

OPEN SCIENCE – DONNEES EXPERIMENTALES, OBSERVATION, ENVIRONNEMENT

Proposition de recommandations pour le partage des données issues d'expérimentations, d'observations ou de simulations

Composition du groupe de travail

Sandrine Auzoux (Cirad, expérimentation végétale, interaction bioagresseurs et insectes), Anna Cohen-Nabeiro (Fondation pour la Recherche sur la Biodiversité, ECOSCOPE, écologie, ressources génétiques, observations), Marc Deconchat (INRA, écologie forestière et écologie du paysage), Aurélie Delavaud (Fondation pour la Recherche sur la Biodiversité, ECOSCOPE, écologie, ressources génétiques, observations), Robin Goffaux (Fondation pour la Recherche sur la Biodiversité, ECOSCOPE, écologie, ressources génétiques, observations), Wilfried Heintz (INRA, écologie et forêts), Jean-François Hocquette (INRA, expérimentation animale), Daniel Jacob (INRA, métabolomique végétale), François Laperruque (INRA, SI, expérimentation animale), Isabelle Lebert (INRA, expérimentation animale, épidémiologie), Denis Loustau (INRA, infrastructure ICOS), Mélanie Martignon (INRA, CNUE et expérimentation animale, co-animatrice du groupe de réflexion), Jérôme Molénat (INRA, ORE/SOERE), Sylvie Nugier (INRA, RSSI), Sébastien Paradis (Cirad, expérimentation végétale, interaction bioagresseurs et insectes), Christian Pichot (INRA, expérimentation et observation, écosystèmes continentaux, SI AnaEE-France), Cyril Pommier (INRA, plantes, SI, expérimentations, co-animateur du groupe de réflexion).

Table des matières

| | | |
|----------|---|----------|
| 1 | INTRODUCTION | 1 |
| 2 | QUELLES DONNEES PARTAGEONS-NOUS ?..... | 2 |
| 2.1 | POUR LES METADONNEES | 2 |
| 2.2 | POUR LES DONNEES | 2 |
| 2.3 | NATURE DES DONNEES..... | 3 |
| 3 | SECURITE..... | 4 |
| 4 | DEFINIR LES CIBLES STRATEGIQUES..... | 4 |
| 5 | CYCLE DE VIE | 5 |
| 6 | COMMENT PARTAGE-T-ON ? | 6 |
| 6.1 | A PROPOS DU « DIFFUSEUR » DE DONNEES..... | 6 |
| 6.2 | A PROPOS DE L'UTILISATEUR DE DONNEES | 7 |
| 7 | ANNEXES..... | 8 |
| 7.1 | ENTREPOTS DE REFERENCE..... | 8 |
| 7.2 | STANDARDS, ONTOLOGIES ET VOCABULAIRES CONTROLES. | 10 |
| 7.3 | REFERENCES, SOURCES ET LEXIQUE | 12 |

1 Introduction

On entend par données « issues d'expérimentations » celles acquises dans le cadre de dispositifs où certains facteurs sont contrôlés ou manipulés (phénotypage, essais agronomiques, caractérisations de collections de types ressource génétiques, caractérisation écophysiologique en phytotron...). Les données « issues d'observations » sont acquises sur des objets non manipulés, le plus souvent en milieu naturel et dans le contexte INRA, fréquemment sur le long terme. Il peut s'agir par exemple de mesures de flux au-dessus d'un couvert forestier ou prairial, de variations de potentiel hydrique des plantes ou d'autres mesures effectuées au sein d'observatoires labellisés [1][2][3]. Enfin, les données « issues de simulations » sont produites par des expériences de calcul numérique mobilisant un ensemble organisé d'algorithmes et équations.

Pour être partageables, diffusables et (re)utilisables, les données doivent être accompagnées de métadonnées. Les métadonnées, ou données sur les données, telles que définies par exemple dans le cadre du Dublin Core [4] permettent de préciser la source des données (titre ou nom, auteur, date, thématique, etc...), leur nature (expérimentation, observation, résultat de simulations), leur contenu (descripteurs, mots clefs, résumé). Dans le cas des données scientifiques produites dans les organismes de recherche comme l'INRA, les métadonnées recouvrent aussi les conditions et les environnements/lieux d'expérimentation et d'observation. Les métadonnées permettent également de pérenniser l'intelligibilité des données (compréhension du contexte d'acquisition et de leur organisation lorsque le référent n'est plus là), de répondre aux requis du législateur (Directive européenne INSPIRE) voire des financeurs. Elles constituent la première étape de la mise en partage des données (Open Data). La description des données via la production de métadonnées apporte ainsi de multiples utilités :

- Découverte et valorisation des jeux de données :
 - Rechercher et identifier des jeux de données hétérogènes (origine, format, contenu)
 - Obtenir des DOI pour les jeux de données puis les publier dans des entrepôts, sous forme d'articles (« data paper ») dans des revues mixtes (publications scientifiques et de données) ou dédiées (Data Journal)
- Ré-utilisation des données :
 - Évaluer le contenu, la pertinence, les emprises et la qualité des jeux de données
 - Utiliser de façon appropriée les données (prise en compte des limites d'utilisation)
 - Réaliser des synthèses (semi-)automatiques sur les données
- Accès aux données :
 - Savoir comment et où accéder aux données - directement (lien ou automatisation) ou indirectement (personne contact) et sous quels formats
 - Rôle dans l'interopérabilité via les standards d'échanges (XML)
- Pérennisation des données :
 - Structurer et décrire les données pour une gestion optimale
 - Documenter le contexte d'interprétation des données
- Réponse aux législateurs et bailleurs
 - Mettre en conformité vis-à-vis de la réglementation (Exemple : directive INSPIRE)
 - Réaliser une étape de préconisations « open access »

2 Quelles données partageons-nous ?

2.1 Pour les métadonnées

Nous proposons de classer les métadonnées en trois niveaux, découverte / exploration / exploitation [5]: (i) les métadonnées de découverte sont utilisées dans les mécanismes de recherche et identifient la ressource décrite, (ii) les métadonnées d'exploration, ou essentielles permettent d'en savoir plus sur les ressources et de sélectionner des cibles d'intérêt et (iii) les métadonnées d'exploitation, spécifiques, qui sont généralement attachées aux données elles-mêmes et répondent aux besoins de groupes restreints, utilisateurs des données. Les métadonnées d'exploration doivent aussi renseigner les dispositifs expérimentaux associés aux protocoles mis en place, ce qui correspond aux conditions d'environnement dans lesquelles ont été acquises les données. De même, les dispositifs d'observations doivent être décrits. Ainsi, ces métadonnées d'exploration doivent permettre de retracer complètement le processus de production de la donnée (acquisition, curation) en renseignant les protocoles de mesure, capteur utilisé, calcul et traitement de mise en forme des données.

Nous proposons donc de :

- 1) Partager les métadonnées de découverte (Qui ?, quoi ?, où ?, quand ?)

L'objectif est de porter à connaissance l'existence des ressources et plus particulièrement des données, de les identifier et de les faire connaître *via* les moteurs de recherche.

- 2) Utiliser les métadonnées de découverte comme « catalogue » pour accéder aux métadonnées d'exploration et d'exploitation.

D'autre part, il faut respecter les standards de métadonnées quand ils existent. Les métadonnées des données environnementales spatialisées doivent être gérées en conformité avec les standards de la directive européenne INSPIRE [6] visant à la construction d'une infrastructure européenne des données relatives à l'environnement. Elles doivent être produites selon le standard ISO 19115. Pour les métadonnées traitant des travaux en écologie, l'Ecological Metadata Standard permet de décrire différents aspects des jeux de données grâce à un ensemble de modules thématiques¹. L'ensemble des métadonnées devra s'appuyer sur plusieurs vocabulaires contrôlés généralistes ou spécifiques au domaine. Ces vocabulaires existent ou sont à construire selon les cas. Une proposition de recommandations de vocabulaires (thésaurus) et ontologies se trouve en annexe.

2.2 Pour les données

Les données concernées par l'Open Data sont d'une part celles valorisées par une publication et d'autres part celles pouvant elles-mêmes être publiées sans accompagner un article scientifique classique. Il s'agit d'une preuve scientifique. Les recommandations portent sur des données informatisées de manière structurée. Le processus d'informatisation de données historiques ou faiblement structurées est coûteux. Nous recommandons de mettre l'accent sur les données en cours de production dans le cadre de la démarche Open Data. Pour les données disponibles sous forme papier, nous recommandons de prendre conseil auprès de l'archiviste de l'INRA (archives@paris.inra.fr). L'information sur les protocoles d'acquisitions et de traitement des données est renseignée dans les métadonnées d'exploitation.

Ces données se trouvent d'abord sous forme de données brutes non transformées, telles qu'elles sont acquises par les systèmes de mesure (images, volt, ampère, ohm, fréquence, etc) puis en données brutes transformées c'est à dire converties en [cpommier1] grandeur physique et unités S.I. (kg, m, s, °C, °K, etc...). Les données peuvent être obtenues

¹ <http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-language>

directement à ce niveau, par exemple en notant des dates des stades de développement (phénologie) ou via des observations réalisées à la main telles que des mesures physiques ou des notations de maladies. L'utilisation de ces données brutes transformées produit le plus souvent des données élaborées : carte de température interpolée sur un domaine spatial par exemple, flux de chaleur entre compartiments du système étudié, résistance à un stress, etc. Ces données élaborées peuvent être obtenues en combinant plusieurs types de données élaborées ou brutes. A chacun de ces niveaux d'élaboration correspond un ensemble de métadonnées spécifiques documentant leur mode d'obtention, les étapes de calculs appliqués, coefficients d'étalonnage etc... Il faut également noter l'existence de proposition de "standards" de variables, basées sur des données élaborées, pour le climat (Variables Climatiques Essentielles²) et pour la biodiversité (Variables Essentielles de Biodiversité³).

Quel niveau d'élaboration de la donnée partagée ?

Il est souhaitable que le niveau d'élaboration de la donnée partagée soit adéquat et pertinent ; c'est-à-dire qu'il permette l'utilisation de la donnée par d'autres. Une fois publiées, à travers un article ou isolément, les données acquises sur fonds publics doivent être mises à la disposition de la communauté scientifique sans restriction d'accès. Le partage des données nécessite de la part du producteur-gestionnaire [7] de ces données un effort de documentation, de mise en forme et d'alimentation des métadonnées. Le partage de la donnée ne s'accompagne cependant pas de transformation ou d'analyse additionnelle de la part du producteur-gestionnaire. Il est en effet nécessaire d'alléger la charge de travail du producteur-gestionnaire qui partage ses données en limitant son investissement au strict minimum permettant la réutilisation des données.

La diffusion des données et de leurs métadonnées doit être gratuite et simple pour les utilisateurs (consultation et téléchargement). Les données acquises sur fonds publics doivent être mises à la disposition de la communauté scientifique sans restrictions d'accès. Les éventuelles conditions liées à une utilisation à des fins commerciales doivent être précisées par le groupe de travail juridique. Leur réutilisation pour publication implique de citer aisément la source, en se basant sur un système de référencement (identifiant du jeu de données (DOI), revues de publication de données, entrepôt de données, etc...).

2.3 Nature des Données

Les données brutes sont généralement de type numérique (images, spectres, etc....). Elles sont potentiellement très volumineuses, pouvant atteindre une centaine de téra octets par an pour un réseau de phénotypage haut débit. Ces données sont de plus non reproductibles, et dans la mesure du possible doivent être conservées en tant que données patrimoniales.

Les données élaborées produites à partir des données brutes sont des fichiers alphanumériques de types csv, xml, excel, etc.... Ils représentent un volume de quelques giga octets par an et ne posent donc pas de problème de stockage.

Ces données doivent être pérennisées via un stockage dans des entrepôts adaptés. Le choix du ou des entrepôts dépend du consortium produisant les données. Avant leur production, un plan de gestion des données (PGD ou Data Management Plan) adapté doit avoir été conçu via des outils disponibles sur le portail Open Science de l'INRA.

² <https://www.wmo.int/pages/prog/gcos/index.php?name=EssentialClimateVariables>

³ <http://geobon.org/essential-biodiversity-variables/criteria-to-adopt-an-ebv/>

3 Sécurité

La sécurité des systèmes d'information impose d'évaluer les critères de disponibilité, d'intégrité, de confidentialité et de traçabilité concernant les données et les processus les gérant. La criticité des données doit donc être évaluée avant leur partage et détermine un niveau de protection adapté. Celui-ci peut être garanti par des mécanismes technologiques mais aussi par des éléments juridiques.

Certaines données sont protégées par des lois ou règlements. Ainsi les données à caractère personnel sont protégées par la loi Informatique et liberté. Elles doivent être déclarées et être protégées en conséquence. Parmi ces données, certaines sont dites sensibles comme par exemple celles concernant la santé ou la sécurité. De plus, l'hébergement des informations sensibles doit se faire sur le territoire national.

Des données peuvent aussi avoir des restrictions de diffusion, pour des raisons de protection du patrimoine scientifique et technique de la nation.

Il faut être vigilant en cas de réutilisation de ces types de données lors de nouveaux projets notamment lorsque la finalité du traitement informatique change.

C'est au responsable des données de s'assurer que la disponibilité, l'intégrité et la confidentialité sont au niveau requis de protection pour ses données, ceci en s'assurant que l'offre de service institutionnelle ou partenariale choisie pour leur stockage et leur accès est judicieuse.

Les contraintes de protection des données précitées ne sont pas forcément antinomiques avec le partage des données. Il suffit de savoir situer l'étape de gestion de la donnée en cours dans celles de la vie de la donnée (ou du plan de gestion de la donnée). Ce qui est protégé à un instant T pourra être accessible à un instant T+1 ou à un niveau d'agrégation différent ou à défaut partagé sous condition ou convention.

4 Définir les cibles stratégiques

Les enjeux sont que l'INRA fasse reconnaître et valorise sa capacité d'acquisition/production de données d'expérimentation et d'observation afin (i) de se positionner comme référence internationale pour les chercheurs, décideurs et gestionnaires, (ii) d'accroître la valorisation de ses données à travers leur utilisation et leur ré-utilisation, de favoriser leur pérennité, et (iii) d'augmenter les opportunités de partenariat scientifique, académique et socio-économique. En plus de cette diffusion des données de l'INRA vers la communauté scientifique internationale, il s'agit d'être à la pointe du transfert recherche-société et de démontrer l'expertise et l'indépendance de l'INRA dans ce domaine. Le partage de données est aussi un moyen de populariser la science et la recherche auprès de différents publics et gestionnaires. L'ouverture aux données et métadonnées et la promotion de leur (ré)utilisation par les acteurs ciblés est une voie privilégiée pour promouvoir la recherche coopérative partagée. Le choix de la langue des métadonnées et données (français, espagnol, chinois ou anglais principalement mais pas exclusivement) n'est pas simple et dépend du public visé (communauté scientifique internationale, société civile, ...) et du consortium produisant les données. La langue utilisée peut être elle-même renseignée dans les métadonnées.

Les cibles stratégiques sont nombreuses et peuvent être regroupées ainsi :

- Recherche nationale, Recherche internationale, Enseignement supérieur, Instituts

techniques

- Partenaires privés, PME/PMI, Industries : agro-alimentaires, pharmaco-chimie, eau, ...
- Associations, Gestionnaires publics (collectivités territoriales,...), Politiques/décisionnaires, Bureaux d'études, Syndicats agricoles, Organismes agricoles professionnels (chambre d'agriculture, institut technique, ...), citoyens, Enseignement secondaire général ou professionnel

La diffusion large et rapide des données aux groupes d'utilisateurs permet de les valoriser de façon optimale en permettant leur réutilisation dans le cadre d'un travail de recherche, d'enseignement, d'éducation ou de valorisation appliquée. Elle n'est en aucun cas un abandon de propriété ou une perte de reconnaissance pour les auteurs (autorat, authorship) mais bien une démultiplication du potentiel de développement de connaissances qu'offre chaque type de données.

Chaque donnée dans sa discipline et son domaine de recherche apparaît comme un cas particulier. Le choix du niveau de données à partager (brutes, transformées) et les conditions d'accès (libre, sur demande, restriction pour usage commerciale) sont des points critiques mais qui ne dépendent pas uniquement du public visé. Ainsi la nature des données partagées et leurs conditions d'accès sont à définir par le chercheur (et son institution ?) selon le type de la donnée : cela dépend de l'origine de la donnée (lieu d'obtention, moyen, ...), sa nature, son niveau d'élaboration (brute non transformée, brute transformée, élaborée) et le public visé. S'il est difficile de dire quelle donnée peut être partagée et pour qui, il est possible de définir comment : la diffusion de la donnée est toujours accompagnée des métadonnées, les données sont en accès libre mais sous condition selon le public considéré (juridique, citation, etc...).

Certains contextes de partage présentent des particularités nécessitant une adaptation des données diffusables selon le public visé (exemple : données de géolocalisation concernant des données épidémiologiques, cultures OGM, espèces rares et/ou menacées). De plus, certains protocoles expérimentaux ou analytiques constituent l'avantage compétitif, le savoir-faire spécifique que le chercheur/expérimentateur ne peut diffuser publiquement sans contrôle.

5 Cycle de vie

Le cycle de vie de la donnée commence par la conception de la question scientifique, suivie par la conception du projet et de l'expérimentation qui permettront sa production et la gestion de son ouverture. La réutilisation des données peut en effet impliquer des questionnements scientifiques très différents de ceux qui ont mené à leur production. A toutes ces étapes il y a un protocole permettant de faire passer la donnée d'un état à l'autre. Les informations relatives au protocole constituent les métadonnées d'exploitation.

Etapes définies du cycle de vie d'une donnée :

- 1) Planification de la production de la donnée⁴
- 2) Production de la donnée brute et des métadonnées associées
- 3) Traitement à donnée transformée, à donnée élaborée. Les métadonnées associées doivent être renseignées à chaque étape.
- 4) PARTAGE [8]
 - a. Publication et valorisation scientifique en lien avec l'objectif primaire.
 - b. Valorisation du jeu de données (publication dans un entrepôt, dans une revue généraliste ou spécialisée en tant que « data paper », ...). Ce mode de publication est très intéressant pour les jeux de données liés à un projet ou programme précis. Pour les chroniques de données du type SOERE, qui par définition sont toujours en « acquisition », il faut profiter d'un événement ou d'une publication scientifique particulière ou prévoir une publication

⁴ De type préconisé par <https://www.dataone.org/best-practices>

régulière du jeu (5 ou 10 ans...) et informer du fait que la chronique continue d'incrémenter la production de données.

5) 2ème vie :

- a. Réutilisation dans un autre projet
- b. Complément des données produites par d'autres données
- c. Stockage/archivage sur le long terme
- d. Fin de vie éventuelle

Etape du partage des données

- Nous recommandons que le partage ait lieu le plus tôt possible, au plus tard avant la fin de la période d'exclusivité mentionnée ci-après. En effet, il nous paraît important que le producteur-gestionnaire soit encore actif sur ces données, car 1/ il possède une connaissance et une expertise sur les données acquises et 2/ différentes études⁵ montrent l'importance du lien entre producteur et ré-utilisateur de données pour le partage des données, notamment en sciences de l'écologie. Le renseignement complet des métadonnées - avec la description du contexte de collecte, du protocole et de l'organisation des données - permet toutefois de pérenniser les données en limitant le besoin d'un recours au producteur initial.

- Le partage peut se faire à partir du moment où le jeu de données est cohérent et complet. Ce dernier point est à l'appréciation du producteur-gestionnaire.

- pour une expérimentation bien délimitée dans le temps, annuelle par exemple, le moment du partage correspond à la fin de l'expérimentation, après transformation et élaboration des données. Si le nettoyage et l'élaboration des données sont trop longs, une mise à disposition des données, clairement notée non nettoyées, est recommandée. (voir annexe)
- pour les observations continues (réseaux multiannuels, etc...), le choix est à l'appréciation des producteur-gestionnaires, avec une fréquence *a priori* au minimum annuelle.
- pour les SOERE et ORE, le partage est recommandé dès que la donnée est validée par le processus mis en place au sein de l'ORE/SOERE

- Période d'exclusivité : à partir du moment où un jeu de données est acquis, le producteur dispose d'une période d'exclusivité (par exemple de six mois) pour nettoyer et valoriser ce jeu de données, sauf dans le cas de schémas expérimentaux complexes qui doivent faire partie d'un accord de consortium spécifique précisant explicitement la période d'exclusivité ou être inscrit dans une procédure type ORE visant la mise à disposition publique.

- Nous recommandons par ailleurs que le partage de données soit reconnu et valorisé par la mention « production de données primaires » explicitée lors de la diffusion des données en faveur des auteurs des données.

6 Comment partage-t-on ?

6.1 A propos du « diffuseur » de données.

La mise à disposition de la donnée est sous la responsabilité du scientifique, producteur ou gestionnaire à l'origine de sa production selon des règles partagées de diffusion.

Nous pensons qu'il est nécessaire de mettre en place des systèmes d'informations (SI) qui gèrent et affichent des métadonnées de découverte avec leurs sources, permettant ainsi l'accès aux métadonnées d'exploration, d'exploitation et aux différents jeux de données. Ces

⁵ voir Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B. and Gemeinholzer, B. 2012. The user's view on biodiversity data sharing – Investigating facts of acceptance requirements to realize a sustainable use of research data. *Ecological Informatics*, 11: 25-33, doi:10.1016/j.ecoinf.2012.03.004

et Tenopir C., Allard S., Douglass K., Aydinoglu A.U., Wu L., Read E., Manoff M. and Frame M. 2011. Data sharing by scientists: practices and perceptions. *PLoS ONE* 6(6): e21101 doi:10.1371/journal.pone.0021101

systèmes d'informations peuvent être distribués avec des portails interrogeant plusieurs sources de données, éventuellement thématiques, plus un portail INRA dédié à la vue d'ensemble des données produites par l'institut. Ceci permettrait de réutiliser les SI existants. Cette idée est par exemple implémentée dans les portails GBIF [9], Transplant [10], Metabolomexchange [11], AnaEE-France, ECOSCOPE⁶ et sera utilisée pour d'autres infrastructures comme Elixir/IFB ou Phenome.

Les différentes personnes ayant participé à la production de la donnée (techniciens, unités expérimentales, dispositifs,...) doivent être associées au partage des données, par exemple en les nommant dans les métadonnées. Nous recommandons que le partage des données ne soit pas un abandon de propriété, bien au contraire il doit aussi permettre d'identifier les contributions et définir une ou plusieurs licences d'utilisation. Si une diffusion efficace des données doit impérativement rester simple et rapide (« en deux clics »), leur utilisation par l'utilisateur doit s'assortir de la souscription par celui-ci d'une obligation (morale) de citation et l'invitation à proposer aux auteurs une valorisation en partenariat.

Se rattacher aux standards internationaux de données existants ou en cours de construction sera nécessaire. Par exemple le standard ISO19115/19139 dans le cadre de la directive européenne INSPIRE pour les données géoréférencées, DublinCore pour la description des ressources numériques (norme internationale [ISO 15836](#)), les recommandations de type MIAME [12], Darwincore [13] pour les données sur la biodiversité. A noter que plusieurs projets et communautés (WheatIS, Transplant, AnaEE France, ECOSCOPE, Phenome, ...) travaillent sur cette question et que de nouvelles propositions peuvent arriver.

6.2 A propos de l'utilisateur de données

Il est important de rendre compte de l'utilisation des données, de contrôler le respect des règles d'utilisation et d'identifier des partenariats potentiels pour le producteur-gestionnaire de données. C'est pourquoi, au moment de la diffusion, nous recommandons que, dans la mesure du possible, la personne téléchargeant un ou des jeux de données ait besoin de s'identifier et que les téléchargements de données soient tracés. Par ailleurs, outre s'identifier, cette personne doit avoir d'autres devoirs tels que :

- référencer les données comme source
- se conformer à la licence d'utilisation
- accepter une politique d'utilisation éthique (fair use policy), à définir, expliquant comment citer la source de données et si leur auteur doit être associé à la valorisation prévue (case à cocher « je reconnais avoir pris connaissance de etc. »).

Les composantes juridiques de la re-utilisation des données et le choix des licences adaptées devront être précisés à partir des travaux du 'groupe juridique' du chantier Open Data.

⁶ <http://www.fondationbiodiversite.fr/ecoscope/fr/>

7 Annexes

7.1 Entrepôts de référence

Il s'agit des entrepôts non liés à un projet ou une plateforme spécifique. Ils sont en mesure d'héberger des données provenant de divers projets ou de fédérer des systèmes d'informations projets ou plateforme.

| | Phénotypage | Environnement | Métabolomique | ORE et SOERE | Biodiversité |
|---|-------------|---------------|---------------|--------------|--------------|
| Climatik https://intranet.inra.fr/climatik/do/welcome | | X | | | |
| Metabolomexchange http://metabolomexchange.org | | | X | | |
| Agrosyst https://agrosyst.fr | | X | | | |
| Anaee France http://www.anaee-s.fr/ | X | X | | X | X |
| MetaboLights (EMBL-EBI) http://www.ebi.ac.uk/metabolights/ | | | X | | |
| GnpIS https://urgi.versailles.inra.fr/gnpis/ (Phenome portal) | X | | | | |
| SOERE RBV : portail des métadonnées des observatoires hydrologiques français: http://portailrbv.sedoo.fr/ | | X | | X | |
| Infosol http://www.gissol.fr/outils/donesol-web-336 | | X | | | |

| | Phénotypage | Environnement | Métabolomique | ORE et SOERE | Biodiversité |
|---|-------------|---------------|---------------|--------------|--------------|
| PlantNet http://www.plantnet-project.org/papyrus.php | X | | | | |
| GBIF http://www.gbif.org/ | | | | | X |
| Système national d'information génétique GA | X | | | | |
| Bases de données plateforme, ex : Phenopsis, Vitphe, Phenoarch/PHIS | X | X | | | |
| ECOSCOPE (Métadonnées) | X | X | X | X | X |
| Gentoul Bioinfo http://get.genotoul.fr/ | X | | | | |
| CRB Anim (infrastructure nationale en biologie et santé animale) http://www.crb-anim.fr/ | X | | | | X |
| SIDEx (protocoles expérimentaux animaux / GA - PHASE) | X | | | | |
| Plateforme SIGENAE (génomique animale) http://www.sigenae.org/ | X | | | | |

7.2 Standards, ontologies et vocabulaires contrôlés.

| | Phénotypage | Environnement | Phénotypage moléculaire | Générale |
|--|-------------|---------------|-------------------------|----------|
| Agrovoc | | | | X |
| Plant Ontology http://www.plantontology.org/ | X | | | X |
| Crop Ontology www.croponontology.org | X | X | | X |
| Agroportal http://agroportal.lirmm.fr/ | X | X | X | X |
| Wheat Data Interoperability (RDA et Wheat Initiative) http://ist.blogs.inra.fr/wdi/ | X | X | X | X |
| Brenda | | | X | |
| Golm Metabolom | | | X | |
| Chebi | | | X | |
| Environment Ontology (XEO) https://xao.codeplex.com/ | | X | | |
| Extensible Observation Ontology (OBOE) | | X | | |

| | Phénotypage | Environnement | Phénotypage moléculaire | Générale |
|---|-------------|---------------|-------------------------|----------|
| Socio-Ecological Research and Observation oNTOlogy (SERONTO) | | X | | |
| EnvThes - Environmental Thesaurus (LTER-Europe) | | X | | |
| AnaEE-France thesaurus | X | X | | |
| Environmental Applications Reference Thesaurus (EARTH) | | | | X |
| Ontology for Biomedical Investigations (OBI) | | | | X |
| Animal Trait Ontology for Livestock (ATOL) http://www.atol-ontology.com/ | X | | | |
| Ecological Metadata Language (EML) http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-language | X | X | X | X |
| ISO19115/19139 (INSPIRE) | | X | | X |
| Typologie descripteurs observation de la biodiversité (ECOSCOPE) | X | X | X | X |

Annuaire de standards de métadonnées par discipline :
<http://www.dcc.ac.uk/resources/metadata-standards>

7.3 Références, sources et lexique

- [1] Integrated Carbon Observation System (ICOS)
- [2] Observatoires de Recherche en Environnement (ORE)
- [3] Systèmes d'Observations et d'Expérimentation au long terme pour la Recherche en Environnement (SOERE)
- [4] Dublin Core basics : <http://dublincore.org/metadata-basics/>, Dublin Core bibliographic citations : <http://dublincore.org/documents/dc-citation-guidelines/>
- [5] Métadonnées Edition 1.2. Descriptif de contenu. Métadonnées des Référentiels Géographiques, IGN, décembre 2008. Page 6 paragraphe 2.1
- [6] Site Infrastructure for Spatial Information in the European Community : <http://inspire.jrc.ec.europa.eu/> Directive 2007/2/CE du Parlement européen et du Conseil du 14 mars 2007 établissant une infrastructure d'information géographique dans la Communauté européenne (INSPIRE). <http://inspire.ign.fr/directive/presentation>
- [7] La gestion de la donnée est un rôle assuré par le producteur de la donnée et/ou par un gestionnaire de données. D'où la définition du producteur-gestionnaire qui englobe les deux rôles.
- [8] Les données de terrain (données d'enquêtes) devront respecter les règles de la CNIL dès lors que des données sur les personnes sont concernées.
- [9] <http://www.gbif.org/>
- [10] <http://www.transplantdb.eu/>
- [11] <http://metabolomexchange.org/>
- [12] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FCP, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. 2001. Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nature Genetics* **29**, 365–371.
- [13] <http://rs.tdwg.org/dwc/>