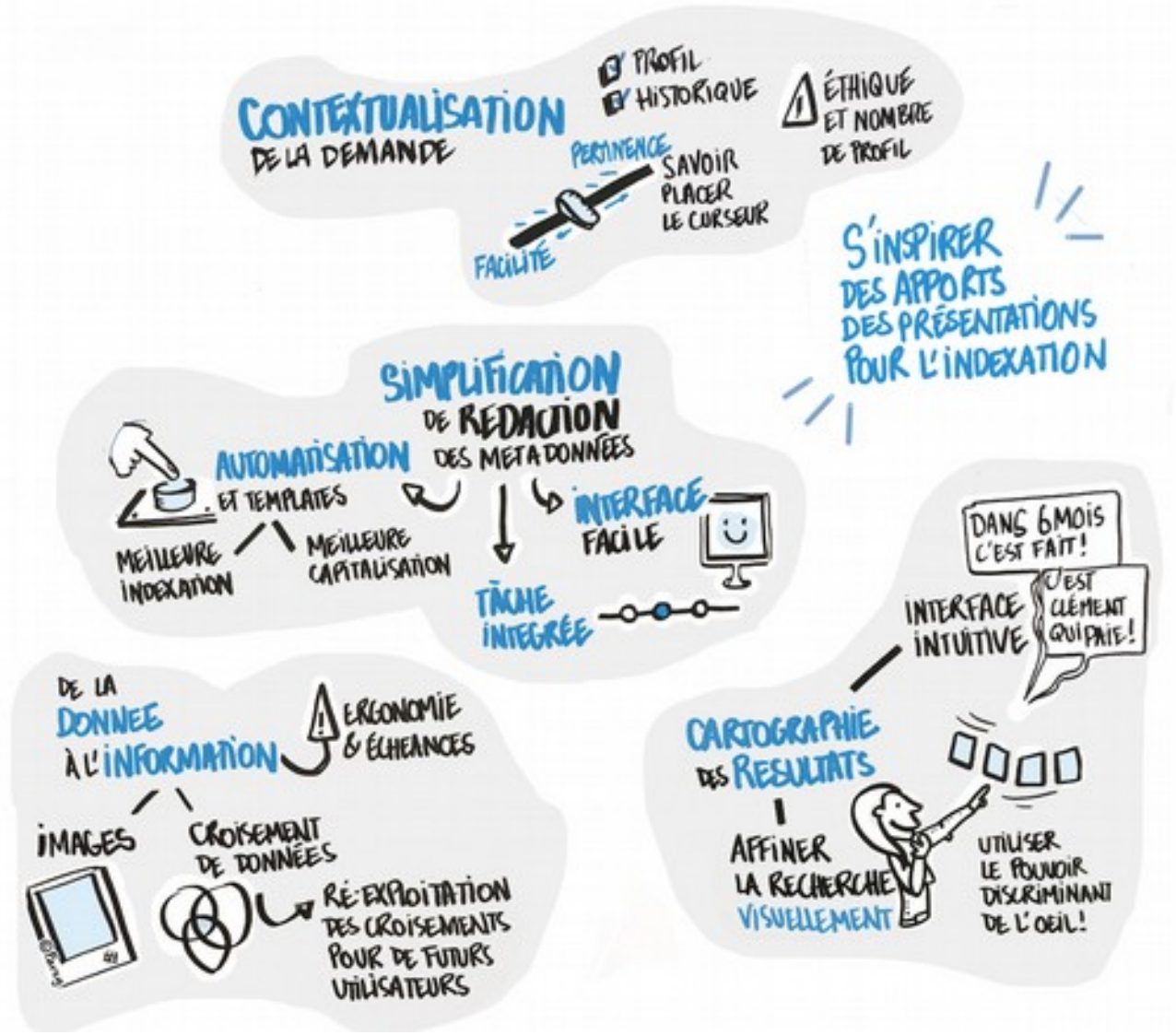
David Pilato de la société Elastic, présent lors de la journée, a fait une présentation du moteur de données Elastic Search

Thierry Joliveau, professeur de géographique , responsable du master de géographie numérique de l'université de Saint-Etienne a fait une présentation vidéo sur un exemple d'extraction de lieux dans les romans parisiens de XiXème siècle

Guillaume Sueur de la société NeoGeo a fait une présentation vidéo du moteur de recherche alliant geonetwork et elastic search développé pour le Grand Lyon...

Ateliers de l'après-midi

Les Ateliers de l'après-midi ont permis de rechercher des pistes de solutions à partir des problèmes et des leviers identifiés par les participants le matin



1/ Contextualiser les recherches

Prendre en compte le contexte de la demande pour mieux la qualifier au regard des mots utilisés

Le bénéfice attendu est une pertinence accrue des résultats, une recherche facilitée ou plus simple

Les conditions de réussite : inofs de profils-usager (ethique) / eviter la bulle informationnelle / bon niveau d'intégration / Avancer progressivement.

Echéance : POC en 6 mois en intégrant des évolutions simples déjà existantes.

2/ De la donnée à l'information

Proposer des services associés (cartes, dataviz, documents, images)
Personnaliser les services, croiser les données
Bénfices : le service au centre de la recherche; gain de temps, anticiper les usages
Condition de réussite : Qualité de l'indexation, pertinence de l'info proposée, ergonomie des services
Echéance : Pre-requis de disposer d'un moteur de recherche permettant l'indexation

3/ Navigation graphique dans les résultats de recherche

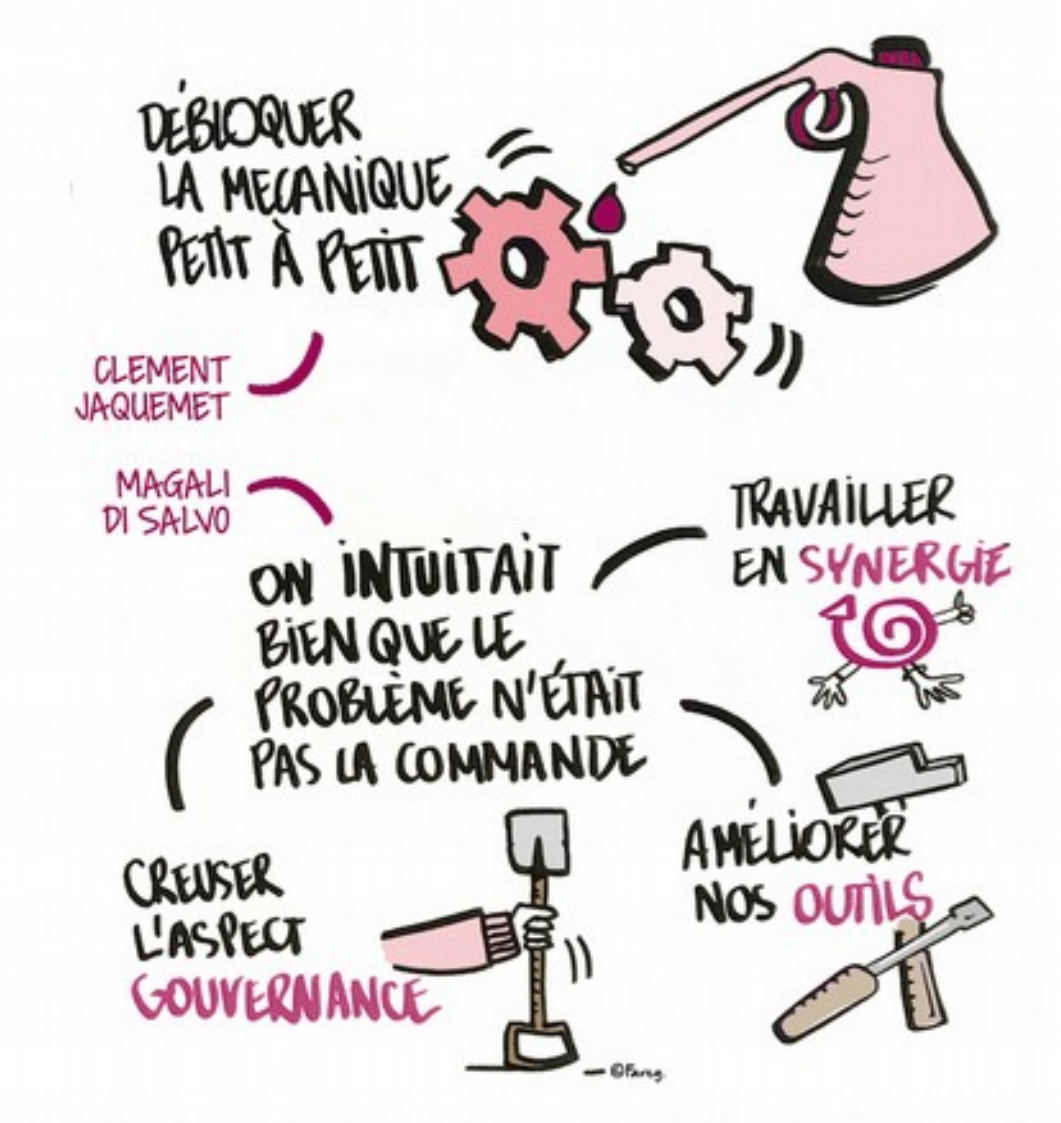
Présentation des résultats de rercherche distinguant cartes et données avec fenêtre de visualisation interactive avec comparaison de données (slifer, loupes...)
Faire le lien cartes et données
Bénéfice : Permet d'affiner la recherche visuellement (controle qualité) et de faire une exploitation sémantique des cartes pour trouver les bonnes données
Condition de réussite : Cartes doivent être décrites, service de visualisation opérationnels et interopérables / interface intuitive et efficace
Echéance : 6 mois

4/ Simplifier la saisie des métadonnées

Utilisation de référentiels pour les champs à saisir
aide contextuelle compréhensible
automatisation
templates pour les texte libres
bénéfice : meilleure indexation, découvrabilité, traçabilité, capitalisation réelle des connaissance
Condition de réussite : Soigner l'interface homme-machine
rapprocher les outils de production de données et de métadonnées
alimentation incrémentale des métadonnées
Echéances : le volet informatique peut prendre 1 mois selon le financement

Rapport d'étonnement de Françoise de Blomac

Conclusion par le CGDD (Clement Jaquemet) et la DREAL (Magali Di Salvo)



Au final, ce n'est effectivement pas le "moteur de recherche" le coeur du problème mais surtout la manière dont nous remplissons les métadonnées et les outils peu ergonomique et trop "ouverts".

Il s'agirait donc dans un premier temps de travailler en synergie pour améliorer les outils existants. Après l'identification des problèmes et leviers, le CGDD dispose donc des informations nécessaires lui permettant d'orienter de futurs développements, et pourquoi pas de poursuivre le travail avec le réseau d'acteurs.

Merci aux participants qui ont répondu présents à l'invitation de la DREAL et du CGDD

Alkante
Conseil Régional Auvergne-Rhône-Alpes
Préfécture de Région / DATARA
DDT 01
DDT 03
DECRYPTAGEO
DREAL PACA
Elasticsearch
Axel Haustant
IGN
INRA
INSA
CGDD
OPENDATASOFT
OpenIG

Cet évènement a été financé par un appel à projet du Commissariat Général au Développement Durable (**CGDD**). Il a été organisé par la **DREAL ARA** avec l'aide du **DreLab'** ; **la facilitation graphique a été assurée par Fabienne Regnier, la captation vidéo a été assurée par EcoDrone.**

COMMISSARIAT GÉNÉRAL AU DÉVELOPPEMENT DURABLE

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
DIRECTION RÉGIONALE DE L'ENVIRONNEMENT, DE L'AMÉNAGEMENT ET DU LOGEMENT
AUVERGNE - RHÔNE-ALPES