Bonnes pratiques en matière de gestion, de curation et d'utilisation des données

Le présent guide fournit des conseils et des exemples de bonnes pratiques en matière de gestion, de curation et d'utilisation des données dans le cadre du processus d'évaluation nationale des écosystèmes. L'utilisation de données de haute qualité par les auteurs permet de veiller à ce que l'évaluation soit basée sur des preuves solides, augmentant ainsi la crédibilité des résultats clés. Les plans de gestion, de curation et d'utilisation des données doivent être envisagés dès la phase de *cadrage* de l'évaluation afin de garantir l'exactitude, la validation et la disponibilité des données dès le début du processus. Ils doivent également mettre en évidence toute lacune dans les données qui peut être comblée au cours de l'évaluation.

Gestion des données

La gestion des données implique la planification de toutes les étapes du cycle de vie des données. Elle est mise en œuvre tout au long du processus d'évaluation. Une bonne gestion des données facilite l'utilisation efficace des données et peut améliorer l'accès aux données et leur utilisation par les développeurs et les utilisateurs de l'évaluation, accroître l'efficacité opérationnelle, réduire la probabilité d'erreurs ou de pertes de données et contribuer à la sécurité des données. De solides stratégies de gestion des données pour le stockage, l'utilisation, l'élimination et le partage des données sont essentielles pour rendre les processus d'utilisation des données sûrs et efficaces.

Un plan de gestion des données décrivant les processus convenus peut être élaboré lors de la phase de *cadrage* de l'évaluation, avant la collecte et l'analyse des données et des ressources. Ce plan peut garantir la cohérence au sein d'une organisation et réduire les coûts liés aux données tout en augmentant l'efficacité.

Tenez compte de votre budget : la gestion des données entraîne des coûts liés au stockage et à la sécurité des données, ainsi qu'au temps de travail du personnel. Ces coûts doivent être pris en compte dès le début du processus d'évaluation nationale des écosystèmes.

Considérations importantes pour un plan de gestion des données :

De solides conventions en matière de dénomination et de catalogage des fichiers peuvent être définies et mises en œuvre au sein de l'équipe d'évaluation. Cela permettra aux auteurs et aux autres membres de l'équipe de facilement trouver, utiliser et mettre à jour les données. Dans l'idéal, ces conventions doivent être à la fois conviviales et évolutives. Quelques conseils sur les bonnes pratiques en matière de dénomination de fichiers :

- Une hiérarchie de dossiers adaptée à la structure de l'évaluation ;
- Utilisez des noms de fichiers courts mais descriptifs de manière cohérente ;
- Utilisez un format cohérent pour inclure les dates dans les noms de fichiers. Le format AAAAMMJJ est suggéré ;
- Incluez un numéro de version ;
- Utilisez des majuscules incorporées et des traits de soulignement au lieu d'espaces.

La gestion des données relève de la responsabilité de chacun. Le groupe de support technique est généralement responsable de l'organisation et du stockage des données utilisées dans la réalisation de l'évaluation. Cependant, tous les auteurs sont responsables de l'approvisionnement en données et en métadonnées, et du respect du plan de gestion des données. Les rôles et les responsabilités en matière de gestion des données pourraient être clairement attribués à des personnes et à des équipes spécifiques au cours du processus de planification de l'évaluation.

Enregistrez et mettez à jour les métadonnées tout au long du cycle de vie des données. Les métadonnées sont des informations sur les données. Elles sont essentielles pour préserver la qualité des données. Elles comprennent des informations sur le contenu des données, leur structure et leur autorisation, et contribuent à garantir que les données soient correctement utilisées et réutilisées. Idéalement, les plans de gestion des données encouragent l'équipe de l'évaluation à adhérer aux principes FAIR[1] (trouvable, accessible, interopérable, réutilisable) pour le traitement des métadonnées. Un accord sur les normes relatives aux métadonnées peut également contribuer à une utilisation cohérente au sein de l'équipe de l'évaluation.

Un plan de stockage des données est nécessaire pour garantir la sécurité, l'accessibilité et la disponibilité des données pendant et après la réalisation de l'évaluation. Ce plan doit porter sur le stockage, la sauvegarde, le partage et la conservation, autant d'aspects interdépendants du stockage des données.

Stockage des données : le stockage des données peut inclure des données numériques ou non numériques. L'équipe peut déterminer à l'avance les exigences concernant les informations à conserver et celles à détruire. Les options de stockage comprennent le stockage local (ordinateurs portables, lecteurs réseau, stockage optique et dispositifs de stockage externes) et le stockage sur cloud. Le stockage sur cloud, tel que Dropbox, Sharepoint et Onedrive, a gagné en popularité parce qu'il permet de réduire les investissements en capital et l'espace de stockage physique, et offre plusieurs niveaux de sécurité des données. Les questions de confidentialité, la facilité d'accès en fonction de l'emplacement et le coût doivent être évalués lors du choix des options de stockage.

Meilleures pratiques en matière de stockage des données numériques[2] :

- Utilisez des systèmes de stockage de haute qualité ;
- Utilisez des formats non propriétaires ;
- Migrez les fichiers de données tous les deux à cinq ans vers un nouveau stockage ;
- Vérifiez régulièrement les données stockées pour vous assurer que rien n'a été perdu ;
- Utilisez différentes formes de stockage pour les même données ;
- Étiquetez et organisez logiquement les fichiers stockés ;
- Pensez au chiffrement.

Sauvegarde : les bonnes pratiques soulignent la nécessité de sauvegarder les données pendant et après les réalisations de l'évaluation (voir « Conservation » ci-dessous). Les données peuvent être perdues en raison d'une défaillance matérielle, de problèmes logiciels, d'une infection virale, d'une panne de courant, d'une erreur humaine, d'un vol ou d'une perte, ou d'un dommage matériel[3]. Une stratégie de sauvegarde idéale pourrait inclure une protection contre tous ces risques, mais l'accent pourrait être mis sur la protection contre les risques les plus probables selon le contexte spécifique. Le plan de sauvegarde pourrait préciser quelles données sont sauvegardées, à quelle fréquence et où les données sont sauvegardées, qui est responsable de l'exécution des sauvegardes et si des sauvegardes complètes ou incrémentielles sont nécessaires. Une pratique recommandée pour la sauvegarde et le stockage des données est la règle 3-2-1 qui recommande d'avoir *trois* copies des données, sur *deux* supports de stockage différents, dont l'*une* doit être stockée hors site[4]. Tester le système de sauvegarde lors de son installation initiale et à des intervalles réguliers par la suite permet de s'assurer que les fichiers sauvegardés peuvent être récupérés et lus.

Partage : les besoins en matière de partage des données sont un élément important à prendre en compte lors de l'élaboration d'un plan de stockage des données pour une évaluation nationale des écosystèmes. Un plan de stockage des données bien conçu peut simplifier le partage des données et la collaboration avec les membres de l'équipe pour une utilisation actuelle et future. De nombreuses options de stockage sur cloud servent également de services de partage de fichiers, permettant aux utilisateurs de partager et de collaborer sur des données, des documents et d'autres fichiers. Lors de l'élaboration du plan de gestion des données, il est important de déterminer qui aura besoin d'accéder aux données et à partir de quel endroit, car cela peut être déterminant pour choisir la meilleure façon de stocker et de partager les données. L'élaboration

[1] Principes FAIR https://www.go-fair.org/fair-principles/

[2] British Ecological Society. 2018. Data Management. British Ecological Society. Disponible à l'adresse suivante : https://www.britishecologicalsociety.org/wp-content/uploads/2019/06/BES-Guide-Data-Management-2019.pdf?utm source=web&utm medium=web&utm campaign=better science

[3] British Ecological Society. 2018. Data Management. British Ecological Society. Disponible à l'adresse suivante : https://www.britishecologicalsociety.org/wp-content/uploads/2019/06/BES-Guide-Data-Management-2019.pdf?utm source=web&utm medium=web&utm campaign=better science

[4] Preston, W.C. 2020. For secure data backup, here's how to do the 3-2-1 rule right. Disponible à l'adresse suivante : https://www.networkworld.com/article/3527303/for-secure-data-backup-here-s-how-to-do-the-3-2-1-rule-right.html

d'accords de partage de données et de restrictions potentielles sur l'utilisation des données constitue une étape essentielle, avant qu'ils ne soient rendus publics. Si les données doivent être partagées avant la publication de l'évaluation, des documents de confidentialité et d'autorisation devront être signés. Cette démarche est clé pour garantir la protection des données et de leurs sources pour l'évaluation.

Conservation : un plan de conservation des données est essentiel pour garantir l'utilisation des données au-delà de la réalisation de l'évaluation. Il implique l'archivage des données afin de garantir leur viabilité et leur accessibilité futures. Un plan indiquant comment et où les données seront archivées peut être élaboré au cours du processus de planification et être inclus dans le plan de gestion des données. Les données archivées pourraient comprendre des fichiers de données accompagnés des noms de fichiers descriptifs dans une structure de fichiers bien organisée, des métadonnées solides, ainsi que la documentation et des fichiers supplémentaires qui aideront les futurs utilisateurs à utiliser et à comprendre les données[5]. Une autre solution consiste à déposer les données dans une archive ou dans un dépôt de données dès l'achèvement de l'évaluation afin de les conserver et de faciliter leur utilisation dans le cadre de l'élaboration des politiques, de recherches ou d'autres activités futures.

La sécurité et la protection des données sont essentielles pour garantir la confidentialité, la disponibilité et l'intégrité des données. Une infrastructure de sécurité solide pourrait être planifiée et mise en œuvre pour réduire le risque de violation des données. Par exemple, des logiciels antivirus et anti-espions pourraient être installés sur tous les postes de travail. De même, la conformité aux réglementations en matière de confidentialité pourrait être essentielle et devra être prise en compte dès la phase de cadrage de l'évaluation.

Identification des sources de données, acquisition et curation des données

Identification des besoins en données
Avant de collecter des données, il est important de considérer les questions politiques clés qui seront axées sur le besoin existant d'informations relatives à l'état et aux tendances de la biodiversité et des services écosystémiques, qui rendent compte de leurs contributions au bien-être humain dans la prise de décision à travers les secteurs. Afin d'identifier la nécessité d'une évaluation, il est important de comprendre le contexte environnemental, social et économique national actuel, par exemple à travers le cadre conceptuel et les défis connexes auxquels sont confrontés les décideurs. L'identification de ces lacunes lors de la phase de cadrage permet d'orienter les auteurs sur les besoins spécifiques en données de l'évaluation.

Qualité et validation des données
Avant de collecter des données, il est également important de considérer le niveau acceptable de qualité des données. L'utilisation de données de haute qualité permet de s'assurer que l'évaluation est basée sur des preuves solides, augmentant ainsi la crédibilité des résultats clés. Des indicateurs de qualité des données peuvent être utilisés pour déterminer si la qualité d'un ensemble de données est acceptable. Les indicateurs clés de qualité des données à prendre en compte sont les conditions d'utilisation des données, l'accessibilité des données, la fréquence des mises à jour, la couverture géographique des données, l'échelle des données, l'inclusion de la méthodologie de collecte des données et des métadonnées, ainsi que l'authenticité du fournisseur de données.

Conseils essentiels pour la collecte et la curation des données

- Conservez une trace de toutes les sources de données, y compris les coordonnées des différents fournisseurs de données et les sites web d'où proviennent des ensembles de données accessibles au public. Gardez à l'esprit que les données provenant des fournisseurs de données peuvent nécessiter des accords formels de partage de données.
- Discutez de l'utilisation que vous comptez faire des données avec les fournisseurs de données. Cela permettra de clarifier les droits dont vous disposez pour utiliser les données.
- Utilisez des identifiants uniques pour chaque point de données dans un ensemble de données. Cela

[5] Axiom Data Science. 2017. Metadata and Documentation. Disponible à l'adresse suivante : https://www.axiomdatascience.com/best- practices/MetadataandDocumentation.html#metadata-metadata

permettra de suivre les données au fil du temps, même si d'autres informations descriptives concernant le point de données changent.

- Gérez les versions. Si un ensemble de données est mis à jour ou modifié au fil du temps, chaque nouvelle version peut être enregistrée dans un nouveau fichier avec la même convention de nom et le même horodatage.
- Il est recommandé d'utiliser un système de coordonnées cohérent lorsqu'il s'agit de données spatiales.
- Les titres de champs et des valeurs cohérents peuvent être utilisés lors de la mise à jour des ensembles de données.

Utilisation des données de manière efficace

Tous les aspects de la gestion des données conduisent à leur utilisation et à leur réutilisation. Une fois les sources de données identifiées, il est important de réfléchir à la manière d'utiliser efficacement les données dans les analyses.

Conseils des auteurs pour une utilisation efficace des données

- Prenez votre temps. Il peut être utile de suivre la règle 70-20-10 utilisée dans le cadre du traitement des données. Vous devez consacrer 70 % de votre temps à la recherche et au catalogage des données, 20 % au nettoyage et à la préparation des données et 10 % aux analyses proprement dites.
- L'organisation est capitale et le fait d'avoir un plan pour chaque étape du processus peut aider à rester concentré. Parmi les conseils, figure un système de dénomination des produits et la structuration des produits. Vous pouvez également prendre des notes détaillées pour chaque étape des analyses.
- Assurez la traçabilité de votre travail. Cette démarche sera utile pour l'utilisation future des informations et des données collectées dans le cadre de la réalisation de l'évaluation. Github est une plateforme puissante et utile pour le partage du code d'analyse avec des partenaires internes et/ou externes.
- Il est crucial de respecter les conditions d'utilisation des ensembles de données et de consigner ces informations dans les métadonnées.
- Les accords d'utilisation des données, tels que les droits de propriété intellectuelle, les informations liées à l'autorisation et les licences doivent être lus et appliqués. Il est recommandé d'utiliser un logiciel de gestion des références pour citer les sources.