Enquête téléphonique sur les pratiques et les besoins des ZAs en termes de gestion des données (FAIRisation) – janvier à mars 2020

Résultats intermédiaires (principalement relevant du niveau 1)

Objectif de l'enquête

L'objectif de cette enquête est d'avoir un meilleur aperçu des connaissances, compétences et solutions techniques des ZAs et de ses membres de la gestion des données et des métadonnées. Cette enquête s'apparente à **une analyse des besoins**, même si elle n'est pas formulée en ces termes.

Pour mémoire, à l'issue de cette enquête, il s'agira de fournir et proposer des **formations et des documents supports ad hoc** pour avancer sur la production de données FAIR, c'est à dire:

- production de fiches de métadonnées aux standards du RZA (ISO 19115 ou compatibles; avec vocabulaires contrôlés)
- accès à des bases de données
- valorisation des jeux de données avec DOI / publication de data paper
- réutilisation des bases de données existantes

(cf. feuille de route 2020 : Géocatalogue)

Ce travail aliment également en parallèle les échanges avec OZCAR/THEIA dans le cadre du projet eLTER.

Précautions !!! L'hypothèse a été faite que les directeurs disposaient d'une vue d'ensemble complète du fonctionnement de la gestion des données et des comportements des membres de leur ZAs. Il s'avère qu'au fil des entretiens, on perçoit assez rapidement que les connaissances à l'échelle d'une ZA sont plutôt partielles et d'autant plus s'il existe un référant compétent en la matière (délégation complète). On perçoit également que les propos rapportés concernent souvent la personne interrogée. Ce faisant, les propos recueillis témoignent plutôt d'un profil type de personne convaincue de l'intérêt de la ZA mais en apprentissage ou volontaire pour apprendre dans le domaine de la gestion des données.

Préambule et vue d'ensemble

De manière générale, les directeurs ont **répondu favorablement à cette sollicitation**. Une partie d'entre eux ont **connaissance du projet et du collectif BED**, notamment avec l'appui technique de Christine et Oton, particulièrement apprécié même si les éléments transmis/échangés relevés parfois d'une technicité élevée. En revanche, la **maitrise de l'environnement des données** au sein des ZAs est très incertaine et repose pour beaucoup sur le correspondant « données » au sein de la ZA – en général, membre du collectif BED, mais pas que.

Les questions et ressentis globalement sont les suivants :

- Sur-sollicitation des chercheurs qui empêche / **réduit drastiquement le temps de recherche** (cela va au-delà des productions des fiches de métadonnées)
- Injonction qui vise tous les chercheurs mais **toutes les données ne se valent pas**, certaines peuvent être partagées de manière **brutes**, d'autres nécessitent d'être **contextualisées**. Important de cibler dans le discours et dans les stratégies de FAIRisation les données à retenir, la recherche concernée (ex. suivis long terme). Notamment, **quelles données partagées ?**
    - Issues de modélisation (ex. ZAS) ?
    - Issues de plusieurs BDD (ex. ZATA) ?
    - Celles ayant une p-value mais laquelle ? Risque de multiplier les fiches MD (ex. ELVIS – OSR)
- Craintes d'un **partage des données avant leur publication** ; trois stades de données (1) intéressante mais insuffisante pour les publier (2) intéressante et en cours de publication (3) intéressante et publiées.
- Identifier la « carotte » pour les chercheurs car ce n'est pas un travail valorisant et **le fonctionnement en huit clos**, en sachant qui a la donnée, fonctionne plutôt bien! (ex. ZAHW ; ZAS). Toutefois, la doi-isation (jeux de données et/ou data paper) séduit les chercheurs.

I. Gestion des métadonnées

1. Géocatalogue

En pratique peu de personnes parmi les interrogés utilisent des catalogues de données. Pour celles qui l'ont eu fait, il s'agit surtout d'avoir un aperçu du travail produit ailleurs (à l'international ou dans le cadre d'un réseau de mesures) mais en aucun cas une volonté de réutiliser la donnée produite. La communauté RZA est plutôt composée de **producteurs de données**. La **réutilisation des données** semble motivée par des axes de recherches transversaux, tel que celui de l'analyse des trajectoires des socio-écosystèmes, actuellement en cours.

Concernant les géocatalogues, on distingue **trois profils** de ZA: (1) avancé (ZABR, ZAAJ, ZAA…) (2) intermédiaire, ie. existant mais difficile à faire fonctionner sans RH (ZAEU, ZABrI…) (3) en construction avec 2 sous catégories (3a) ceux disposant de moyens / personnes compétentes (3b) ceux ne disposant pas de moyens.
La plupart des directeurs ont **connaissance du geonetwork** existant, mais étonnamment pas tous en raison d'une confiance « aveugle » envers le correspondant donnée et peu de pratiques de consultation/production de fiche de MD. On rencontre parfois des difficultés techniques ou d'organisation, notamment avec les géocatalogues associés à des partenaires académiques (ZAM) ou non académiques (ZAEU).

Parmi les géocatalogues cités, des **services ou interfaces webcarto** sont mentionnés, soit en complément du géocatalogues soit comme première étape vers le geonetwork (ex. ZAS, ZATU, ZAAr).

(cf. tableau avec le listing des geocatalogues)

2. Rédaction des métadonnées

Pour les plus anciennes ZAs pratiquant la production de fiche de métadonnées, le **fichier excel** du RZA est bien connu, mais a parfois été mis de côté, au bénéfice d'une visualisation de la fiche produite depuis l'interface du geonetwork. L'interface du géonetwork n'étant pas simple d'accès, plusieurs ZAs ont produit des tutoriels (cf. tableau).

La production de MD s'appuie sur **une animation continue**, soit sous la forme d'une **Métadonnées Party**, soit des rdv individuels. La présence de **doctorant** facilite la mise en œuvre de la production des fiches de MD. La plupart du temps les « animateurs » invitent les chercheurs à **dupliquer leur première fiche de métadonnées** (ZAAJ) ou à dupliquer celle des collègues qui s'apparentent à leur projet (ZABR).

D'autre part, les ZAs accordent un financement de projet **dans le cadre des AAP** à condition de **produire des fiches de MD**.

L'usage de R pour certaines personnes peut s'avérer inhabituel et reposerait sur une formation.

NB. L'ensemble des ZAs qui anime cette dimension « données » pensent que la diversité des outils n'est pas un obstacle à la réalisation des fiches de MD.

La qualité des métadonnées semble assurer en particulier pour les ZAs disposant d'un personnel dédié à cette tâche. Notamment, on notera les pratiques : (1) de la ZABR : **stagiaire de 6 mois**, pour pré-remplir les fiches et solliciter les chercheurs afin de les valider (2) de la ZAAJ : présence d'une **documentaliste** qui contrôle le vocabulaire contrôlé ! Leur présence assure une **homogénéité des informations fournies**. Enfin, bien que les fiches soient dans des formats normées (INSPIRE, ISO 19 115/19 139), l'interopérabilité reste difficile en raison de l'absence d'un vocabulaire contrôlé.

Le **vocabulaire contrôlé** est une notion plus ou moins connue des directeurs de ZAs. Ils parleront plus facilement des thesaurus, que des référentiels disciplinaires utilisés (seulement quatre ont été évoqués : WARMS, SANDRE, TAXREF, champ disciplinaire OST).

Le **choix de la langue des fiches produites** a été partiellement réglé pour certaines ZAs, ie. en choisissant l'anglais pour la fiche. Reste le problème que pour les partenaires non académiques nationaux, la langue française est importante voire nécessaire. Ainsi, certaines ZAs doublent le texte en FR et en EN. Toutefois, il existe **une option multilingue sur géonetwork**, qui permet de renseigner en FR et en EN le texte libre (les mots clefs, s'ils appartiennent à un thesaurus sont automatiquement traduit), sur une même fiche. La mise en pratique au sein de la ZAAr met en évidence un doublon dans l'affichage des mots clefs – problème technique qui reste à régler.

Concernant DEIMS, pour ceux qui l'ont pratiqué – peu de personnes, la liste des variables sur deux niveaux est difficile à s'approprier, et des problèmes pour les shape ont été mentionnés.
NB. L'ensemble des codes d'accès ont été récupérés.

II. Gestion des données

Le stockage des données représente un réel enjeu pour la communauté, notamment l'**absence d'espace de stockage centralisé**. Beaucoup de directeurs ont mentionné ce point comme une priorité. En revanche, on perçoit bien que le partage des données n'est pas une évidence et les directeurs mentionnent souvent de la réticence de la part des collègues (« vieux modèles » de chercheurs).

Souvent les directeurs précisent que la nature des données est très hétérogène, et que les données sont plutôt compilées dans des **jeux de données sur fichiers excels** plutôt que des bases de données relationnelles.

Sans surprise la production de **Plan Gestion de la Donnée**, s'il est connu (peu de personnes le pratique ou en ont entendu parlé) reste entachée de difficultés. Le témoignage de certaines ZAs telles que la ZABR et la ZAA, pourrait être encourageant pour avancer sur ce champ là...

Les **pôles de données** appartiennent à une nébuleuse pour la plupart des directeurs. Alors que certains souhaitent être mieux informer sur leurs missions, fonctions, etc. d'autres soutiennent plutôt que les interlocuteurs à privilégier sont les **centres de données locaux / régionaux** ; la proximité avec la donnée étant particulièrement importante (efficacité du transfert, assurer la qualité).

La carotte dans cette « entreprise FAIR » est clairement la capacité d'afficher sa paternité pour les jeux de données. Ainsi, la **DOI-isation** est plutôt bien perçue voire en cours de mise en œuvre (auprès des OSU et/ou INIST, discussion pour produire des DOI) ou déjà possible.

La DOI-isation se porte même sur de la littérature grise afin de recenser les données historiques (ZAS).

Les **données SHS** sont plutôt mal connues des directeurs des ZAs, notamment pour les aspects de stockage et d'application du RGPD. Dans le cadre de l'application du RGPD, des zones d'ombres subsistent en particulier pour le partage entre labos, et sur les enquêtes courtes (à la volée à l'extérieur vs. en salle). Ainsi, bien que le RGPD semble plutôt bien appliqué pour les personnes/ZAs des domaines SHS, un partage d'information et des formations sur le sujet sont attendus. La ZAAr et la ZAEU semblent actives sur le sujet.

III. Site internet ZA

Une grande majorité des ZAs partage des **informations sur la gestion des données sur leur site internet**, et l'ensemble des ZAs sont d'accord pour créer une page si elle n'existe pas ou enrichir le contenu existant, notamment en indiquant le lien vers la page du RZA-BED.

Dans certains cas, l'information existe mais n'est pas facile d'accès. Un **template pour guider l'architecture des site internet ZAs** a donc été demandé.

Concernant la **gestion des sites internet**, les situations diffèrent : parfois, une personne de la ZA a accès au site pour faire les modifications, parfois elle passe par un service de communication ou un informaticien de l'université…

IV. Réseau BED

Comme indiqué en préambule, le réseau BED et ses actions sont plus ou moins connus. Une **information régulière dans la newsletter** peut suffire, ou via l'intermédiaire du correspondant.

Concernant les formations, la tendance – faute de temps – est clairement d'envoyer le correspondant se former et redistribuer les informations.

Parfois, des personnes ont été suggérées pour intégrer la liste de BED.

V. Conclusions partielles

Cette enquête confirme la vision déjà partagée au sein du collectif BED, mais permet de mieux **évaluer les priorités** (stockage vs. production de métadonnées sans carotte) pour mettre en place une dynamique au sein du RZA.

Dans l'ensemble, les **réticences sont faibles** et on peut percevoir des **changements structurels ou un changement de direction/renouvellement de dossier favorables** (ex. IRL pour ZAHW ; ZATU, ZAEU).

A partir de ces éléments et de ces trois profils, on peut imaginer différents niveaux d'interactions pour BED.

(1) ZAs avancées : un travail sur les fiches de MD (vocabulaire contrôlé) et les BDD
(2) ZAs intermédiaire : la production de fiches de MD + établir une stratégie de gestion de la donnée sur la base de l'existant
(3) ZAs débutante et/ou en construction : identifier et partager les outils, logiciels pour démarrer + établir une stratégie de gestion de la donnée

Enfin, cette enquête met en évidence que des points importants n'ont pas été mentionnés (ou ponctuellement), laissant probablement la place à des pratiques incertaines :

- les **travaux et données qui relèvent d'une ZA**,
  un texte clair et précis mériterait d'être diffusé – seule la ZAPYGAR procède à une labellisation ;
- les **variables essentielles** qui caractérisent nos jeux de données et celles qui peuvent être attendues à l'échelle du RZA,
  une liste à minima des données que les ZAs disposent devraient être établies. Ce point rejoint les actions en cours du projet eLTER. L'enjeu est de taille, en particulier, pour les sites qui peuvent prétendre à devenir TNA.
  un lien est à faire avec celles déjà renseignées dans DEIMS
- les **BDD ou jeux de données** n'ont été que faiblement abordés par manque sans doute de connaissance au sein des ZAs. Un travail d'inventaire semble nécessaire (cf. discussion en cours avec THEIA/OZCAR et Data Terra dans le cadre du PIA3 ;