
Partage des données de la recherche

Chantier Entrepôt - Annuaire



Septembre / 2014

Ce document est un rapport intermédiaire des travaux du groupe de travail sur les chantiers Entrepôt et Annuaire.

Coordination : Dzale Yeumo Esther, Soler Lydie

Membres du groupe de travail : Aubin Sophie, Boizot Szantai Christine, Buche Patrice, Carrere Sébastien, Dervaux Stéphane, Dibe Juliette, Guinet Nicolas, Hocquette Jean-François, Lassalle Gilles, Legeai Fabrice, Passouant Michel, Pichot Christian, Pommier Cyril.

1	LES CHANTIERS ENTREPOT ET ANNUAIRE : INTRODUCTION	3
1.1	PERIMETRE DU CHANTIER	3
1.2	COMPOSITION DU GROUPE DE TRAVAIL	3
1.3	GLOSSAIRE	3
1.4	ORGANISATION DU DOCUMENT	3
2	UN ENTREPOT DE DONNEES INRA	4
2.1	UN ENTREPOT DE DONNEES INSTITUTIONNEL : DEFINITION, BESOIN, OPPORTUNITES ET RISQUES	4
2.1.1	DEFINITION	4
2.1.2	LES BESOINS EXPRIMES PAR LES FAMILLES DE DONNEES	4
2.1.3	ÉCOSYSTEME INRA ET INTERNATIONAL	5
2.1.4	ENTREPOT INSTITUTIONNEL	5
2.1.5	OPPORTUNITES ET RISQUES D'UN ENTREPOT INSTITUTIONNEL	6
2.2	CARACTERISATION DE L'ENTREPOT INSTITUTIONNEL	7
2.2.1	PERIMETRE DES DONNEES AYANT VOCATION A ETRE STOCKEES DANS L'ENTREPOT	7
2.2.2	LA POLITIQUE DE STOCKAGE ET D'ARCHIVAGE A LONG TERME (PRESERVATION)	8
2.2.3	SYSTEME DE CONSULTATION / RECHERCHE	11
3	LE CHANTIER ANNUAIRE	13
3.1	UN ANNUAIRE INSTITUTIONNEL : DEFINITION, OPPORTUNITES ET RISQUES	13
3.1.1	DEFINITION	13
3.1.2	OPPORTUNITES ET RISQUES	13
3.2	CARACTERISATION DE L'ANNUAIRE INSTITUTIONNEL	14
3.2.1	PERIMETRE DES DONNEES AYANT VOCATION A ETRE REFERENCEES DANS L'ANNUAIRE	14
3.2.2	PRESENTATION DES DONNEES REFERENCEES	14
3.2.3	ALIMENTATION DE L'ANNUAIRE – INTEGRATION DANS L'ECOSYSTEME OPEN DATA	15
3.2.4	SYSTEME DE CONSULTATION/RECHERCHE	15
	ANNEXE	16
1	LISTE INITIALE DES FORMATS PRESERVABLES ET ACCEPTABLES	16

1 Les chantiers Entrepôt et Annuaire : introduction

1.1 Périmètre du chantier

Afin de répondre aux enjeux de l'« Open » et du « Big » data, l'INRA souhaite se doter d'une offre de services en direction des équipes de recherche pour les aider à rentrer dans le mouvement de l'open science. Dans cette démarche de mise en place de services en faveur du partage des données, le séminaire du 2 et 3 décembre 2013 a mis en évidence la nécessité de mener une étude de faisabilité concernant les dispositifs suivant :

- un annuaire des sources de données de l'INRA,
- une infrastructure de dépôt (entrepôt de données),
- et la gestion des identifiants numériques (DOI).

Ce document présente les premières réflexions menées par le groupe de travail sur les deux chantiers Entrepôt de données et Annuaire.

1.2 Composition du groupe de travail

Ce groupe de travail est composé de :

Coordination : Dzale Yeumo Esther, Soler Lydie

Membres du groupe de travail : Aubin Sophie, Boizot Szantai Christine, Buche Patrice, Carrere Sébastien, Dervaux Stéphane, Dibe Juliette, Guinet Nicolas, Hocquette Jean-François, Lassalle Gilles, Legeai Fabrice, Passouant Michel, Pichot Christian, Pommier Cyril.

1.3 Glossaire

Glossaire des principaux termes utilisés dans ce document	
Données partagées	Données citables c'est-à-dire accessibles au-delà des partenaires initiaux ayant permis leur création. Cet accès peut être gratuit ou payant et concerner l'ensemble du public ou un groupe plus restreint (par exemple uniquement des chercheurs).
Déposant	Personne qui dépose des données dans l'entrepôt de données.
Archivage à long terme (préservation)	Stockage de données pour une durée dépassant plusieurs années (ordre de grandeur : 10 ans) avec contrôle de lisibilité des données, opérations de conversion de formats et de migration de supports, restaurations, traçabilité des opérations effectuées sur les données. ¹

1.4 Organisation du document

Le chapitre 2 présente les premières réflexions du groupe de travail concernant le chantier Entrepôt de données et le chapitre 3 celles sur le chantier Annuaire.

¹ D'après le « Nouveau glossaire de l'archivage » :

http://extranet.ucanss.fr/contenu/public/EspaceDeveloppementDurable/pdf/Nouveau_glossaire_de_l_archivage.pdf

2 Un entrepôt de données INRA

2.1 Un entrepôt de données institutionnel : définition, besoin, opportunités et risques

2.1.1 Définition

Un entrepôt **numérique** de données (Digital Repository) est un espace de stockage de données **numériques** et des métadonnées associées. Il offre à minima les services de dépôt, gestion, recherche et extraction.

L'entrepôt dont nous parlons dans ce document s'inscrit dans le mouvement de l'Open Science. Il a vocation à accueillir les données concernées par le partage des données (la suite du document apportera des précisions).

2.1.2 Les besoins exprimés par les familles de données

Le besoin d'un entrepôt institutionnel diffère d'une famille de données à l'autre.

La famille « **Données expérimentales, observations** » a identifié une liste d'entrepôts de données de référence dans lesquels les données produites ou co-produites par l'Inra pourraient être déposées. Elle exprime aussi le besoin de mettre en place à l'INRA « des systèmes d'information qui gèrent et affichent des métadonnées (...) avec leurs sources, permettant ainsi l'accès à différents jeux de données. Ces systèmes d'information peuvent être distribués avec un portail unique interrogeant plusieurs sources de données, éventuellement thématiques. Ceci permettrait de réutiliser les SI existants. »²

La famille « **Données d'enquêtes et analyse textuelle** » a identifié des plateformes proposant des services de partage des données avec une ouverture vers l'extérieur, et à travers lesquelles les données en sciences humaines et sociales pourraient être partagées. Il s'agit des plateformes Quételet et TGIR Huma-Num. En revanche, cette famille de données attend de l'Inra la mise à disposition d'un service opérationnel qui répond aux attentes des éditeurs dans le cadre du partage des données à des fins de reproductibilité. Une telle solution « garantirait la sécurité des données et donnerait la preuve de l'exactitude des résultats »³.

La famille « **Ressources génétiques et génomiques** » a identifié de nombreux entrepôts Inra, nationaux et internationaux dans lesquelles les données de cette famille peuvent être déposées à des fins de partage. Elle fait cependant le constat que les entrepôts centralisés tels que Genbank ont atteint leurs limites, et recommande que l'Inra investisse dans l'hébergement de certains entrepôts internationaux en s'inscrivant ainsi dans l'écosystème du domaine⁴.

En résumé, les familles expriment le besoin d'un entrepôt institutionnel qui s'intégrerait dans les systèmes d'information existants INRA et hors INRA.

² 2013 et 2014. Open Science - Données expérimentales, Observations - Proposition de recommandations. Marc Deconchat, Jean-François Hocquette, Daniel Jacob, François Laperruque, Isabelle Lebert, Denis Loustau, Mélanie Martignon, Jérôme Molénat, Christian Pichot, Cyril Pommier, Jean-François Rami.

³ 2014. Groupe données d'enquêtes et analyse textuelle. Eric Cahuzac (ODR, SAE2), Christine Boizot (ALISS, SAE2), Christophe Bontemps (GREMAQ, SAE2), Pierre Triboulet (AGIR, SAD), Véronique Batifol-Garandel (IST, SAD), Patrice Buche (IAT, CEPIA), Lydie Soler (MIA 518, MIA), Philippe Breucker (SENS, SAE2/SAD), Annie Hofstetter (LAMETA, SAE2), Jean-Marc Rousselle (LAMETA, SAE2) Jean-Loup Dupuy & Cédric Lanu (GAEL, SAE2).

⁴ 2014. Partage des données relatives aux ressources génétiques et génomiques. P. Bessièrès, S. Carrère, S. Casaregola, G. Lassalle, F. Legeai, N. Mohellibi, F. Moreews, J-L. Noyer (CIRAD), H. Quesneville, L. Ranjard, S. Sidibe-Bocs (CIRAD), M. Tixier-Boichard (rédaction)

2.1.3 Écosystème INRA et international

Dans le cadre de l'Open Science, il existe de nombreux entrepôts de données :

- Thématiques ou pluridisciplinaires;
- Portés par des institutions publiques, des organisations à but lucratif ou non lucratif.

Quelques exemples d'entrepôts sont disponibles ici :

<https://wiki.inra.fr/wiki/donneesrechercheist/Main/Comparaison+d%27entrep%C3%B4ts+de+donn%C3%A9es>

D'autre part, différents systèmes d'information permettant le stockage de données existent à l'INRA et pourraient évoluer pour permettre l'ouverture des données dont ils ont déjà la charge. C'est le cas notamment :

- des ORE (Observatoire de Recherche en Environnement) et des SOERE (Systèmes d'Observation et d'Expérimentation pour la Recherche en Environnement) qui assurent la gestion, la cohérence et la mise à disposition des données acquises;
- des plateformes telles que Migale, Genotoul, plateforme bioinformatique de l'URGI, GenoSol, etc.

Une liste d'entrepôts spécifiques aux ressources génétiques et génomiques disponibles à l'Inra est fournie en annexe au rapport du groupe de travail « Ressources génétiques et génomiques » sur le partage des données.

De même, une liste d'entrepôts recommandés pour les données expérimentales et d'observation est disponible en annexe du rapport du groupe de travail correspondant.

2.1.4 Entrepôt institutionnel

Au regard des besoins exprimés par les familles de données et de l'écosystème interne et externe, il nous paraît souhaitable que l'Inra se dote d'un entrepôt institutionnel pour répondre aux besoins suivants :

- stocker, gérer, préserver et rendre accessibles (notamment aux revues) les données qui sont liées à des publications scientifiques dans le respect des droits des auteurs ;
- stocker, gérer, préserver et rendre accessibles les données (non liées à des publications) partagées ou qui ont vocation à l'être mais qui ne disposent pas d'un entrepôt national ou international adéquat pouvant les héberger ;
- fournir des indicateurs sur les accès et les téléchargements des données à leurs auteurs.

Suivant les recommandations de l'organisme de certification DSA (Data Seal of Approval)⁵ pour des entrepôts durables et fiables, les données stockées dans l'entrepôt devront être :

- accessibles sur Internet tout en respectant la législation en vigueur par rapport aux données personnelles et à la propriété intellectuelle ;
- disponibles dans des formats utilisables ;
- fiables ;
- citables (identifiées avec des identifiants pérennes tels que les DOI)

L'entrepôt institutionnel doit s'inscrire dans l'écosystème INRA et hors INRA qui diffère d'une famille de données à une autre : des entrepôts nationaux ou internationaux existent mais atteignent leurs limites pour certaines familles et d'autres sont sous-pourvues.

Pour les familles sous-pourvues, il est nécessaire de créer un entrepôt où les chercheurs viendront déposer leurs données. Pour les autres familles, d'autres systèmes sont probablement à réfléchir : prévoir un dépôt automatisé des données depuis les plateformes d'origine vers l'entrepôt de données institutionnel (cas des données liées à des publications), inscrire l'entrepôt comme un maillon d'une infrastructure internationale, etc.

⁵ <http://datasealofapproval.org/en/information/guidelines/>

Le groupe de travail attire l'attention sur l'importance de mener une réflexion approfondie sur l'architecture technique de l'entrepôt : il est probable que celui-ci ne soit pas unique et centralisé mais composé de plusieurs nœuds (thématiques ?) interopérables.

Ces questions sont à étudier plus précisément lors de la phase de mise en œuvre de l'entrepôt institutionnel.

2.1.5 Opportunités et risques d'un entrepôt institutionnel

D'une manière générale, la mise en place d'un entrepôt de données à l'Inra présente les risques et les opportunités présentés dans le tableau ci-dessous.

	Aspects positifs/opportunités	Aspects négatifs/risques
Pour l'INRA	<ul style="list-style-type: none"> • Permettre aux chercheurs de satisfaire les exigences des agences de financement en termes de gestion, de partage et de préservation des données • Prendre en compte les besoins spécifiques à l'INRA (chercheurs, direction, outils expérimentaux, plateformes) • Garantir un niveau minimal de qualité des données partagées sous le « label INRA » • Identification facilitée du patrimoine de l'INRA • Permettre la découverte et l'accès aux données à tous les acteurs de la recherche. • Servir les domaines/disciplines « sous-dotés » • Assurer la pérennité des données identifiées par un DOI INRA 	<ul style="list-style-type: none"> • Le niveau minimum de qualité exigé pour les données ne doit pas rendre le dépôt trop contraignant • En fonction du niveau de service souhaité, les moyens humains, matériels, financiers doivent être en cohérence. • L'entrepôt pourrait ne pas satisfaire toutes les spécificités de chaque famille de données.
Vis-à-vis de l'écosystème international	<ul style="list-style-type: none"> • Plus de poids pour discuter avec les éditeurs (notamment concernant l'hébergement des données INRA liées à des publications dans le respect des droits et attentes des chercheurs INRA) • Plus d'outils de négociations avec des partenaires dans le cas par exemple de montage de programmes européens • Plus de facilité pour des dépôts dans des entrepôts internationaux • Viser la mise à disposition de toutes les données associées aux publications. 	<ul style="list-style-type: none"> • Pour les chercheurs qui doivent de toute manière déposer leurs données dans des entrepôts hors Inra (entrepôts thématiques, entrepôts éditeurs, projets internationaux, etc.), il est important d'éviter le multi-dépôt.

2.2 Caractérisation de l'entrepôt institutionnel

2.2.1 Périmètre des données ayant vocation à être stockées dans l'entrepôt

Le but de ce chapitre consiste à définir de manière plus précise les données qui ont vocation à être stockées dans l'entrepôt. Le choix du périmètre aura des conséquences, notamment sur la politique de préservation à mettre en place et les coûts associés. Nous détaillerons ces points dans la suite de ce document.

2.2.1.1 *Respect des règles de propriété intellectuelle et du cadre juridique*

Les données qui seront déposées dans l'entrepôt à des fins de partage doivent respecter les règles éthiques, juridiques et de propriété intellectuelle. Ces règles dépendent notamment de la nature des données, des méthodes d'obtention des données, de la discipline scientifique concernée (anonymisation statistique ...) ...

Le groupe de travail recommande que l'entrepôt de données suive les recommandations qui seront faites par le groupe « Propriété intellectuelle/cadre juridique ».

2.2.1.2 *Nature des données*

L'entrepôt de données acceptera des données **numériques** de toutes natures (un jeu de données⁶, une image, un document, un document audio, une ontologie, un thésaurus, etc.) à condition que l'on souhaite les **partagées**. En effet, l'entrepôt de données institutionnel a vocation à accueillir des données concernées par le mouvement Open Science c'est-à-dire des données que l'on souhaite **partager au-delà des partenaires initiaux**. Pour simplifier la gestion de données par les déposants, le groupe de travail propose d'autoriser le dépôt de données en début de projet. Cependant, ces projets devront disposer d'un plan de gestion des données précis indiquant les conditions de partage de ces données et notamment une échéance (par exemple : partage des données 3 ans après la fin du projet).

De plus, ces données devront être **validées** et **documentées** (c'est-à-dire accompagnées de métadonnées). C'est le cas des données liées à des publications ou des données provenant des plateformes « Open Science ». Il faudra envisager des règles de validation au sein de l'entrepôt pour les autres cas. Il faudra aussi réfléchir à une documentation/des métadonnées minimales à exiger (des premières propositions sont faites dans un prochain paragraphe).

Il est important d'assurer un niveau de qualité suffisant aux données stockées sur l'entrepôt et permettre ainsi aux chercheurs déposant ou consultant des données d'avoir confiance dans l'entrepôt. Ceci est d'autant plus vrai que l'entrepôt constituera à terme une vitrine du patrimoine de l'INRA.

2.2.1.3 *Granularité des données*

Il n'y a a priori aucune restriction sur le niveau de granularité des données qu'il serait possible de stocker dans l'entrepôt. Cependant, le choix du niveau de granularité est une question complexe : ce grain peut varier d'une famille de données à une autre notamment pour respecter le cadre juridique et les règles du secret statistique.

Un accompagnement sera doit être mis en place pour aider les déposants à choisir le niveau de granularité.

⁶ Un ensemble de fichiers contenant chacun des données de la recherche

2.2.1.4 Maturités des données

Afin de favoriser le dépôt et le partage de données, le groupe recommande de permettre un dépôt de données au plus tôt dans le cycle de vie d'un projet à condition qu'un plan de gestion des données précis indiquant les conditions de partage soit établi. Bien qu'il ne soit pas souhaitable de restreindre l'entrepôt en fonction du niveau de maturité des données, il conviendrait d'envisager plusieurs niveaux de services :

- Un niveau « basique » (stockage informatique sécurisé) pour les données qui ne sont pas encore partagées au-delà des partenaires initiaux (par exemple, pour des données en début de projet). En contrepartie, le déposant fournira suffisamment de métadonnées pour que son jeu de données soit identifiable.
- Un niveau de service « élevé » assurant la pérennisation et l'archivage à long terme des données partagées et incluant leur identification par un DOI.

2.2.1.5 Conditions d'accès aux données de l'entrepôt

L'entrepôt de données laissera la liberté au déposant dans la définition et l'attribution des conditions d'accès à ses données. Les données partagées pourront ainsi être :

- En accès libre ou à un groupe restreint (au-delà des partenaires initiaux)
- En accès gratuit ou payant (en précisant les conditions de paiement).

2.2.2 La politique de stockage et d'archivage à long terme (préservation)

2.2.2.1 Enjeux

La politique de stockage et d'archivage à long terme est un des enjeux majeurs de l'entrepôt. Les choix faits auront des répercussions directes sur les moyens techniques, humains et financiers dont l'entrepôt devra être doté. Ces choix auront aussi des conséquences sur le positionnement stratégique de l'institut et sur les possibilités offertes aux chercheurs. Par exemple, un stockage de certaines données sur plusieurs années permettrait de nouvelles recherches tirant profit de la profondeur historique de la source ou de son aspect longitudinal. Cela est notamment le cas en sciences humaines et sociales, par exemple, pour le suivi d'individus dans le temps.

2.2.2.2 Cycle de vie des données de la recherche et premières propositions de service

La durée de vie des données est souvent supérieure à celle des projets qui leur ont donné naissance comme le montre le schéma ci-dessous (figure 1). Bien organisées, documentées, préservées et partagées, ces données peuvent générer de nouvelles opportunités pour la recherche et l'innovation.

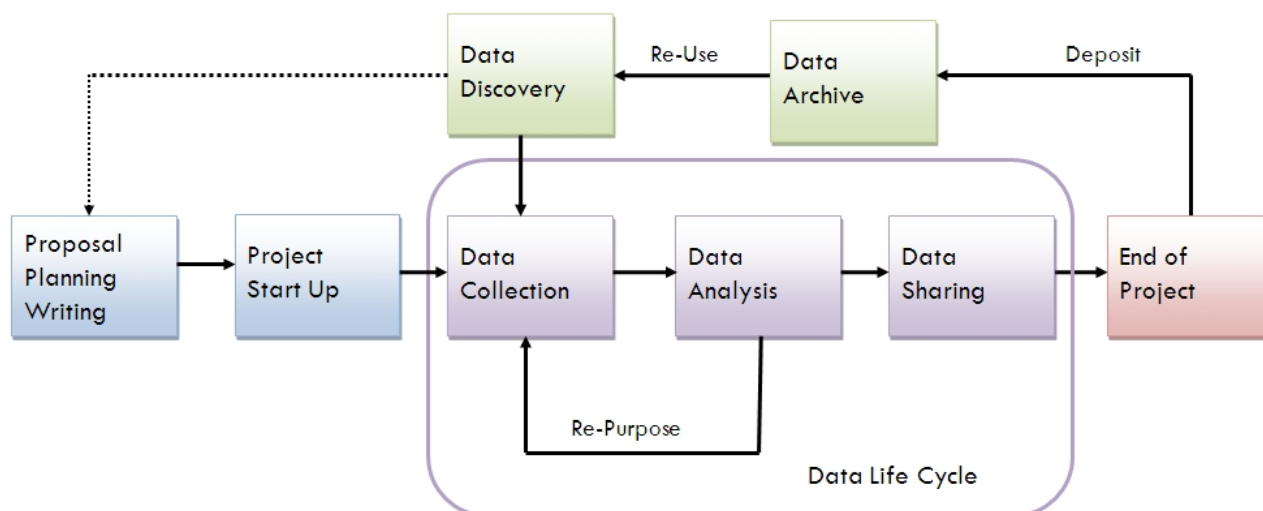


Figure 1 : Représentation du cycle de vie des données de la recherche, University of Virginia Library⁷

La Figure 1 ci-dessous, réalisée par le TGE Adonis, présente le cycle de vie des données de la recherche et les problématiques de stockage et d'archivage.

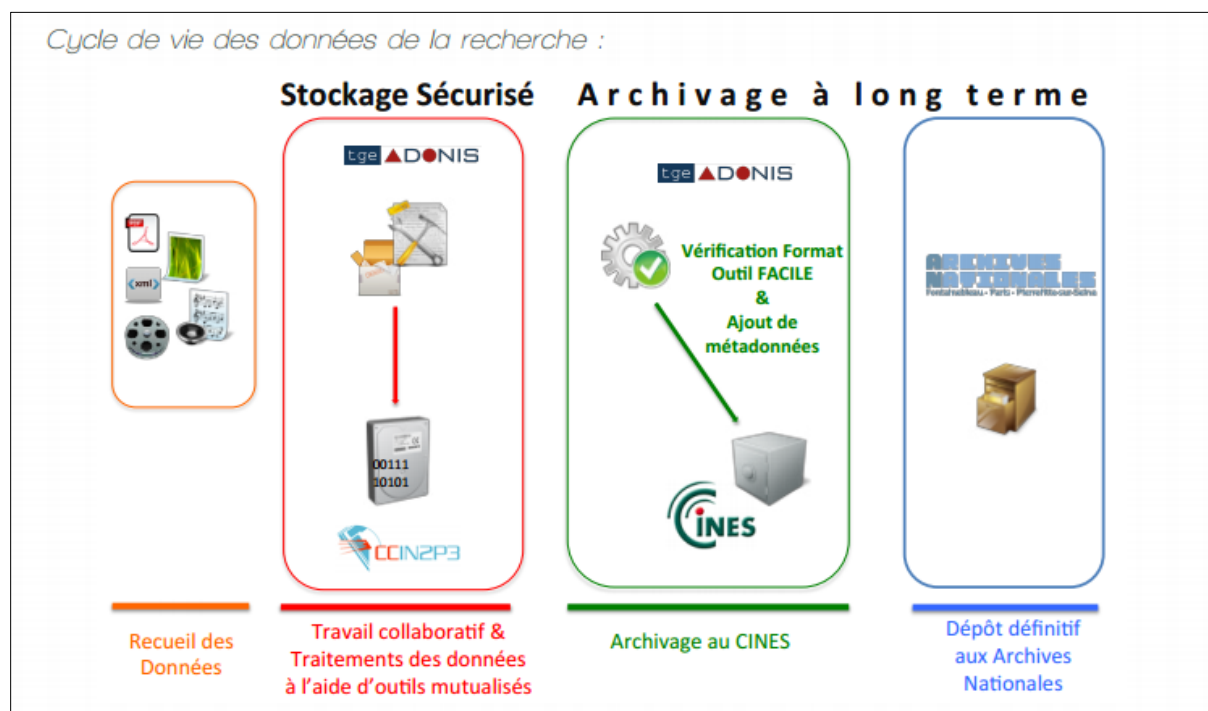


Figure 1 Cycle de vie des données de la recherche, point de vue du TGE Adonis

Après avoir analysé différentes stratégies dont celle du TGE ADONIS⁸ et celle de PURR⁹ (Perdue University Research Repository), le groupe de travail préconise que l'entrepôt propose deux niveaux de service « basique » et « élevé » décrits dans le tableau ci-après.

⁷ <http://dmconsult.library.virginia.edu/lifecycle/>

⁸ <http://www.huma-num.fr/sites/default/files/ressourcesdoc/la-lettre-fev-mars2013.pdf>

⁹ <https://purrr.purdue.edu/legal/preservation-strategies>

Niveaux de service	<p>Niveau « basique » : stockage sécurisé des données qui consiste en des activités de sauvegarde (backup) et de vérification d'intégrité (checksum). Il ne permet pas de partager (au sens Open Science) des données ni de les pérenniser.</p> <p>Niveau « élevé » : archivage à long terme ou préservation des données. En plus des activités du niveau basique, des activités de migration et une veille technologique pour détecter les changements et les obsolescences au niveau des formats et des technologies est assurée. C'est le niveau de service offert aux données partagées.</p>
Données concernées par chaque niveau de service	<p>Niveau « basique » : données en début de projet ou n'ayant pas encore rempli les conditions nécessaires à leur partage conformément au plan de gestion de données, aux règles éthiques, juridiques et de propriété intellectuelle.</p> <p>Niveau « élevé » : données partagées ou partageables sans délai, identifiées à l'aide d'un identifiant pérenne (DOI...). Ces données doivent être disponibles sous des formats éligibles à la préservation.</p>

2.2.2.3 Archivage à long terme et problématiques des formats de fichiers

Lorsque l'on archive à long terme une donnée, une des problématiques est d'être capable de lire cette donnée plusieurs années après son dépôt. Les risques auxquels il faut faire face sont de plusieurs natures dont notamment :

- La détérioration des supports physiques,
- La perte d'information décrivant le contenu,
- L'obsolescence des formats de fichier,
- La disparition des logiciels ou matériels de lecture.

Le groupe de travail a mené une première réflexion sur la problématique des formats de fichiers et propose quelques recommandations. Cette réflexion sera à reprendre et poursuivre lors de la phase de mise en œuvre de l'entrepôt avec les spécialistes du domaine (scientifiques, ingénieurs systèmes, archivistes, etc.).

Critères d'éligibilité des formats à la préservation (archivage long terme)	<p>Les formats éligibles à l'archivage à long terme doivent :</p> <ul style="list-style-type: none"> • Être libres de droits • Avoir des spécifications accessibles • Être supportés par un grand nombre de plateformes • Être largement adoptés
Classement des formats	<p>Les formats de fichiers pourront être classés comme suit pour faciliter leur suivi :</p> <ul style="list-style-type: none"> • A l'étude : formats en cours de vérification • Potentiellement préservables : formats qui répondent positivement aux critères d'éligibilité ci-dessus mentionnés, mais qui ne sont pas effectivement pris en charge soit par manque de besoin, soit par limitation en raison des charges de maintenance. • Préservables : formats répondant aux critères d'éligibilité et effectivement pris en charge. • Formats acceptables : formats ne répondant pas à tous les critères d'éligibilité mais qui sont pris en charge pour diverses raisons (large adoption, absence d'alternatives, etc.) • En voie d'obsolescence : formats candidats à la migration vers un nouveau format. • Obsolescents : formats inexploitable.

En annexe 1, vous trouverez une liste initiale de formats préservables et des formats acceptables.

2.2.2.4 *Durée de préservation*

- Pour les données partagées, concernées par le niveau de service « élevé » :

Le groupe recommande d'identifier à l'aide d'un DOI les jeux de données partagées et stockées dans l'entrepôt. Lorsqu'on identifie un jeu de données à l'aide d'un DOI, on s'engage à assurer sa pérennité pendant 10 ans. En conséquence, le groupe de travail recommande que la durée de préservation soit de 10 ans minimum.

- Pour les données non partagées, concernées par le niveau de service « basique » :

Toutes les données stockées dans l'entrepôt doivent être accompagnées d'un plan de gestion des données indiquant les conditions et à quelle échéance elles seront partagées. Pour éviter l'encombrement de l'entrepôt avec des données non partagées, le groupe de travail propose de mettre en place un système de relance envers les déposants une fois cette échéance passée. Si dans les 3 ans qui suivent la première relance les données n'ont toujours pas été partagées, le groupe recommande leur suppression de l'entrepôt (en prévenant bien évidemment le déposant).

2.2.3 **Système de consultation / recherche**

2.2.3.1 *Le système de recherche*

L'entrepôt de données institutionnel doit disposer d'un système permettant de consulter les données qui y sont stockées, en d'autres termes, il est nécessaire de prévoir un système de recherche de données. Le groupe de travail recommande que ce système de recherche repose sur l'indexation des métadonnées et non sur le contenu des données.

Ce système n'est pas à confondre avec l'annuaire qu'il n'a pas vocation à remplacer.

Remarque : cet entrepôt institutionnel peut se traduire « techniquement » par la mise en œuvre de plusieurs entrepôts. Chaque entrepôt pourrait utiliser des technologies différentes et disposer de son propre système de recherche.

2.2.3.2 *Documentation des données - Les métadonnées*

La documentation des données permet de comprendre (à court, moyen ou long terme) le contenu, le format et le contexte des données collectées ou générées au cours d'un projet de recherche. La documentation facilite également la recherche et la localisation des données. La documentation des données joue par conséquent un rôle important dans leur réutilisation, que ce soit à des fins de reproductibilité, de preuve ou de synthèse (méta-analyse).

Les métadonnées permettent de documenter les données de manière structurée.

On distingue habituellement 3 types principaux de métadonnées :

- les métadonnées descriptives : elles permettent de décrire le contenu et le contexte des jeux de données. Exemples de métadonnées descriptives : titre, auteur, description, résumé, mots, clés, etc.
- les métadonnées techniques ou de structuration : elles regroupent les informations sur le format, les processus, les relations entre les différents fichiers ou parties d'un jeu de données. Exemples de métadonnées techniques d'une donnée de type image : appareil photo, ouverture, format du fichier, paramètres, etc.
- les métadonnées administratives : elles comprennent les informations utiles pour gérer et utiliser les données décrites. Exemples de métadonnées administratives : date de création, droits d'usage, logiciels requis, provenance, contrôles d'intégrité, etc.

Quel que soit leur type (descriptif, technique ou administratif), les métadonnées peuvent être catégorisées suivant leur objectif :

- Métadonnées de découverte permettant la découverte des jeux de données ;
- Métadonnées de contexte permettant de mettre les données en contexte, d'évaluer leur pertinence et leur qualité ;
- Métadonnées d'interopérabilité permettant l'interopérabilité (requêtes homogènes sur des jeux de données non homogènes).

Les métadonnées d'interopérabilité et certaines métadonnées de contexte dépendent des types des données.

Pour garantir une consistance au niveau du contenu et du format des métadonnées, ainsi que l'interopérabilité de l'entrepôt, il convient de se baser sur des standards internationaux. Le groupe de travail recommande de se baser sur le schéma de DataCite et de le compléter au besoin par des métadonnées exigibles par l'Inra et les différentes familles de données.

3 Le chantier Annuaire

3.1 Un annuaire institutionnel : définition, opportunités et risques

3.1.1 Définition

Un annuaire se présente comme un répertoire constitué de rubriques et de sous rubriques et qui recense des liens vers des ressources. Dans le cadre de l'Open Science, ces ressources seront les données partagées. Un annuaire est donc un outil de recherche indexant et classant des données. Il ne stocke pas directement les données qu'il référence. Ces données sont stockées ailleurs (notamment dans l'entrepôt institutionnel).

Le référencement doit s'appuyer sur des standards internationaux partagés. Il doit être conforme aux contraintes et normes en vigueur (à décliner selon les types de données. Exemple : Directive UE INSPIRE pour les données géoréférencées et d'intérêt environnemental).

3.1.2 Opportunités et risques

D'une manière générale, la mise en place d'un annuaire à l'Inra présente les risques et les opportunités présentés dans le tableau ci-dessous.

	Aspects positifs/opportunités	Aspects négatifs/risques
Pour l'INRA	<ul style="list-style-type: none">• Point central pour trouver les données partagées de l'INRA• Identification et lisibilité de la production de « données » de l'institut• Permettre une recherche unifiée	<ul style="list-style-type: none">• Système trop normatif pour les chercheurs• Garder l'annuaire à jour nécessite une veille importante• Mise en œuvre et pérennité de l'alimentation• Importance de maintenir un lien valide et à jour vers les données
Vis-à-vis de l'écosystème international	<ul style="list-style-type: none">• Vitrine du patrimoine INRA	<ul style="list-style-type: none">• Risque d'augmenter les sollicitations des chercheurs quant à l'utilisation de leurs données

3.2 Caractérisation de l'annuaire institutionnel

3.2.1 Périmètre des données ayant vocation à être référencées dans l'annuaire

3.2.1.1 Nature et provenance des données référencées

L'objectif de cet annuaire est de référencer toutes les données **numériques INRA** de toutes natures (un jeu de données¹⁰, une image, un document, un document audio, une ontologie, un thésaurus, etc.) à condition qu'elles soient **partagées**. Par données Inra, nous comprenons données pour lesquelles une structure ou une personne affiliée à l'Inra est auteur/contributeur ou co-auteur.

Pour ce qui concerne les données de l'entrepôt institutionnel, seules les données identifiées par un DOI (données effectivement partagées) seront référencées par l'annuaire pour des raisons évidentes de compétitivité, de confidentialité et de sécurité. Concernant les données Inra stockées dans des entrepôts hors Inra, seules les données ayant des identifiants pérennes (DOI, numéros d'accès Genbank, identifiants PDB, etc.) seront référencées. L'enjeu ici est de s'assurer que l'annuaire pointerait uniquement vers des données pérennes.

L'annuaire référencera toutes les données INRA quel que soit l'endroit où elles sont physiquement stockées (dans l'entrepôt Inra à construire, dans tout autre système d'information Inra, ou hors Inra).

3.2.1.2 Politique de droit d'accès

L'annuaire référencera les données quels que soient leurs conditions d'accès :

- En accès libre ou à un groupe restreint (au-delà des partenaires initiaux)
- En accès gratuit ou payant (en précisant les conditions de paiement).

Les pages descriptives des données préciseront ces conditions d'accès (voir paragraphe suivant).

3.2.2 Présentation des données référencées

Pour chaque jeu de données référencé dans l'annuaire, une page descriptive sera disponible. Cette page est en quelque sorte une carte d'identité du jeu de données. Celle-ci sera construite sur le modèle des landing pages exigées par DataCite pour les données identifiées par des DOI. Cette page descriptive contiendra donc :

- La citation complète du jeu de données ;
- Les métadonnées associées ;
- Les informations concernant l'accès à l'objet scientifique (URL d'accès, conditions d'obtention, restrictions, etc.) ;
- Les informations pour lire l'objet scientifique (logiciels, contexte, autres informations nécessaires à l'interprétation....).

L'annuaire renverra systématiquement vers les entrepôts d'origine où les pages de description des données sont susceptibles d'être plus riches que celles de l'annuaire.

¹⁰ Un ensemble de fichiers contenant chacun des données de la recherche

3.2.3 Alimentation de l'annuaire – Intégration dans l'écosystème Open Data

L'annuaire sera alimenté à la fois manuellement et automatiquement (API, ...). Des interconnexions entre le service d'attribution de DOI, l'entrepôt institutionnel mais aussi des entrepôts et annuaires hors INRA sont à prévoir.

L'annuaire devra ainsi utiliser les standards d'architecture et de protocoles existant (exemples : REST, OAI-PMH, SPARQL) pour être interopérable avec l'écosystème existant.

3.2.4 Système de consultation/recherche

L'annuaire institutionnel doit disposer d'un système permettant de rechercher les données qui y sont référencées. Ce système de recherche s'appuiera sur les métadonnées associées aux données référencées. Son efficacité dépend donc de la richesse des métadonnées associées aux données référencées.

Pour faciliter l'indexation des données et donc la recherche future, il est important de coordonner les schémas de métadonnées utilisés pour l'entrepôt et l'annuaire institutionnels. Le groupe de travail recommande comme pour l'entrepôt de données de se baser sur des standards internationaux tels que le schéma de DataCite.

Deux niveaux de recherche sont ainsi envisageables : un premier niveau basé sur les métadonnées de découverte, et un deuxième niveau basé sur les métadonnées de contexte et d'interopérabilité (spécifiques aux différents types de données).

Annexe

1 Liste initiale des formats préservables et acceptables

Le tableau ci-dessous propose une liste initiale des formats préservables et des formats acceptables. Cette liste est appelée être complétée et validée par les différentes familles de données et les informaticiens système, puis faire l'objet de mises à jour régulières sur un mode collaboratif.

Pour certains formats propriétaires, il sera peut être nécessaire de stocker/préserver le logiciel associé pour permettre la lecture des données dans les années à venir.

Ce tableau est dérivé de l'analyse de plusieurs recommandations (exemples : celles de l'université d'Edinburgh¹¹, du TGE ADONIS et de PURR) et du retour des groupes de travail Inra sur les familles de données. Les déposants seront encouragés à convertir leurs données dans un des formats préservables à chaque fois que cela est possible.

Types de données	Préservables	Acceptables
Données quantitatives tabulaires	<ul style="list-style-type: none"> • SPSS format portable (.por) • Fichiers texte .CSV, .tab ou délimités avec un caractère donné (ce caractère doit être absent des données elles-mêmes) • Fichiers de commandes ('setup') contenant des métadonnées : SPSS, Stata, SAS • Fichiers texte structurés ou balisés contenant des métadonnées : DDI XML file 	<ul style="list-style-type: none"> • Formats propriétaires de logiciels de statistiques : SPSS (.sav), Stata (.dta) • Formats populaires : MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf) and OpenDocument Spreadsheet (.ods)
Données géographiques Données vectorielle ou matricielles (raster)	<ul style="list-style-type: none"> • ESRI Shapefile (essential - .shp, .shx, .dbf, optional - .prj, .sbx, .sbn) • geo-referenced TIFF (.tif, .tiff) • Données CAD (.dwg) • Données tabulaires d'attributs GIS 	<ul style="list-style-type: none"> • ESRI Geodatabase format (.mdb) • Format d'échange MapInfo (.mif) • Keyhole Mark-up Language (KML) (.kml) • Adobe Illustrator (.ai), données CAD (.dxf ou .svg) • Formats binaires des logiciels GIS et CAD

¹¹ <http://www.data-archive.ac.uk/create-manage/format/formats-table>

Types de données	Préservables	Acceptables
Données qualitatives Données textuelles	<ul style="list-style-type: none"> eXtensible Mark-up Language (XML) conforme à un Document Type Definition (DTD) ou un schema (.xml) Rich Text Format (.rtf) Données plein texte, ASCII (.txt) 	<ul style="list-style-type: none"> Hypertext Mark-up Language (HTML) (.html) Formats populaires : MS Word (.doc/.docx)
Données numériques images	TIFF version 6 non compressé (.tif)	<ul style="list-style-type: none"> JPEG (.jpeg, .jpg) uniquement s'il s'agit du format d'origine. TIFF (autres versions que le 6 non compressé) (.tif, .tiff) Adobe Portable Document Format (PDF/A, PDF) (.pdf) standard applicable RAW image format (.raw) Photoshop files (.psd)
Données audio numériques	Free Lossless Audio Codec (FLAC) (.flac)	<ul style="list-style-type: none"> MPEG-1 Audio Layer 3 (.mp3) uniquement s'il s'agit du format d'origine Audio Interchange File Format (AIFF) (.aif) Waveform Audio Format (WAV) (.wav)
Données vidéo numériques	<ul style="list-style-type: none"> MPEG-4 (.mp4) motion JPEG 2000 (.mj2) 	
Documents et scripts	<ul style="list-style-type: none"> Rich Text Format (.rtf) PDF/A or PDF (.pdf) HTML (.htm) OpenDocument Text (.odt) 	<ul style="list-style-type: none"> plein texte (.txt) Formats populaires : MS Word (.doc/.docx) ou MS Excel (.xls/.xlsx) XML marked-up text (.xml) conforme à un DTD ou un schéma XML
Séquences lues ADN-ARN	fastq, sff	
Géotypes SNP	Illumina (matrice marque*organisme) + métadonnées en en-tête, VCF	
Géotypes SSR	CSV - Excel (GeneMapper)	
Données d'expression (arrays, qPCR)	MIAME	
Profils protéiques (quantitatifs)	EC number, SBML	
Séquences protéines		

Types de données	Préservables	Acceptables
Séquences alignées/assemblées	fasta, asn, embl	
Données d'expression RNAseq	fasta, bam, sam	
Polymorphismes (SNP)	bam, sam, sff, bed, wig	
Polymorphismes SSR	VCF, fasta	
Variants structuraux	gff3	
Patrons de méthylation	VCF (V4.1+)	
Annotation des gènes	bed	
Orthologues, paralogues, familles de gènes	gff3, asn, embl	
Cartes (génétiques, QTLs, physiques)	tables xml	
Données passeport populations/souches	acp, text	
Données passeport croisements temporaires		
Données passeport banques génomiques		