CR du Copil “Bancariser Ensemble les données”(BED)
Visioconf - lundi 16 et mardi 17 novembre 2020

Lieu : Visioconférence avec zoom - https://univ-grenoble-alpes-fr.zoom.us/j/3860669552

Nombre de participants : D. Sarramia (ZAEU), C. Pignol (ZAA), I. Charpentier (ZAEU), W. Heintz (ZAPYGAR), V. Girard (eLTER FR)

Ordre du jour sur ces 2 journées

1. Rappel des enjeux
2. Bilan 2020 : budget / actions
3. Feuille de route 2021
4. Préparation du Codir du RZA (16 et 17 décembre)

***

<u>Lundi 16 novembre</u> (09h30- 17h30 Visio)

09:30 - 10:00 :	Partage ordre du jour et définition du programme
10:30 - 12:30 :	Shapes / CCTP Géocatalogue / ROZA / ZA Timeline
12:30 - 13:30 :	Repas
13:30 - 14:45 :	ZA Timeline / DOI et dataverse / Voc contrôlé
15:00 - 16:00/30 :	Discussion avec THEIA/OZCAR, avec Isabelle B. Sylvie G. Véro. C. / Voc contrôlé
16:30 - 17:30 :	Lien avec PNDB / Programmation jour suivant

<u>Mardi 17 novembre</u> (09h00- 16h00 Visio)

09:00 - 10:30 :	Geoflow communication / Voc contrôlé suite
10:30 - 12:30 :	Échanges avec la direction RZA (MN Pons, C. Schrive)
12:30 - 13:30 :	Repas
13:30 - 15:00 :	Plan gestion de données & politique de données
15:00 - 16:00 :	Budget et actions 2021

Sommaire

Shape des contours des ZAs 2

Géocatalogues CCTP 3

ROZA 3

ZA Timeline 4

Dataset search Google 7

Vocabulaire contrôlé / thésaurus 4

Geoflow 5

Lien avec PNDB - EML/Metacat 7

Plan de gestion de données du RZA 7

eLTER FR 8

Ecole thématique e-ENVIR 2021 8

Budget 2020 8

Veille 9

Budget 2021 9

Présentation au comité direction du RZA 10

Annexes 11

Shape des contours des ZAs

6 réponses à notre demande. Exercice réalisé de manière partiel. Rendre visible le travail demandé via un affichage dans le geonetwork.

Aller vers des fiches de métadonnées pour les contours des ZAs, avec des flux de données (1 par ZA et 1 général pour le RZA).

A faire :

- Wilfried : réalisation du fichier de configuration json pour geoflow associé à la ZAPYGAR à titre d'exemple;
- Cécile, Isabelle etc. : idem avec les autres ZAs;
- Impliquer MN pour réaliser la carte du RZA et remplir les infos sur geoflow;

Deadline : 15/12 pour présenter l'affichage aux directeurs des ZAs à l'occasion du Codir RZA (16-17/12).

Géocatalogues CCTP

Lien vers le CCTP = https://docs.google.com/document/d/1bTUX399cu9Y18GfRBGgIJDYnfWOg8ReDAkRsG6rpb8/edit?usp=sharing

(attention annexe 3 sur les thésaurus, faire un choix parcimonieux si difficile à installer)

Dictionnaires des données et catalogues des attributs : regarder la différence, les pistes d'amélioration (cf. def en annexe). Voir également https://help.osf.io/hc/en-us/articles/360019739054-How-to-Make-a-Data-Dictionary

A faire :

- Virginie soumet pour avis le cctp à Mathias (ok envoyé le 16/11; rdv le 27/11) + BBEES. Retour attendu d'ici décembre ;
- Cécile et Wilfried prennent contact début décembre avec le prestataire Titilus pour connaître ses dispos et interagir sur les éléments du CCTP (hiérarchisation des demandes et coûts associés).

ROZA

Nécessité d'avoir des données formatées suivant un standard simple. Or, les données sont produites par différents analyseurs, et livrées avec dans un format constructeur (sans standard) difficile à manipuler. Les chercheurs doivent faire évoluer leurs pratiques pour faciliter la gestion des données...

Courant 2020 : difficultés à mobiliser les chercheurs. Analyse : Stade de prototype après plusieurs années de développement, épuisement des chercheurs? investissement important nécessaire de la part des chercheurs, nécessité d'un appui.

Participation de Mathias Rouan (ZABrI) dans le développement de l'outil : appui expertise au début, puis progressivement investissement conséquent, non soutenable actuellement.

Quel rebond ?

- soumettre un projet pour obtenir du man-power, en remplacement de Mathias Rouan
- identifier un autre usage car c'est un outil permettant de la **data visualisation** (http://ccwbvps18.in2p3.fr/maps/visualiseur-coyoxhup#project).
    - Un exemple d'une **interface de data visualisation** d'un collègue à Wilfried = http://geowww.agrocampus-ouest.fr/mviewer/?config=/apps/inr2ae/inr2ae.xml
    - Un exemple de visualisation avec plateforme Trajectories : http://lig-tdcge.imag.fr/steamer/trajectories/public/

Aller vers une **interface RShiny** pour faire correspondre des bases/jeux de données "chercheurs" vers les bdd collectives ayant un format différent. Comment intégrer des données dans une bbd ? Geoflow sait le faire! Génération/intégration d'une couche avec postGIS...

Perspectives :

- exploitation d'une **approche avec Elastic** (Cécile/David)
- pour 2021, maintien du prototype de ROZA (serveur RZA à IN2P3) + exploration de l'utilisation de cette expérience au sein de BED
    - essai d'intégration d'une dizaine de jeux de données de carottes; Retour d'expérience et **valorisation avec un article** ? Pour les motiver, il faut faire du croisement de données, sur la base de combien de jeux de données ? (Isabelle/Cécile)

ZA Timeline

Outil construit dans le cadre de l'action transversale Homme-Nature (ATHN) afin de constituer un fichier de données contenant les informations d'un socio-écosystème, puis proposer une représentation graphique de la trajectoire de celui-ci socio-écosystème, croisement des données (cf. présentation Colloque RZA 20 ans, Isabelle Charpentier)

Version actuelle : 1 fichier excel (4 colonnes) qui donne une interface graphique
Version à venir : lien avec d'autres BDD pour ajouter des graphiques

Quel est le devenir de cet outil ?

- une fiche MD, un DOI ou encore voire avec l'agence de protection des programmes ? (Isabelle, logiciel en cours de dépôt avec la Satt Connectus)
- "Software heritage" pour stocker les logiciels (https://www.softwareheritage.org/?lang=fr) ; sinon voir avec Zenodo = https://guides.github.com/activities/citable-code/
- possibilité de produire des fiches de métadonnées "trajectoire" par socio-écosystème

Vocabulaire contrôlé / thésaurus

Choix de s'appuyer sur des thésaurus existant et les recommander suivant les catégories de variables (associés à des thématiques/disciplines) vs. définir son propre thésaurus.

Réaliser une **analyse sémantique** d'un ensemble de ressources (ex. fiches de métadonnées, titre-mots-clés-abstract des publications, etc.). Identifier les algorithmes de recherche les plus adaptés (répétition, association de mots etc.). Compléter les attentes dans la note à destination d'un ou plusieurs stages : https://drive.google.com/file/d/1Mc43I0PpDioMjm1G0yT38GRHyj2dc7HV/view?usp=sharing

Construire une **compartimentation du vocabulaire d'intérêt** associée aux boxes du **schéma conceptuel du RZA** (Bretagnolle et al. 2019; bio-physique, gestion, services, gouvernance, etc.). Isabelle va fournir le fichier correspondant. Pour chacune des box il s'agira d'identifier les variables clefs et le/les thésaurus à privilégier. Pour les pools de variables de

données par boxes, des groupes de travail seront organisés avec les membres des ZAs. (cf. également p7, Figure 3 du CR AGBED2019; https://drive.google.com/file/d/1Y8eqcG0-W1Wm5fZO6c6yoR-ZAZAQHc6b/view?usp=sharing)

Nouvelles ressources à exploiter : les résumés fournis au colloque RZA 20ans

Identifier des **outils permettant de comparer des thésaurus** (nombre de mots qui matchent ou pas). S'appuyer sur l'expérience de l'INIST: **échange à organiser avec Dominique Vachez** INIST.

Logiciel à acheter ? outil qui cherche des mots dans des articles et associations de mots. Voir avec le CESAB / FRB / Collègues SHS. Exemple du logiciel libre de droit : Iramuteq (Interface de R pour les Analyses Multidimensionnelles de Textes et de questionnaires) développé par le Laboratoire d'Etudes et de Recherches Appliquées en Sciences Sociales de l'université de Toulouse Jean-Jaurès.

- IRaMuTeQ : détermine les mots-clés du corpus, les rassemble autour de mots cooccurrents et donc identifie les mondes lexicaux (en fonction de l'intensité de liens) http://www.iramuteq.org/
- TXM : permet d'envisager le corpus dans une perspective diachronique, d'en déduire la temporalité et les évolutions des objets de controverse http://textometrie.ens-lyon.fr/

Appel aux protocoles auprès des directeurs

Thesaurus de OZCAR/THEIA : (sous SKOMOS) https://in-situ.theia-land.fr/skosmos/theia_ozcar_thesaurus/en/page/atmosphericChemistry
Deux entrées possibles : (1) par variables (2) par catégories de variables qui repose sur une hiérarchisation du vocabulaire. A terme sera proposé une troisième entrée : par objet d'intérêts. **Accompagner THEIA/OZCAR** dans la précision de leur thesaurus sur la partie biodiversité (validation de la hiérarchisation des termes), et sur les objets d'intérêts. S'appuyer sur l'existant (ex. GBIF ou le catalogue of life pour les taxons)

A faire :

- lecture simultanée et croisée des résumés pour le colloque RZA 20 ans. Isabelle/Cécile/Virginie ; elaboration d'une méthodologie pour compléter la fiche de stage https://drive.google.com/file/d/1Mc43I0PpDioMjm1G0yT38GRHyj2dc7HV/view?usp=sharing
- regroupement des ressources, dont appel à protocoles
- identification de mots clefs par box_concept-ses-rza

Geoflow

Courant 2020 : prestation de E. Blondel. Développement de packages complémentaires.

Produire un **gabarit du fichier geoflow** à destination des ZAs, avec une première ligne permettant de produire une fiche test toujours visible avec la mention des attentes

Travail d'harmonisation de l'existant: convertir automatiquement les fiches existantes au format geoflow avec un lien vers les anciennes fiches
Rem. Comment optimiser le séparateur {underscore, retour chariot} ? Plutôt qqch d'unique.

Visualisateur : Aller vers un Geoserver Open FAIR Viewer (OFV) pour le RZA; prérequis: avoir un serveur web. Possibilité de voir les données. Geoserver produit un WMS via un shape; donc logique de faire un shape = 1 jeu de données. Un standard pour les flux WMS : le LSD pour garder la mise en forme avec un flux.

Thesaurus / URI automatique : indiquer des mots libre avec identification automatique de l'uri du thesaurus

Module complémentaire possible = Production d'un plan de gestion de données à minima avec les métadonnées initiales fournies dans geoflow. Attendre sans doute une adhésion plus large voire systématique dans les projets des ZAs.

Faciliter la production du fichier contact avec la **consultation d'un annuaire des contacts** via geonetwork. Développement à envisager avec E. Blondel, à discuter avec le prestataire Geonetwork.

Communication :
- texte à envoyer aux directeurs des ZAs à partir de la nouvelle direction (autour 1er décembre) > Virginie
- publication sur aspect méthodo et/ou techno RZA

Guide utilisation : en cours Julien Barde, E. Blondel et Wilfried
- publication / contact Julien Barde & Wilfried ? (guide d'utilisation -> publis spécifiques geoflow)

Guide de recommandations : 1er trimestre

Atelier à organiser dans un premier temps / aide pris en main

Montage vidéo prestation ou outil gratuit (sous Windows 10 application). Scénarios à monter en avant. Identifier un logiciel = Wilfried (ex. Video Prod). Expérience Cécile et Oton : https://vimeo.com/307296991

A faire :
- vidéo support formation
- guide utilisateur
- guide recommandation vs. PGD avec mode opératoire
- article méthodo pour RZA
- Open FAIR Viewer à exploiter
- Harmonisation des fiches MD

Lien avec PNDB - EML/Metacat

Questions de Cécile restées sans réponse, suite à une sollicitation de Yvan. Avoir une discussion en présence d'Emmanuel, de Wilfried. Rappel : E. Blondel est en contact régulièrement avec Yvan pour faciliter le transfert vers le PNDB, dans le cadre du financement de BED.

Depuis métacat (équivalent de geonetwork en EML), digestion EML. Rien ne vient de chez PNDB, et nous envoyons nos fiches ISO au format EML.

Pour les anciennes fiches de MD, avec geometa possibilité de rapatrier les fiches existantes en EML. Nécessité de s'y connaître en R et connaître la norme EML. Travail pour un stagiaire! Ne pas dégrader la fiche existante. Toutes les fiches ISO ne sont pas à transposer en EML; il est nécessaire de faire un tri / pblm quand cela vient d'autres catalogues. Stratégie de versioning pour un certain nombre de fiches (ex. RZA_prgm_prairies-1). En profiter également pour la traduction des fiches.

A faire :

- relancer Yvan
- sélectionner les fiches prioritaires pour être mises au format EML
- écrire dans le plan de gestion de données le mode opératoire
- dimensionner le travail à faire par le stagiaire https://drive.google.com/file/d/1Mc43I0PpDioMjm1G0yT38GRHyj2dc7HV/view?usp=sharing
- traduction des fiches en EN, ou bien du fichier "geoflow" recodant les anciennes fiches

Dataset search Google

https://datasetsearch.research.google.com/

Cet outil permet d'identifier par DOI/mots clefs l'ensemble des données recherchées. Les données publiées par le workflow complet de Geoflow peuvent être visibles dans GoogleDataSearch .

Plan de gestion de données du RZA

Lire les documents de AnaEE (DMP et politique des données) communiqués par C. Schrive et en faire une traduction pour le RZA. Deadline fin 2020.

Avoir un **DMP plutôt opérationnel** avec les codes d'accès etc.

NB. Il en existe un également côté OZCAR en cours de relecture.

eLTER FR

Prendre rdv avec John Watkins pour parler de nos organisations de gestion de la donnée.
Voir avec Mathilde pour participer au groupe de travail de eLTER PPP.

Prendre connaissance de l'interview OZCAR et ICT requirements

Ecole thématique e-ENVIR 2021

Rappel
Le projet entend enfin **favoriser la recherche interdisciplinaire** par la présentation et la structuration des données des chercheurs. Notamment, l'IR OZCAR et l'IR RZA balaient un large ensemble de disciplines (hydro-météo-géologie-géophysique, géochimie, socio-écologie, écologie, archéologie, sociologie…) et présentent la particularité de disposer de données sur le long terme, où la gestion et le partage des données sont des éléments clefs. Six ateliers ont été imaginés (ordre non chronologique) :

1. Un atelier **production de métadonnées** à partir d'un fichier excel et une routine R (Geoflow). Enjeux: automatisation, fiche MD passé, réactualisation, échelle de production (découverte, granularité, objets d'intérêt).
2. Un atelier sur le **vocabulaire contrôlé**. Identification des référentiels clefs des disciplines, exploration du thesaurus THEIA/OZCAR avec discussion pour compléter la partie biodiversité sur la base de l'existant, définitions...
3. Un atelier sur la **DOI-isation et les entrepôts de données** avec un jeu de données fourni pour ceux/celles qui n'en n'auraient pas.
4. Un atelier sur les **pôles de données nationaux et locaux**, et des soutiens à plus ou moins long terme. Comment gérer les données des projets pluridisciplinaires ? les disperser entre plusieurs pôles?
5. Les **données SHS**, d'enquêtes : enjeu de stockage, celles issues des portails nationaux tel que l'INSEE, etc.
6. Rédaction et partage d'une **stratégie de gestion de la donnée** pour les observatoires sur le long terme (**plan gestion de la donnée**) et les étapes pour y parvenir, les ressources nécessaires.

Arbitrage des demandes: mi-décembre 2020

Budget 2020

	Dépenses 2020 engagées (jusqu'à 30/04/21)	Dépenses 2020 à engager (jusqu'à 30/04/21)	Commentaires
1. Prestation E. Blondel - Module Geoflow	**12 000 €**	**6 000 €**	

(a) Dataverse	6 000 €		Création DOI
(b) EML / MetaCat	6 000 €		Fiche EML pour PNDB/GBIF
(c) Interface Shiny		3 000 €	Masquer R / Json config.
(d) nouveau module		3 000 €	
2. Soutien Hackaton-II RoZA (+ ZABR ZAA ZAL ZABri ZAJJ)	0 €		Annulation en raison du confinement
3. AGBED 2020	0 €		En visio vs. rencontres
4. Prestation géocatalogue		**7 000 €**	Marché à venir fin nov. / ~10 jours de presta
TOTAL	**12 000** €	**13 000** €	
Budget 2020 restant	13 000 €	0 €	

Veille

Les **journées SIST 2020** : atelier geoflow programmé. Reporté à 2021? https://sist20.sciencesconf.org/

Budget 2021

- 2 Réunions mai-juin puis sept-oct pour E-ENVIR 2021 : 3500 €
- AGBED 2021 3500 €
- Missions networking - enveloppe de 2000 €
- 1 atelier geoflow approche utilisateurs réparti par quart de France (1 x 4 ateliers ou 4 en simultanée) - 10 pers x 4 - 500 € / atelier soit 2 000 €
- Prestations : modules complémentaires pour geoflow, traduction pour les fiches de données, compilation des données aux formats choisis - enveloppe de 14 k€

Recherche de financements complémentaires (aap).

➤Total demandé : 25 k€

Remarques : suivant la réponse de l'INEE/INSU pour la réalisation de l'Ecole Thématique E-ENVIR 2021 (réponse attendue pour mi-décembre), une partie de cet argent pourra être mobilisée.

Présentation au comité direction du RZA

- Geoflow
- Appel aux protocoles / voc contrôlé / SES conceptuel
- E-ENVIR information & recrutement

Annexes

Catalogue d'attributs (ISO 19110 / Dictionnaire des données (ISO 19126)

3.6.1.2 Le Catalogue d'entités (L'interopérabilité des systèmes d'information géographique Kerhervé, Jean-Christophe, p17)

Le catalogue des entités (ISO 19110) est décrit par le modèle général d'entités défini dans la couche méta-modèle. Le catalogue contient l'ensemble des définitions des types d'entité avec leurs attributs, opérations et relations du monde réel comme par exemple un pont et un lac avec leurs attributs et opérations respectifs tel que la hauteur du pont, la profondeur du lac, le fait que le pont peut se lever et s'abaisser etc.

La norme 19110 définit la méthodologie pour cataloguer les types d'entités et spécifie comment la classification des types d'entités est organisée dans le catalogue et comment elle est présentée aux utilisateurs d'un jeu de données.

Le catalogue d'entités peut se référer au dictionnaire des entités (ISO 19126) qui fournit des définitions basiques et de l'information supplémentaire sur le jeu de concepts. Un même dictionnaire peut être utilisé dans un ou plusieurs catalogues d'entités.

La norme 19110 peut être utilisée comme une base pour la définition de la modélisation du monde réel pour une application particulière ou pour standardiser des aspects généraux de la modélisation du monde réel pour plusieurs applications.

However, the power of ISO 19126:2009 is that it allows different scientific and other disciplines to define and maintain concepts relevant to their field in online registers.

Clause 6 specifies the feature concept dictionary schema, which may include definitions of feature concepts, feature attribute concepts, feature association concepts, feature operation concepts feature role concepts (link between an association and an attribute or operation) and nominal value concepts (category, class, kind or type that may be identified as an element of an enumeration or code list).