Politique de Gestion et d'Ouverture des Données Scientifiques de l'UMR EcoFoG

1. Introduction

Les données scientifiques produites par l'UMR EcoFoG couvrent de nombreux domaines thématiques (inventaires de biodiversité, descripteurs environnementaux, traits écophysiologiques, sciences des matériaux, chimie, génomique, flux de gaz…), tous en lien avec la recherche en écologie. Ces données servent de carburant aux différents projets scientifiques ponctuels ou de longs termes de l'unité et de ses partenaires. Pour optimiser ce rôle, elles doivent être gérées le plus rigoureusement possible afin d'être préservées, disponibles et intelligibles sans ambiguïté. Ces aspects sont synthétisés dans l'acronyme FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable). La FAIRisation des données est l'axe central de cette politique et permet de préparer et d'optimiser, quand c'est approprié, leur publication.
Cette politique s'applique au patrimoine de données anciennement acquises (avant la mise en place d'une politique d'unité sur les données scientifiques), ainsi qu'à la production de nouvelles données en anticipant tous les aspects et les besoins (pratiques, informatiques, infrastructurels, juridiques…).
Quelles que soient leur origine et leur nature, les données doivent donc être centralisées, contrôlées, standardisées, documentées, pérennisées et ouvertes autant que possible, en accord avec l'application du cadre juridique approprié (français, européen, autres) et les politiques d'établissements des différentes tutelles de l'unité qui en possèdent une ou de leurs recommandations (AgroParisTech, Cirad, CNRS, INRAe).
L'agrégation progressive et structurée de toutes les données produites par l'unité s'inscrit dans la philosophie globale de cette politique qui est de parvenir à interconnecter le maximum de données dans un système d'information thématiquement transversal, le principal lien inter-thématique étant le géoréférencement.
La présente politique est établie par la direction de l'unité et le.a responsable de la gestion des données scientifiques de l'unité (ci-après nommé.e « RGDS ») et doit être suivie et appliquée par tous les agents de l'unité. Ils sont accompagnés et formés pour cela par le.a RGDS.

2. Gestion du patrimoine de données

2.1 Centralisation des données anciennes et nouvellement produites

Thématique par thématique, les données produites depuis de nombreuses années, et celles que l'on produit actuellement ou durant les projets à venir, doivent être agrégées et bancarisées afin de les sécuriser. Il s'agit également d'éviter toute perte et toute duplication incontrôlée menant à la multiplication périlleuse des versions. Pour certaines thématiques, ce travail a été initié dès

les premières données créées ; pour d'autres, les données sont dispersées sur plusieurs supports parfois pérennes, parfois non. Il s'agit donc de :

- Faire perpétuellement évoluer les dispositifs existants vers des pratiques vertueuses en termes de sécurité, de fiabilité et de pérennité
- Structurer les thématiques pour lesquelles ce n'est pas encore le cas afin de rassembler toutes les sources de données en utilisant ou développant des procédures adaptées et les réceptacles les plus appropriés.

En pratique : Les producteurs de données de l'unité doivent donc s'appliquer à signaler et transmettre toute source de données ancienne, nouvelle ou future à le.a RGDS afin de viser à leur intégration dans le système d'information transversal et à leur adéquation avec les outils existants ou au développement de nouveaux outils de traitement, de stockage et d'ouverture.

2.2 Curation et enrichissement

Au fur et à mesure de leur agrégation, les données sont explorées, contrôlées et éventuellement corrigées pour assurer :

- Leur homogénéité : mise en adéquation avec des standards (unités, langue…) et des référentiels communs (taxonomiques, géographiques, ontologiques) et régulièrement mis à jour
- Leur complétude : est-ce que tout est renseigné ? les vides sont-ils justifiés ? les valeurs nulles sont-elles autorisées ?
- Leur qualité : précision, exactitude, cohérence (contrôles statistiques et graphiques simples, respect des limites/bornes).

Autant que possible, toutes les données sont enrichies du maximum d'informations pertinentes (géoréférencement, informations nouvellement acquises et potentiellement rétroactives, descriptions environnementales).

En pratique : Ces procédures de contrôle et d'enrichissement sont assurées, en plus ou moins grande partie, voire en totalité, par le.a RGDS, en relation avec les responsables scientifiques de chaque thématique et les producteurs de données (qui sont souvent la même personne) qui sont sollicités pour pré-formater leurs données selon des modèles établis au préalable avec le.a RGDS. L'inspection première des données et le maximum de contrôles doivent cependant être réalisés sur les données avant leur transmission à le.a RGDS.

2.3 Documentation des données

Pour chaque jeu de données, des métadonnées les plus riches possibles doivent être renseignées (origine, genèse, protocole…), des dictionnaires de données décrivant chaque variable (signification, type, unité, origine…) créés et des Plan de Gestion de Données (PGD) établis. Les PGD définissent le contenu des jeux de données, leur origine, leur genèse (auteur, contexte, financement), les précautions à prendre (juridiques, étiques, techniques…) mais surtout la stratégie de gestion qui leur est appliquée (organisation des fichiers, convention de nommage des fichiers et des données, stockage, sauvegarde, partage, publication et le ou les responsable(s) de chaque partie de la stratégie). Des PGD doivent aussi être produits pour un maximum de projets scientifiques, y compris les thèses, et peuvent s'appuyer sur les PGD des jeux de données transversaux.

Un PGD global de l'unité est rédigé par le.a RGDS et régulièrement mis à jour en accompagnement de la présente politique. Il peut servir de ressource pour la rédaction des différents PGD de projet ou de thématique.

En pratique : Cette documentation est supervisée par le.a RGDS et tout est rédigé en collaboration avec les responsables scientifiques des différentes thématiques ou projet. Il est très important que des réflexions soient menées en amont de chaque projet afin de définir l'organisation des données produites, les conventions de nommage et les modes de partage des données durant le déroulement du projet et que ces règles soient appliquées par tous les acteurs du projet. Le.a RGDS est là pour conseiller et accompagner les porteurs de projets dans ces réflexions et la définition de ces règles.

2.4 Application du cadre légal

Chaque producteur de données se doit de s'assurer du respect du cadre légal lors de la manipulation ou de la production de données, notamment en cas de prélèvement de matériel biologique (APA), de manipulation de données personnelles (RGPD) ou de données sensibles.

En pratique : Les agents confrontés à ce genre de données doivent contacter le.a RGDS afin d'être accompagnés et conseillés dans les démarches nécessaires.

3. Fonctionnement technique, infrastructures de données

3.1 Création et gestion des bases de données

Une fois la curation des données faite par le.a RGDS, et validée par les responsables scientifiques de chaque thématique et de chaque projet, elles sont intégrées dans des bases de

données conçues spécifiquement pour chaque domaine. Les schémas de chaque base sont réfléchis en fonction de la teneur des données et de leur finalité. Les bases de données restent évolutives et perfectionnables à tout moment selon l'avancée des projets, l'évolution des besoins et l'arrivée de nouvelles données à agréger, ainsi que de nouvelles variables.

Ces bases de données sont créées avec Microsoft SQL Server, logiciel historiquement utilisé dans l'unité, sur le serveur dédié de l'unité. Une migration vers PostGreSQL, logiciel beaucoup plus répandu, ouvert, gratuit, tout aussi puissant et intégrant une gestion à part entière des données géographiques, est envisagée à moyen terme, en toute transparence pour les utilisateurs.

En pratique : Le.a RGDS gère la conception, l'alimentation et l'évolution des bases de données tout en gardant à l'esprit la nécessité de pouvoir les interconnecter. La maintenance du serveur de base de données est assurée conjointement par le responsable informatique de l'unité et le.a RGDS.

3.2 Stockage

Les données d'écologie étant majoritairement assez légères, notre capacité de stockage, pour ces données-là, n'est pas limitante actuellement et ne nécessite pas d'investissements conséquents. En revanche, pour les données lourdes, telles que la télédétection, la génomique et tout projet impliquant la manipulation ou la création de quantités de données importantes (au-delà de 50Go), des serveurs dédiés existent dans l'unité et sont éventuellement ponctuellement renforcés par des dispositifs appropriés.

Tout nouveau projet ou toute nouvelle acquisition de données lourdes doivent absolument être signalés le plus en amont possible afin de prévoir une stratégie de stockage appropriée, voire externalisée si besoin.

Par ailleurs, chacun doit prendre garde à ne pas accumuler inutilement des copies de données afin d'éviter d'une part, leur circulation incontrôlée et, d'autre part, l'encombrement abusif et écologiquement dommageable de nos dispositifs de stockage.

En pratique : Les besoins de stockage doivent être anticipés et le.a RGDS et le responsable informatique de l'unité doivent être avertis et impliqués dans les réflexions le plus en amont possible de l'arrivée des données.

3.3 Sauvegardes

3.3.1 Locales

La machine virtuelle tenant lieu de serveur de bases de données est sauvegardée automatiquement deux fois par semaine. Ces sauvegardes sont gérées par le responsable informatique de l'unité.

Par ailleurs, si des modifications majeures sont faites sur une base de données, une sauvegarde ponctuelle et manuelle est faite avant et après modifications par le.a RGDS. La sauvegarde antérieure aux modifications est conservée sur une autre instance du campus avec une date et un nom explicite.

3.3.2 Externes

Des sauvegardes automatiques des bases de données sont faites sur un serveur de la DSI du Cirad, à Montpellier :

- Toutes les semaines pour les bases « actives » (i.e. dans lesquelles il y a régulièrement des ajouts ou modifications de données)
- Tous les 6 mois, pour les bases « inactives »
- Chaque année, une copie de toutes les bases est faite et conservée sur ce serveur, années après années.

Des copies de sauvegardes des bases sont également faites régulièrement et stockées de façon sécurisées en Guyane mais en dehors du campus agronomique.

4. Accès aux données pour les agents de l'unité

4.1 Gestion des rôles et des droits

Les seules versions des données faisant foi sont celles contenues dans les bases de données. Afin de ne jamais compromettre ces données, la gestion des droits d'accès et d'actions sur ces bases est donc primordiale. Des groupes d'utilisateurs sont créés par thématique et ces groupes se voient attribuer des droits plus ou moins restreints sur les données selon leurs rôles. Sauf exception, personne ne peut modifier les données hormis le.a RGDS.

En pratique : Le.a RGDS gère la constitution des groupes et l'attribution des droits en accord avec les responsables scientifiques de chaque jeu de données.

4.2 Interfaces de visualisation et d'accès simplifié aux données

Pour plusieurs des bases de données, des interfaces Access ont été développées pour répondre aux attentes des responsables de chaque thématique et aux besoins des utilisateurs. Ces

interfaces, évolutives selon les demandes, permettent essentiellement la visualisation, la combinaison de données par requête et leur extraction de façon simple et rapide. Plusieurs interfaces distinctes peuvent être créées pour une même source de données selon les demandes.
En pratique : Le.a RGDS développe et fait évoluer ces interfaces selon les demandes. Leur fonctionnement repose sur les droits octroyés à l'utilisateur (4.1).

4.3 Scripts d'accès aux données pour analyses avancées

Afin de pouvoir manipuler et analyser directement les données avec les logiciels les plus classiquement utilisés dans l'unité, très majoritairement le logiciel R, des routines d'extraction sont développées. Elles peuvent être dimensionnées pour que les sorties correspondent à des packages d'analyse ciblés. Ces développements concernent chaque thématique mais peuvent également combiner plusieurs thématiques scientifiques croisées et inclure des procédures de SIG.
En pratique : Le.a RGDS développe ces scripts et les ponts de connexions nécessaires entre les logiciels d'analyses et le système de gestion de bases de données. Tous les scripts développés sont centralisés et documentés et beaucoup sont disponibles sur le GitHub de l'unité. Tous les scripts pérennes développés par les agents de l'unité doivent être déposés dans des forges logicielles et documentés. Des forges institutionnelles sont en cours de développement (INRAe, CIRAD). Le.a RGDS est en charge de communiquer les avancées sur ces points aux agents de l'unité.

4.4 Demandes individuelles

Pour toutes les personnes n'utilisant pas couramment R, ne connaissant pas l'utilisation d'Access, n'étant pas sur le campus de Kourou donc n'ayant pas d'accès autorisé au serveur de bases de données, ou pour toutes les requêtes de données plus complexes et/ou transversales que ce que permettent les différents modes d'accès cités ci-dessus, les demandes sont à adresser à le.a RGDS. L'autorisation d'accéder aux données est demandée au responsable scientifique du projet si le demandeur ne fait pas déjà partie d'un groupe autorisé, puis une extraction sur mesure est faite et envoyée au demandeur avec toutes les informations nécessaires à la compréhension des données (2.3) ainsi que les conditions de leur utilisation définies par le responsable scientifique.

5. Ouverture des données

5.1 Plateforme d'accès aux données de Paracou

Pour les données d'inventaires forestiers du site de Paracou (données les plus sollicitées actuellement), une plateforme en ligne permettant leur mise à disposition sur demande scientifiquement motivée a été intégrée au site internet de Paracou. Ce site intègre aussi un module de diffusion des données SIG de Paracou.

Depuis 2023, l'utilisation de la plateforme de données s'amoindrie grâce au développement de l'entrepôt de données EcoFoG (5.3) mais reste en ligne car elle permet des requêtes plus fines et plus ciblées.

A partir de 2025, l'utilisation de la plateforme SIG devrait être abandonnée au profit d'une solution alternative beaucoup plus robuste et pérenne : l'IDG (infrastructure de données géographiques) Cirad sur laquelle nos données SIG vont peu à peu être publiées (avec métadonnées, identifiants pérennes, licence de diffusion). Cette IDG sera pleinement opérationnelle courant 2025 normalement.

5.2 Publication de données dans des entrepôts thématiques

En accord avec les politiques des établissements tutelles de l'unité, une partie des données de l'unité est publiée dans des entrepôts de données thématiques afin de leur donner un affichage ciblé. Ce genre de publication doit être privilégié. Les publications déjà existantes sont régulièrement mises à jour et enrichies, le cas échéant.

En pratique : Lorsque des données sont prêtes à être publiées, les responsables du jeu de données doivent prendre contact avec le.a RGDS afin d'étudier les différentes options de publication et d'être accompagnés dans la démarche.

5.3 Entrepôt de données EcoFoG

Afin d'y publier les données n'ayant pas d'entrepôt thématique approprié, un entrepôt de données EcoFoG, administré par le.a RGDS, a été créé. L'entrepôt institutionnel du Cirad a été choisi pour accueillir cet espace car, d'une part, les fonctionnalités et la sécurité sont sensiblement les mêmes d'un espace institutionnel à l'autre et, d'autre part, la majeure partie, en terme de volume, de nos données est constituée des données de Paracou, qui est un dispositif du Cirad. Par ailleurs, depuis juillet 2022, tous les entrepôts institutionnels français sont rassemblés dans Recherche Data Gouv où toutes les données publiées vont progressivement apparaitre au fur et à mesure des moissonnages.

L'entrepôt EcoFoG a été conçu en arborescence avec des branches thématiques (inventaires forestiers, arthropodes, sciences du bois, traits écophysiologiques…) chacune subdivisée selon ses particularités. Des espaces dédiés à des projets de recherche (exemple) sont créés à la demande, permettant un partage temporaire des données entre partenaires, l'accès aux reviewers des articles scientifiques, puis leur éventuelle publication.

Chaque branche contient les données, les métadonnées, les dictionnaires de données, les PGD et tous les documents et références nécessaires à la bonne intelligibilité et réutilisation des données (scripts d'analyse, protocole, projet, articles…).

Les conditions d'accès aux données sont réglées fichier par fichier donc les données peuvent être publiées avec différents degrés de visibilité. Elles peuvent n'être accessibles que sur requête scientifiquement motivée, rester sous embargo ou être ouvertes librement mais, quel que soit le choix, elles sont toujours accompagnées d'une licence de diffusion définissant leurs conditions de réutilisation (traçabilité, citation des sources, utilisation commerciale autorisée ou non).

En pratique : Le.a RGDS administre l'entrepôt d'unité et attribue des droits d'administration, branche par branche, aux responsables scientifiques appropriés. Tout développement d'un nouvel espace thématique peut se faire en contactant le.a RGDS qui se chargera de concevoir l'organisation de cet espace, son alimentation et la gestion des droits d'accès avec le responsable thématique.

5.4 Publication de DataPapers

Les données publiées doivent être, autant que possible, accompagnées de DataPapers (articles de données). Plusieurs ont déjà été publiés par l'unité.

En pratique : Les responsables de jeu de données peuvent demander le concours de le.a RGDS pour choisir le DataJournal et rédiger le DataPaper. Le.a RGDS doit, a minima, être relecteur de l'article afin de garantir l'adéquation avec les modes de gestion de données appliqués dans l'unité.

6. Formation des agents et diffusion des informations et des ressources

Pour que cette politique d'unité soit appliquée par tous, le plus aisément possible tout au long du cycle de vie des données, il est nécessaire que i) chacun soit formé dans la mesure de ses besoins, ii) ait accès aux ressources d'information nécessaires à la gestion quotidienne de ses données, iii) ait accès aux documents de gestion propres à l'unité.

Le.a RGDS diffuse donc, tout au long de l'année, les informations qu'il.elle juge utiles autour de la gestion et de l'ouverture des données, quelles que soient leurs origines, ainsi que les offres de formation de différents organismes, tutelles de l'unité ou non.

Le.a RGDS propose, deux fois par an, une journée de formation en webinaire (une session en français à l'automne, une en anglais en juin) aux agents et doctorants de l'unité. Les dates de ces formations sont diffusées dès qu'elles sont connues afin que chacun puisse s'organiser pour les suivre. Des formations à la demande peuvent être organisées.

Enfin, une page du site internet de l'unité est dédiée à la thématique des données scientifiques. Une synthèse des données de l'unité disponible en interne, et éventuellement publiées, y est présentée ainsi que des liens vers les documents de gestion des données de l'unité (PGD, politique d'unité) et vers des ressources de formations.

Par ailleurs, le.a RGDS est disponible à tout moment pour des accompagnements personnalisés, individuels ou en groupe (au moment du montage ou du lancement d'un projet par exemple ainsi qu'au démarrage de stages ou de thèses).