Politique des données AnaEE France

Table des matières

Ce document définit la politique des données acquises à partir des services de l'infrastructure nationale AnaEE-France et leur modalité de diffusion. Nous rappelons que ces services couvrent l'accès aux plateformes expérimentales, la production d'analyses pour caractériser les écosystèmes, la mise à disposition d'instruments et l'accès aux plateformes de modélisation. Ce document ne traite pas des données générées par des modèles utilisant les plateformes de modélisation qui font l'objet d'un document spécifique.

Ce document constitue un cadre définissant les **principes d'utilisation et de diffusion des données** qui devra trouver par la suite une traduction juridique dans les documents appropriés (licences, chartes) et s'appuiera sur des outils informatiques de mise à disposition des données. Ce document met à jour les principes énoncés dans l'accord de consortium AnaEE-France. Ce document devra évoluer pour permettre un alignement avec la politique qui sera décidée au niveau Européen tout en restant conforme à la législation Française et aux politiques institutionnelles (CNRS, INRA et autres tutelles des plateformes AnaEE-France).

1 METADONNEES CARACTERISANT LES PROJETS D'EXPERIMENTATION

Toutes les demandes d'utilisation des services font l'objet d'une rédaction de projet définissant la conception de l'expérience et les ressources à mobiliser. Le projet soumis via une interface de soumission de projet permet de collecter des métadonnées qui seront visibles, sans restriction, sur le site AnaEE-FR, dès la fin du projet. Les métadonnées « projet » comportent :

- Titre et acronyme*
- Les services AnaEE-France mobilisés*
- Le projet de recherche (utile pour associer des projets d'expérimentation mobilisant plusieurs services)
- Un descriptif résumé du projet : objectifs, conception de l'expérience ou des expériences réalisées, résultats attendus
- Un résumé grand public (optionnel) *
- Le type d'écosystème
- Les variables mesurées
- Les facteurs manipulés
- Dates d'exécution sur la plateforme et d'exécution du projet
- Responsable du projet* (attention à bien respecter les clauses RGPD en ajoutant une note de clauses standard et un lien à cliquer par l'utilisateur)
- Liens vers les publications (à actualiser au fil de l'eau)

* champs diffusés dès l'acceptation du projet sous réserve d'accord du responsable du projet.

2 PERIMETRE D'UN JEU DE DONNEES ACQUIS SUR UNE PLATEFORME EXPERIMENTALE

2.1 Le cadre général

Un jeu de données produit par une expérience sur une plateforme expérimentale AnaEE-France est composé de données produites par les plateformes et de données produites par les utilisateurs (DU). Dans les données produites par les plateformes, on distingue les données de base (DB) caractérisant l'environnement du site

(météo, sol …) et les données spécifiques à l'expérience collectées par la plateforme (DP) pour le compte de l'utilisateur (mesure de gaz dans une cellule, température d'un bassin …). **Le jeu de données expérimental est alors constitué de l'union des DB (pertinentes pour l'expérience), DP et DU.**

Dans leur cycle de vie, les données font l'objet d'un certain nombre de transformations qu'il convient de bien préciser dans le jeu de données. Ainsi de manière générique on peut préciser plusieurs niveaux dans le traitement de la donnée comme dans d'autres infrastructures du domaine de l'environnement:

- L0 : données brutes (tensions délivrées par un capteur, compte numérique, proxy). Cette étape peut contenir des filtrages standards.
- L1 : données mise à l'échelle dans l'unité des grandeurs cibles (exemple : m/s pour un vent) en utilisant les procédures constructeur. L'objet de ce niveau est de produire rapidement des valeurs ayant un sens physique, chimique ou biologique. Cette étape peut contenir des filtrages standards (type gamme de mesure).
- L2 : données calibrées et filtrées. Les données de niveaux L2 intègrent l'ensemble des procédures de filtrage et d'étalonnage mis en œuvre sur les données délivrées (filtrage automatique et manuel, étalonnage externe, correction des facteurs externes influençant la mesure). Il n'y a pas à ce stade de reconstitution de mesure. La donnée est accompagnée d'un indice de qualité qui peut être différent d'une donnée à l'autre (par exemple, covariables pas toujours disponibles, options de traitement variables, confiance en l'étalonnage, bruit sur la mesure ….).
- L3 : produits élaborés pour faire intervenir un ensemble de capteurs et de modèles pour arriver à un résultat consolidé. Les traitements peuvent être, par exemple :
    - du gapfilling: on utilise une technique permettant de combler les données manquantes;
    - des agrégations de capteurs: on produit une grandeur qui tient compte de répétitions, comme pour l'évaluation d'un stock hydrique à partir de plusieurs mesures sur des capteurs indépendants,
    - assimilation de données dans un modèle : les données sont produites par un modèle dans lequel des données ont été assimilées;
    - ou reconstitution d'une donnée non directement mesurée: par exemple, les grandeurs sont issues d'un bilan.

Un jeu de données peut contenir plusieurs niveaux de traitement mentionnés ci-dessus.

Les métadonnées nécessaires à la compréhension des données et à leur exploitation doivent être associées aux données. Deux types de métadonnées sont à considérer :

- des métadonnées sur le jeu de données pour « le porter à connaissance ». Ces métadonnées ont pour vocation à alimenter les catalogues de données. Elles respecteront donc les standards internationaux afin de faciliter leur moissonnage par les catalogues nationaux et internationaux. Ces métadonnées de découverte devront renseigner la nature du jeu de données (contexte expérimental, écosystèmes étudiés, contenu du jeu de données, emprise spatiale et temporelle, contacts et droits sur l'utilisation). Ces métadonnées peuvent en partie être produites à partir des métadonnées sur les projets (voir section précédente)
- des métadonnées permettant de décrire précisément les données (métadonnées métier) en donnant pour chaque variable ou type d'observation des informations sur :
    - l'objet mesuré (quoi, où, étendue spatiale, échantillonnage, contexte expérimental (traitement)),
    - la grandeur considérée (grandeur, unité)

- les caractéristiques temporelles (représentativité temporelle de l'information)
- la provenance (qui ?, méthodes de mesure et de traitement).
- et le propriétaire

Le périmètre du jeu de donnée sera précisé dans le plan de gestion de données, document définissant l'ensemble des modalités et moyens mis en œuvre pour la gestion (depuis la collecte jusqu'à la valorisation ou la destruction) des données dans le cadre du projet et initié lors de la conception de l'expérience[1].

2.2 Les cas particuliers

2.2.1 Expérimentations à long terme

Dans les expérimentations à long terme, la collecte des données de bases (DB) par la plateforme est inhérente à sa mission. Les données produites par les utilisateurs constituent des données complémentaires qui ont vocation à être intégrées au corpus des données gérées par la plateforme, réutilisées par la plateforme et mises à disposition.

2.2.2 Plateformes analytiques et les instruments partagés

La mission des plateformes analytiques et des instruments partagés est de fournir des mesures qui ont vocation à s'intégrer dans les jeux de données des expériences auxquelles ces données sont rattachées, expériences pouvant être réalisées ou non sur des plateformes expérimentales AnaEE-France. Nous considérons donc que les jeux de données doivent alors suivre les règles proposées par les financeurs et celles des dispositifs expérimentaux associés. Par contre, la capitalisation sur les analyses et les métadonnées associées permettant de caractériser les conditions de l'expérience est essentielle pour l'amélioration du service analytique/instrumental produit (par exemple, pour produire un référentiel d'interprétation). L'accès aux données et métadonnées des utilisateurs par les responsables des plateformes de même que leur utilisation doit donc être garantis pour améliorer les mesures et leur interprétation.

3 LA LOI ET LES RECOMMANDATIONS AUX INFRASTRUCTURES

En France, l'ouverture par principe des données publiques disponibles au format électronique est posée par la loi pour une république numérique 2016-1321 du 7 octobre 2016. Cela concerne en particulier les bases de données des organismes publics de recherche et les données « dont la publication présente un intérêt économique, social, sanitaire ou environnemental. Les données produites dans le cadre de AnaEE-France sont donc pleinement concernées. De même, la loi prévoit un principe de libre réutilisation de ces données.

En particulier, « dès lors que les données issues d'une activité de recherche financée au moins pour moitié par des dotations de l'Etat, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu'elles ont été rendues publiques par le chercheur, l'établissement ou l'organisme de recherche, leur réutilisation est libre. » Enfin, toute diffusion des données doit « se fait dans un standard ouvert, aisément réutilisable et exploitable par un système de traitement automatisé ».

L'ouverture par principe des données publiques de recherche et donc leur libre réutilisation ne pourra être limitée qu'en cas d'exceptions encadrées par la loi et notamment :

- le secret en matière commerciale et industrielle et le secret professionnel ;

[1] Pour l'instant un plan de gestion de données est disponible. Il sera proposé aux responsables des projet de recherche.

- le secret de la défense nationale et des impératifs relatifs à la sécurité de l'Etat ou de l'établissement (ex. sécurité des systèmes d'information des administrations), PPST ;
- les droits de tiers ;
- les documents non encore librement communicables au regard du code du patrimoine.

Récemment la France s'est dotée d'un plan national pour la science ouverte[2] qui réaffirme les principes d'ouverture des résultats de la recherche. Les données de recherche produites par la recherche publique française doivent progressivement être structurées en conformité avec les principes FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable). La notion d'embargo n'est pas présente dans la loi qui prévoit une ouverture par défaut et sans délai. En revanche, des délais d'ouverture des données peuvent se justifier au regard de la qualité et de la fiabilité des données attendues au regard des bonnes pratiques de la communauté scientifique concernée et des besoins des utilisateurs futurs des jeux de données, ou de questions éthiques par exemple.

En parallèle, au niveau français, les organismes de recherche ont édité des chartes sur les Infrastructures de Recherche (IR) pour diffuser librement les données qu'elles produisent (par exemple, Charte des infrastructures INRA)[3]. Au niveau européen, les chartes sur les infrastructures[4] encouragent une politique d'ouverture des données et pour ce faire, la recherche d'accords entre utilisateurs des plateformes et les responsables des plateformes de recherche.

4 POLITIQUE D'OUVERTURE DES DONNEES

4.1 Principes

Les métadonnées descriptives des expérimentations sont libres d'accès dès la fin du projet. Elles sont disponibles sur le portail AnaEE-France.

Les jeux de données (DP, DU, DB) produites avec les services d'AnaEE-France et par leurs utilisateurs ont vocation à être réutilisées librement selon les principes de l'Open Data. Toute personne pourra avoir un accès libre et gratuit aux données avec pour seul engagement de citer l'origine du jeu de données et la date de dernière mise à jour. Les producteurs de données (gestionnaires des services et utilisateurs des services) s'engagent à appliquer ce principe aux données qu'ils génèrent.

Un jeu de données est ouvert lorsque les données ont été traitées, vérifiées et annotées avec les métadonnées appropriées.

4.2 Aménagement possible

L'utilisateur pourra demander au responsable de la plateforme de retarder cette ouverture pour des motifs légitimes prédéfinis par ladite plateforme, éventuellement dans son Data Management Plan (maturité vérifiabilité / traçabilité, données personnelles non encore anonymisées, raisons éthiques, …etc). Le temps nécessaire pour mener ces opérations ne pourra pas excéder 2 années à partir de la fin du projet de recherche (ci-après le « délai d'ouverture »).

[2] http://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf
[3] https://inra-dam-front-resources-cdn.brainsonic.com/ressources/afile/368466-19bb9-resource-charte-des-infrastructures-de-recherche-inra.pdf
[4] European Charter for Access to Research Infrastructures : Principles and Guidelines for Access and Related Services, doi:10.2777/524573

Pour les expérimentations à long terme, une mise à jour régulière des jeux de données est envisagée sur une base annuelle. Le délai maximum est alors compté à partir de la date de fin de la période de mise à jour (par exemple si la période est l'année 2015, le délai est compté à partir du 31 décembre de cette année).

Pendant le délai d'ouverture, le jeu de données est uniquement accessible aux personnes associées au projet de recherche. Passé ce délai, le jeu de données contenant l'ensemble des données et métadonnées sera publié et ainsi faire l'objet d'un DOI. Le jeu est rendu accessible sur un entrepôt de données ouvert.

Les publications scientifiques ne se substituent pas au jeu de données publié. En effet, les publications ne traitent en général que d'un sous ensemble du jeu de données et ne fournissent par ailleurs pas toujours l'intégralité des informations nécessaires à la réutilisation des données.

Si un utilisateur des services est défaillant dans la diffusion des données, notamment en l'absence de restitution du jeu de données dans le délai maximum imparti, la plateforme pourra diffuser les données acquises pour le compte de l'utilisateur.

Dans le cas des plateformes analytiques et des instruments partagés (voir section ci-dessus), les responsables de ces dispositifs pourront bénéficier, quelle que soit l'origine du financement (fonds privés, fond publics), des données collectées et des métadonnées associées pour des fins d'amélioration du service rendu (enrichissement d'un référentiel par exemple).

Remarques : De telles dispositions se démarquent de celles de l'accord de consortium AnaEE-France par une ouverture accrue des données[5] pour se placer dans le cadre du plan national pour la science ouverte et de la loi pour une république numérique.

5 IMPLEMENTATION DE LA POLITIQUE DES DONNEES

L'implémentation de la politique s'appuie sur des documents à caractère contractuel comme des chartes d'accès aux services ou des licences sur les jeux de données, et sur les systèmes d'information assurant la mise à disposition des données.

5.1 Chartes d'accès aux services

La charte a deux rôles :

- régir les échanges de données entre la plateforme et l'utilisateur afin d'assurer la réalisation de l'expérimentation et la possibilité d'utiliser les données à des fins d'amélioration des services délivrées par la plateforme.
- rappeler à l'utilisateur les règles de diffusion des jeux de données contribuant à l'Open Data conformément au plan national pour la science ouverte et à la législation (loi Lemaire) et en accord avec la politique des données de l'infrastructure nationale.

Dans la suite, les principes généraux sont déclinés par type de plateforme et l'objet de ce chapitre est de proposer une méthodologie pour aborder les questions relatives à l'usage des données dans les chartes. La question de la diffusion des données à un tiers n'est pas l'objet des chartes, si ce n'est le rappel des principes AnaEE-France et les engagements pris par les utilisateurs pour diffuser leurs données au terme de l'expérimentation.

.

5.2 Nature des jeux de données et règles d'utilisation

Dans le tableau ci-dessous, la nature des jeux de données ainsi que les principes de leur utilisation sont donnés par catégorie de plateforme de l'infrastructure AnaEE-France.

Catégorie de service	**Nature des jeux de données**	**Règles sur la fourniture et l'utilisation des métadonnées et données**
Plateforme d'expérimentation à façon (Ecotrons, dispositifs semi-contrôlés et plateformes d'écologie expérimental des Nouragues et du Lautaret)	Une expérience ➔ Un jeu de données composé de 3 sous-entités : - données de base (DB) de la plateforme caractérisant par exemple l'environnement (météo, sol …) - Données expérimentales collectées par la plateforme (DP) pour le compte de l'utilisateur (mesure de gaz dans une cellule, température d'un bassin … - Données expérimentales additionnelles produites et collectées par l'utilisateur (DU)	1- L'utilisateur s'engage à fournir à la plateforme les métadonnées descriptives de l'expérience (écosystème, manipulation réalisée, observations réalisées) pour alimenter la base de données des projets hébergés par la plateforme. 2- Dans le respect du cadre juridique relatif au partage des données, l'utilisateur responsable scientifique d'une expérience s'engage à diffuser le jeu de données librement et sans contrepartie financière. 3- Les DU pourront être gérées par le SI de la plateforme lorsque cela sera techniquement possible. Les utilisateurs bénéficieront ainsi des services de publication des données offerts par la plateforme*. 4- La plateforme garde toute liberté pour l'usage de ses données de base (DB) et peut utiliser librement les données expérimentales qu'elle a collectées pour le compte de l'utilisateur (DP) pour les besoins d'amélioration du service. 5- Si l'utilisateur est défaillant dans la publication du jeu de données et sa diffusion, la plateforme pourra diffuser les données qu'elle aura acquises (DP) après le délai éventuel d'ouverture fixé avec l'utilisateur, ainsi que toutes les données mises à disposition via la publication d'articles scientifiques
Plateformes d'expérimentation à long terme (dispositifs in natura)	Données de bases produites par la plateforme**(DB) sur le dispositif expérimental à long terme Données additionnelles produites par l'utilisateur sur le dispositif expérimental (DU)	1- La plateforme s'engage à fournir à l'infrastructure AnaEE-France les métadonnées descriptives du dispositif expérimental de long terme (écosystèmes, manipulations réalisées, observations réalisées) pour alimenter en métadonnées la base AnaEE- France. 2- L'utilisateur s'engage à compléter les métadonnées de l'expérimentation à long terme avec les informations relatives aux DU.

		3- Les données DB (historiques et données en cours) sont mises à disposition des utilisateurs à partir moment où elles sont exploitables***. 4- Les utilisateurs mettent leurs données (DU) à disposition de la plateforme pour leur intégration dans son SI. 5- Dans le respect du cadre juridique relatif au partage des données, l'utilisateur responsable scientifique de l'acquisition des DU s'engage à diffuser le jeu de données librement et sans contrepartie financière dès leur achèvement ou au plus tard après le délai d'ouverture fixé avec la plateforme. Les DU seront diffusées selon les mêmes supports que les DB.
Plateformes analytiques et instruments partagés	Un jeu de données composé de deux sous-entités : Données produites par les moyens analytiques**** (DP) Métadonnées fournies par l'utilisateur (MDU) pour l'interprétation des mesures	1- L'utilisateur s'engage à fournir à la plateforme les métadonnées descriptives de l'expérience (écosystème, manipulation réalisée, observations réalisées) pour alimenter la base de données des projets hébergés par la plateforme. 2- La plateforme dispose de l'usage des DP et des MDU pour les besoins d'amélioration du service (amélioration du référentiel) 3- Si les données DP sont acquises dans le cadre d'une expérimentation sur les plateformes AnaEE France, les utilisateurs s'engagent à intégrer les DP dans le jeu de données expérimental et les diffuser selon les principes de la ou les plateformes AnaEE-FR concernées. 4- La plateforme peut diffuser le référentiel selon les règles fixées avec l'utilisateur.

* Pour la publication des données produites par les expérimentations financées majoritairement sur fonds publics, il semblerait naturel que cela soit sous la coresponsabilité de l'utilisateur du service et de la plateforme. Il faut sans doute ici prévoir un article précisant que l'ouverture des données (conformément au cadre législatif) fait l'objet d'un engagement de l'utilisateur du service et qu'en cas de défaillance, la plateforme ou l'infrastructure est en droit/devoir de les publier. L'utilisateur pourrait aussi, d'emblée, transférer à la plateforme (si elle dispose de ce service) la responsabilité de publier les données.
** Notamment toutes les données du dispositif expérimental de base des expérimentations à long terme des services SOERE.
*** lorsque la donnée brute a subi les traitements nécessaires pour être compréhensible, interprétable et exploitable
**** Données produites avec les ressources humaines de la plateforme ou par celles de l'utilisateur

5.3 Licences

Les jeux de données produits pourront faire l'objet d'une licence. La licence à utiliser pour les données produites par les services de AnaEE-France devrait être la 'Licence Ouverte' ((licence LO 2.0, ODBL équivalente à la licence CC-BY). L'ensemble des activités d'AnaEE-Fr et en particulier la production des données s'inscrivant d'emblée dans un cadre international, l'utilisation de la licence CC-BY 4.0 serait la licence recommandée.

5.4 Les Systèmes d'Information (SI)

Les SI associés aux plateformes d'AnaEE-France jouent un rôle important pour la diffusion des données, mais la publication des jeux de données conduit à les mettre à disposition via les entrepôts de données. Deux voies de diffusion sont à considérer.

Dans AnaEE-France, deux familles de SI sont en cours de développement :

- Le SI Ecoinfo a vocation à rassembler les DB des expérimentations à long terme ainsi que les données complémentaires (DU) des utilisateurs. Le SI offre ainsi plusieurs services :
  - Un environnement de stockage propre et sécurisé;
  - Une interface de requêtes pour extraire des jeux de données à façon;
  - Une gestion des droits et d'attribution des licences;
- Le SI ISIA est en cours de développement et destiné aux expérimentations à façon. Il aura dans un premier temps les mêmes fonctions que le SI Ecoinfo pour les données de DB et DP. Un service aux utilisateurs pourra être fourni pour gérer également les DU et ainsi obtenir un environnement unique pour gérer l'ensemble du jeu de données d'une expérience et faciliter l'accès aux données et leur publication à la fin du processus.

Le SI global d'AnaEE-France valorise ces SI 'locaux' en développant leur l'interopérabilité au travers de l'annotation sémantique des ressources. Cette annotation utilise une ontologie partagée basée sur OBOE. Des pipelines informatiques sont développés pour automatiser les traitements sémantiques. Un premier est consacré à l'annotation et à la production des données et métadonnées. Les métadonnées décrivent les ressources (ici données) gérées dans les SI locaux et alimentent le portail AnaEE-France d'accès à ces ressources. Un deuxième pipeline est consacré à l'exploitation des données sémantiques à travers la génération i) de métadonnées ISO et GeoDCAT normalisées et ii) de fichiers de données (NetCDF) à partir de périmètres sélectionnés (sites expérimentaux, années, facteurs expérimentaux, variables mesurées ...). Ces pipelines sont développés dans le contexte d'ENVRIplus et contribuent à son portefeuille de services. La généricité des outils doit garantir leur réutilisation dans différents contextes d'ontologies et de bases de données.

L'utilisation des SI locaux pour diffuser les données est un environnement plutôt approprié. Il souffre néanmoins d'être en perpétuelle évolution et il peut être difficile de raccorder une extraction à un DOI qui fige l'état d'une base à un état donné. Il faut donc établir une stratégie qui tienne compte des différentes voies de diffusion avec néanmoins un principe partagé qui est le référencement avec un DOI de toutes les données récupérées par un utilisateur, quelle que soit la voie de diffusion. On peut ainsi imaginer plusieurs scénarios (liste non exhaustive) :

- Les jeux de données publiés et déposés dans les entrepôts constituent la seule voie de diffusion ouverte des données. Les SI des plateformes demeurent un outil interne permettant de faciliter le travail des utilisateurs et les gestionnaires des plateformes. Les jeux de données déposés dans les entrepôts sont produits soit directement à partir des SI soit grâce au pipeline d'exploitation des données sémantiques qui offre un service de génération du jeu de données (extrait depuis le SI de la plateforme), d'attribution d'un DOI et de dépôt sur un entrepôt.
- Le SI dans son ensemble est figé et publié lorsque le jeu de données est achevé. Toutes extractions du SI doivent alors faire référence au DOI associé. On peut imaginer un versionnage du SI (en particulier pour les expérimentations à long terme) incluant les mises à jour. Cela suppose d'avoir

en permanence deux versions du SI, une de travail dans lequel se font les mises à jour et une version ouverte et figée qui est rattachée au DOI.

- Les deux systèmes précédents co-existent avec l'inconvénient qu'une même donnée peut être rattachée à deux DOI.