



Cycle de vie des données : stockage sauvegarde conservation à long terme des données

Journée Archivage numérique des données de recherche

20 NOVEMBRE 2019 – AUDITORIUM IMAG GRENOBLE

CNRS – Inist/DVDR/Service Formation-DoRANum





CYCLE DE VIE DES DONNÉES

CYCLE DE VIE DES DONNÉES DE RECHERCHE

C'est l'ensemble des étapes

- de gestion,
- de conservation
- et de diffusion des données de recherche,

associées aux activités de recherche



Il n'y a pas de début et de fin distincts à la gestion des données. C'est pourquoi le cycle de vie des données est représenté sous forme de cercle plutôt que de ligne droite.

D'après Research data lifecycle – UK Data Service
<https://www.ukdataservice.ac.uk/manage-data/lifecycle>

DÉFINITION ET DIVERSITÉ DES DONNÉES DE RECHERCHE

Les données de recherche sont « l'ensemble des informations, spécimens et matériaux produits, recueillis et documentés par les chercheurs, et qui sont collectées et exploitées à des fins de recherche et de preuve par les chercheurs et leurs équipes. »

Définition des archivistes de la Section AURORE de l'AAF*

Les données de recherche peuvent être :

- **produites**, lors de campagnes de recherche (observations, mesures...)
- **collectées** : données déjà existantes (corpus, archives...)

DÉFINITION ET DIVERSITÉ DES DONNÉES DE RECHERCHE

Données d'observation

- capturées en temps réel
- habituellement uniques, impossible à reproduire

Relevés météo, images
Enquêtes sociales
Fouilles archéologiques



Données expérimentales

- obtenues à partir d'équipements de laboratoire
- souvent reproductibles, parfois coûteuses

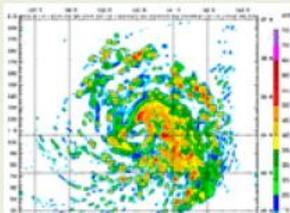
Poids biomasse,
Séquence peptide



Données de simulation numérique

- générées par des modèles informatiques
- souvent reproductibles si le modèle est correctement documenté

Modèle climatique
Modèle économique



[Wikimedia, CC-BY-SA 3.0](#)

Données dérivées ou compilées

- issues du traitement ou de la combinaison de données "brutes"
- souvent reproductibles mais coûteuses

Base de données compilées
Fouille de texte



[Heiti Paves, CC-BY-SA 3.0](#)

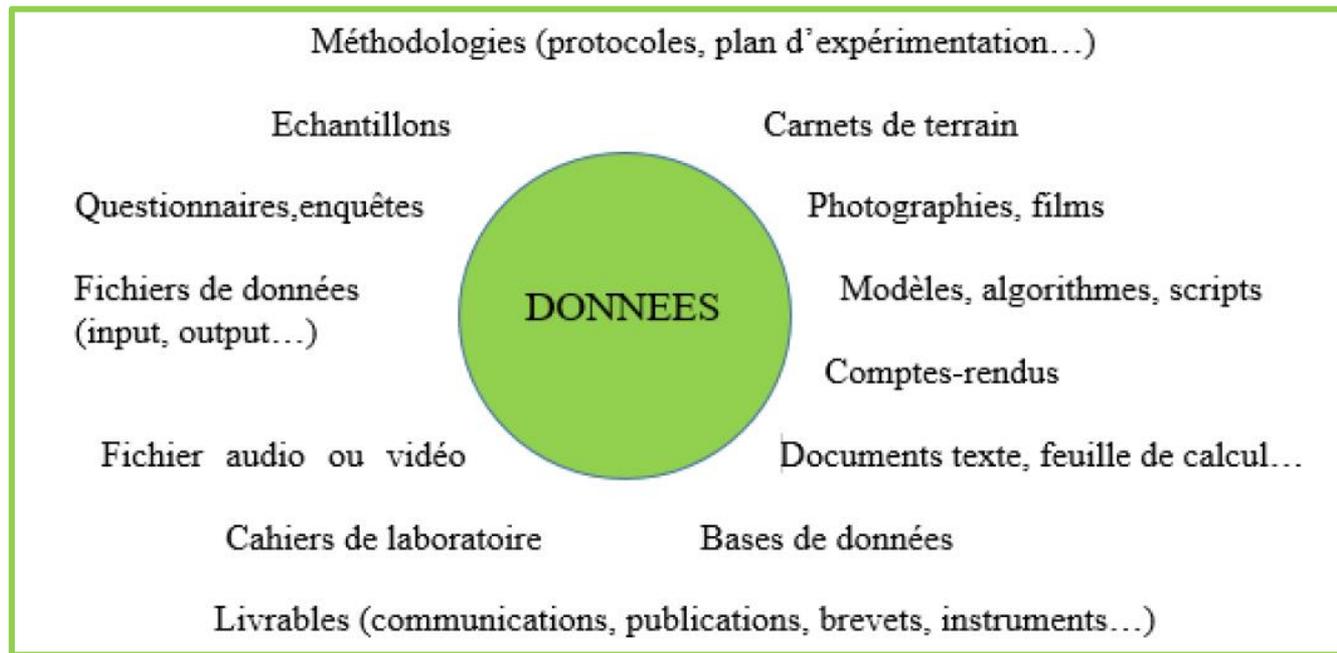
Données de référence

Séquence gènes ,TP53, Structures chimiques



(Source : Gaillard, 2014)

DÉFINITION ET DIVERSITÉ DES DONNÉES DE RECHERCHE

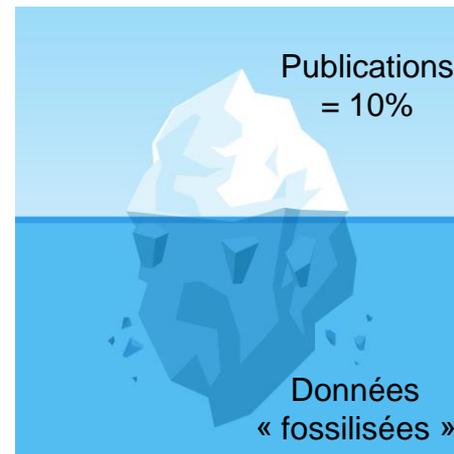


Ressources :

Alain Rivet, Marie-Laure Bachèlerie, Auriane Denis-Meyere et Delphine Tisserand - Traçabilité des activités de recherche et gestion des connaissances – Guide pratique de mise en place – 2018 - http://qualite-en-recherche.cnrs.fr/IMG/pdf/guide_tracabilite_activites_recherche_gestion_connaissances.pdf

POURQUOI GÉRER ET PRÉSERVER SES DONNÉES

- Gestion nécessaire face à l'accroissement de la quantité de données
- Exhumation de données « fossilisées »
- Evite la perte de données uniques
- Gain de temps et d'argent
- Facilite la reproductibilité, **la réutilisation** et le croisement de données provenant de différentes disciplines



"Designed by vvstudio / Freepik"



PLAN DE GESTION DES DONNÉES DE RECHERCHE (PGD OU DMP)

CYCLE DE VIE DES DONNÉES DE RECHERCHE

PLAN DE GESTION DE DONNÉES

DMP



Document synthétique et évolutif qui

- facilite une bonne organisation de ses données tout au long du projet
- décrit la façon dont les données seront obtenues, traitées, organisées, stockées, sécurisées, préservées, partagées...
- garantit des données fiables et bien gérées, compréhensibles, disponibles et préservées sur le long terme



Ressources :

- [DoRanum - Plan de gestion de données](#)
- [Nathalie Reymonet, Magalie Moysan, Aurore Cartier, Renaud Délémontez - Réaliser un plan de gestion de données FAIR](#)

D'après Research data lifecycle – UK Data Service
<https://www.ukdataservice.ac.uk/manage-data/lifecycle>



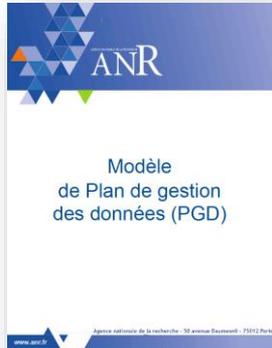
PLAN DE GESTION DES DONNÉES - MODÈLE ANR

EXIGENCES MINIMALES STOCKAGE SAUVEGARDE CONSERVATION LONG TERME



Data Management Plan pour une Optimisation
du Partage et de l'Interopérabilité des
Données de la Recherche
<https://dmp.opidor.fr/>

- Modèle composé de **6 grandes thématiques** illustrant les bonnes pratiques de gestion et de partage :



Établi à partir du guide
Science Europe. (2018). *Practical
guide to the international
alignment of research data
management.*

Description des
données et collecte
ou réutilisation des
données existantes

Documentation et
qualité des données

Stockage et
sauvegarde pendant
le processus de
recherche

Exigences légales et
éthiques, codes de
conduite

Partage des
données et
conservation à long
terme

Responsabilités et
ressources en
matière de gestion
des données

PLAN DE GESTION DES DONNÉES - MODÈLE ANR

EXIGENCES MINIMALES STOCKAGE SAUVEGARDE CONSERVATION LONG TERME

- a. Comment **les données et les métadonnées** seront-elles **stockées et sauvegardées** tout au long du processus de recherche ?
- b. Comment la **sécurité des données et la protection des données sensibles** seront-elles **assurées** tout au long du processus de recherche ?

Stockage et sauvegarde pendant le processus de recherche



- a. Comment et quand les données seront-elles partagées ? Y-a-t-il des **restrictions au partage des données** ou des raisons de définir un embargo ?
- b. Comment les **données à conserver** seront-elles **sélectionnées** et où seront-elles **préservées sur le long terme (entrepôt, archive)** ?
- c. Quelles **méthodes** ou quels **outils logiciels** seront nécessaires **pour accéder et utiliser les données** ?
- d. Comment l'application d'un **identifiant unique et pérenne** (comme le DOI) sera réalisée pour chaque jeu de données ?

Partage des données et conservation à long terme



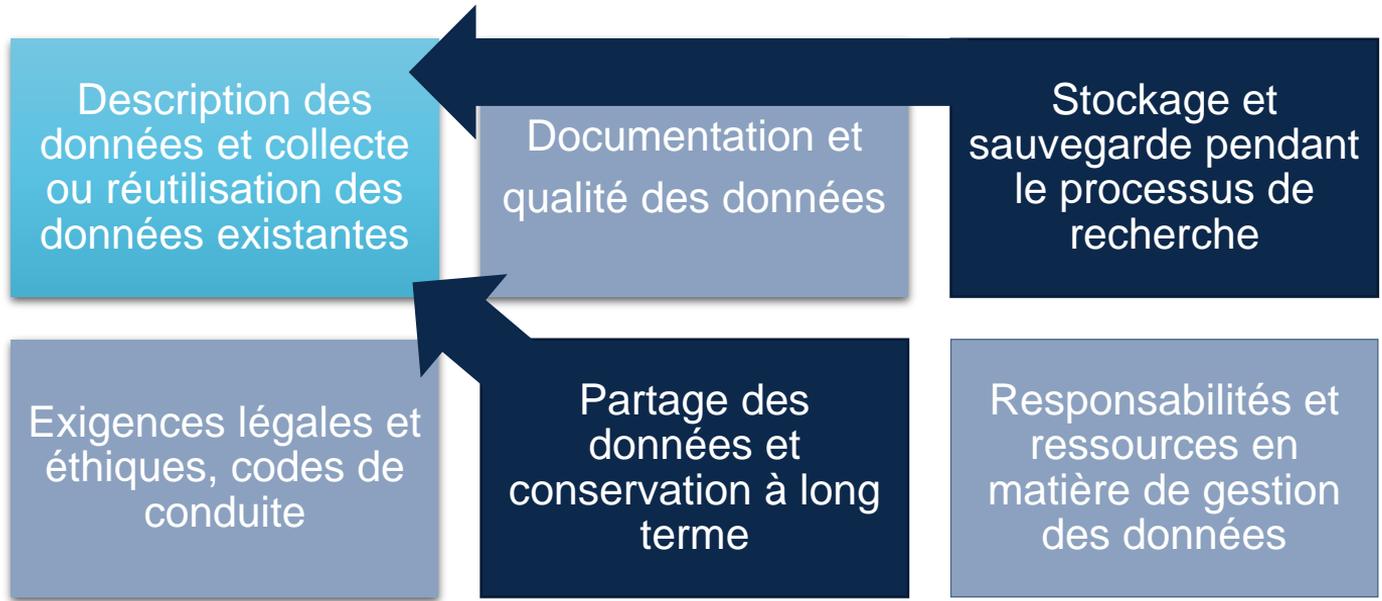
PLAN DE GESTION DES DONNÉES - MODÈLE ANR

EXIGENCES MINIMALES STOCKAGE SAUVEGARDE CONSERVATION LONG TERME

- Modèle composé de **6 grandes thématiques** illustrant les bonnes pratiques de gestion et de partage :



Établi à partir du guide Science Europe. (2018). [Practical guide to the international alignment of research data management.](#)



PLAN DE GESTION DES DONNÉES - MODÈLE ANR

EXIGENCES MINIMALES STOCKAGE SAUVEGARDE CONSERVATION LONG TERME

- Modèle composé de **6 grandes thématiques** illustrant les bonnes pratiques de gestion et de partage :



Établi à partir du guide Science Europe. (2018). [Practical guide to the international alignment of research data management.](#)



PLAN DE GESTION DES DONNÉES - MODÈLE ANR

EXIGENCES MINIMALES STOCKAGE SAUVEGARDE CONSERVATION LONG TERME

- Modèle composé de **6 grandes thématiques** illustrant les bonnes pratiques de gestion et de partage :



Établi à partir du guide Science Europe. (2018). [Practical guide to the international alignment of research data management.](#)



PLAN DE GESTION DES DONNÉES - MODÈLE ANR

EXIGENCES MINIMALES STOCKAGE SAUVEGARDE CONSERVATION LONG TERME



PLAN DE GESTION DES DONNÉES

RESPONSABILITÉS ET RESSOURCES

Ressources

(budget, temps alloués)

En fonction des ressources des institutions, infrastructures...

- Coût matériel
- Frais de stockage
- Coûts des entrepôts de données
- Coût attribution de DOI
- ...

RH (rôle et responsabilités)

Archivistes

Informaticiens



Communauté scientifique
(chercheurs du même domaine)

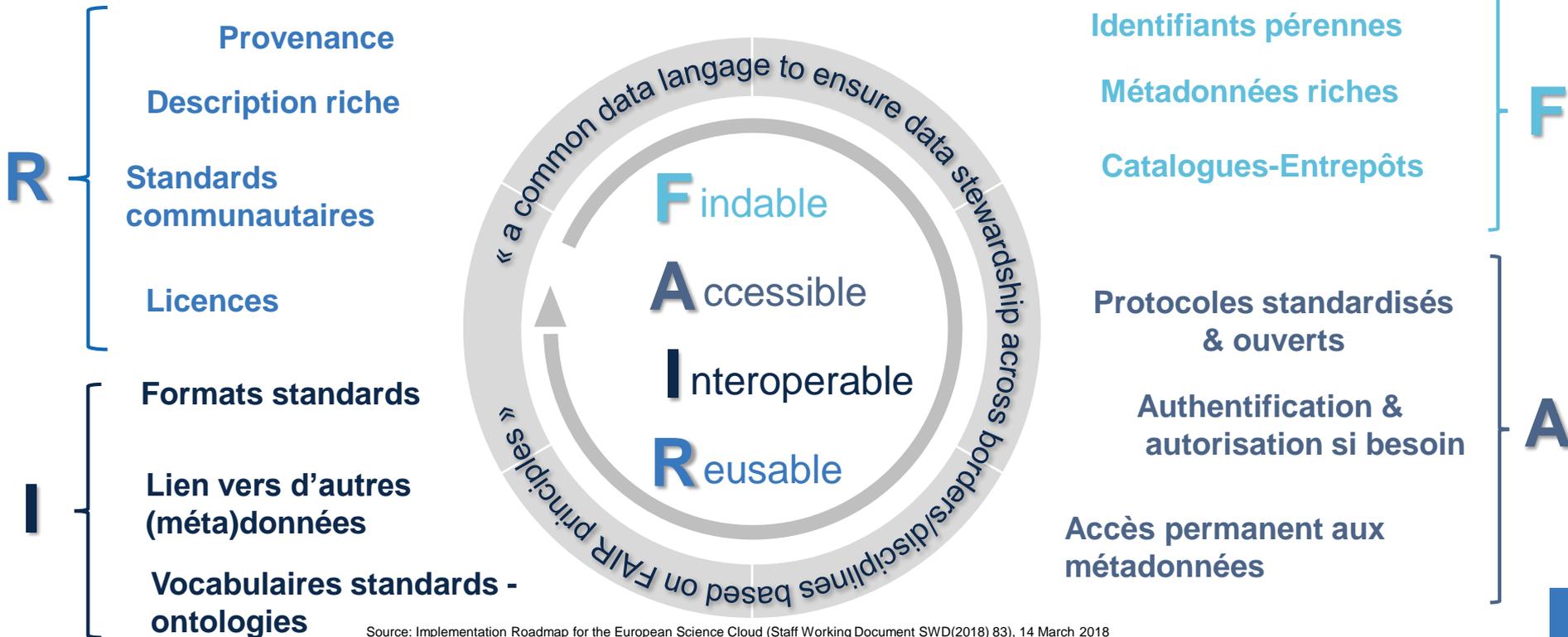
Ingénieur-projet

Services juridiques

Délégué Protection
Données

Spécialiste de l'IST

COMPATIBILITÉ AVEC LES PRINCIPES FAIR



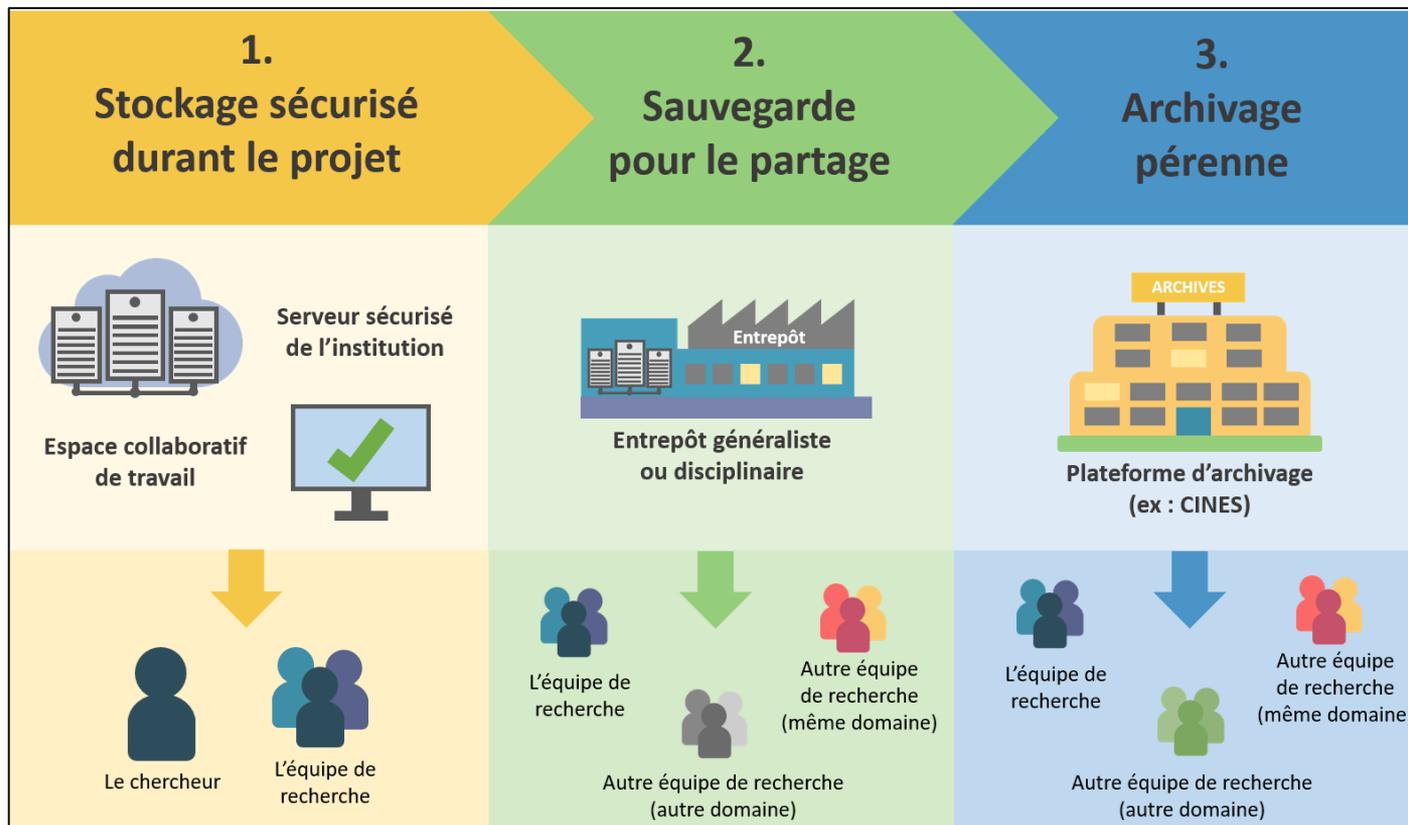
Traduction Inra <https://www6.inra.fr/datapartage/Produire-des-donnees-FAIR>

Des données FAIR plus faciles à partager et réutilisables
par les hommes et par les machines



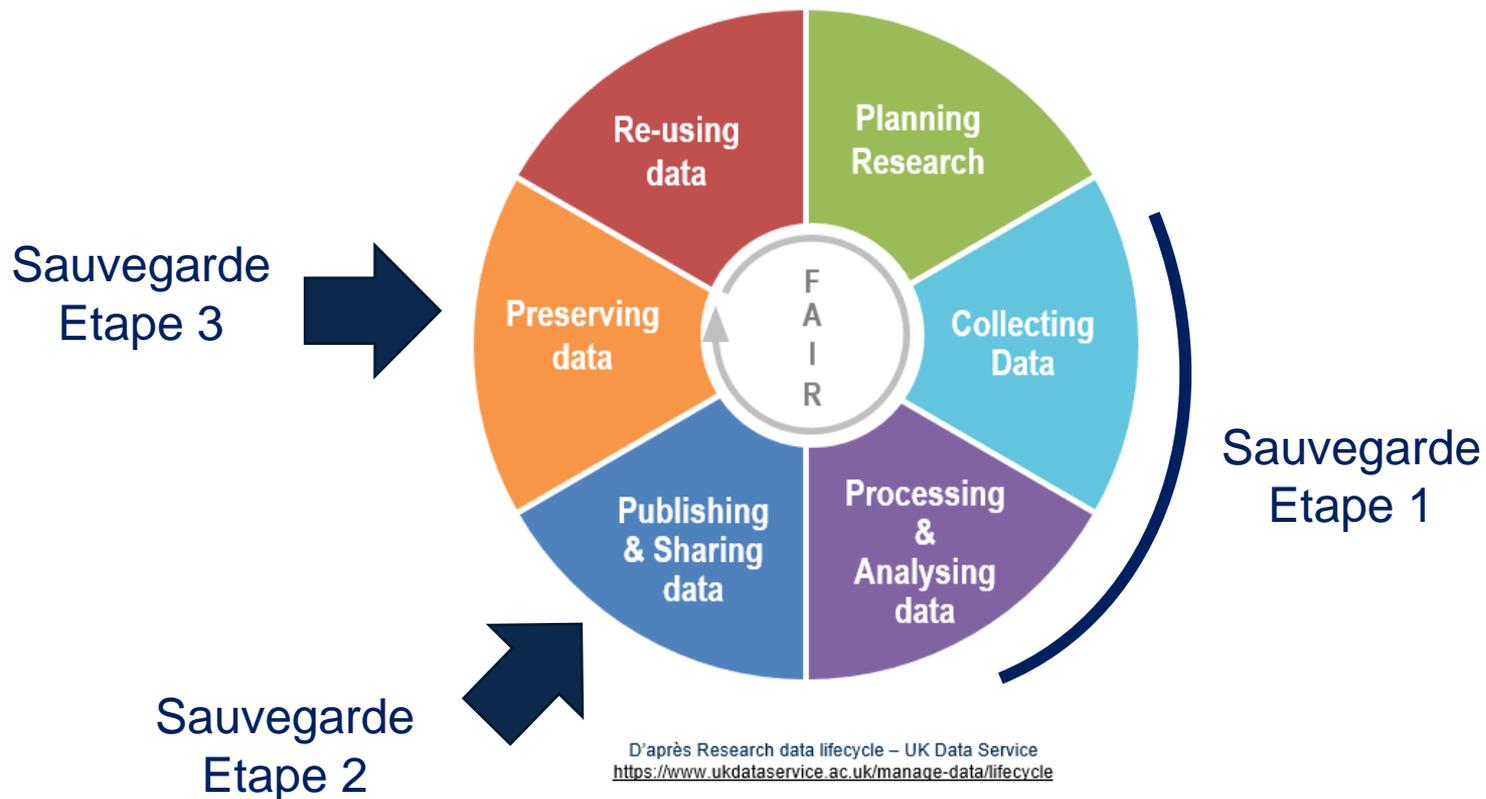
SAUVEGARDE DES DONNÉES TOUT AU LONG DU PROJET

3 ÉTAPES DE SAUVEGARDE DES DONNÉES



CYCLE DE VIE DES DONNÉES DE RECHERCHE

SAUVEGARDE DES DONNÉES



D'après Research data lifecycle – UK Data Service
<https://www.ukdataservice.ac.uk/manage-data/lifecycle>

L'obsolescence technologique : depuis plus de 50 ans, l'évolution rapide des matériels informatiques (nouvelle génération tous les 5 ans en moyenne) engendre une obsolescence des logiciels qui doivent sans cesse rester compatibles avec les nouveaux matériels

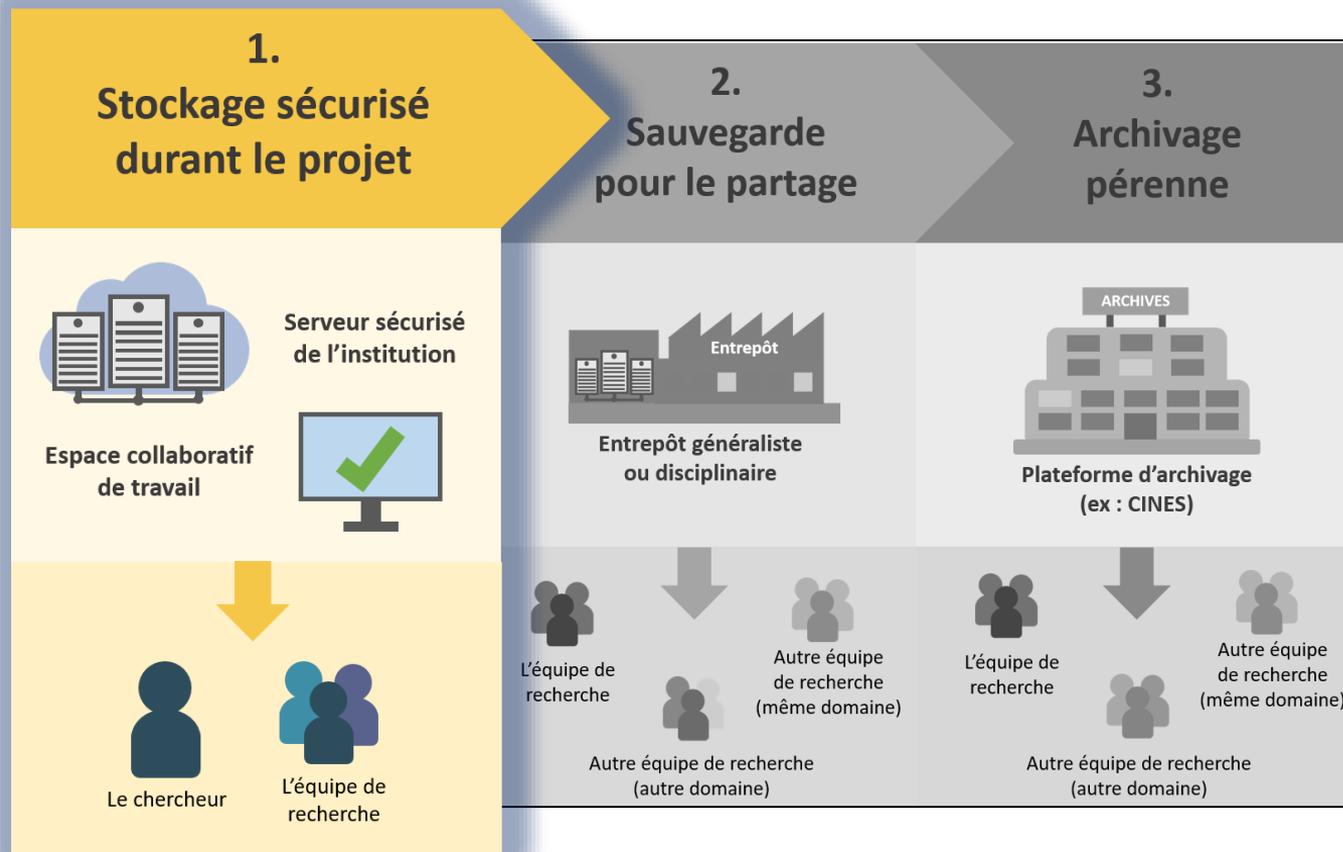
La disparition des formats : on observe en permanence la disparition de formats devenus obsolètes et l'apparition de nouveaux formats d'enregistrement des données

Le manque de documentation : afin de préserver l'intelligibilité des données, il est indispensable de les accompagner de toutes les métadonnées nécessaires à la bonne compréhension du contexte de leur production

ETAPE 1

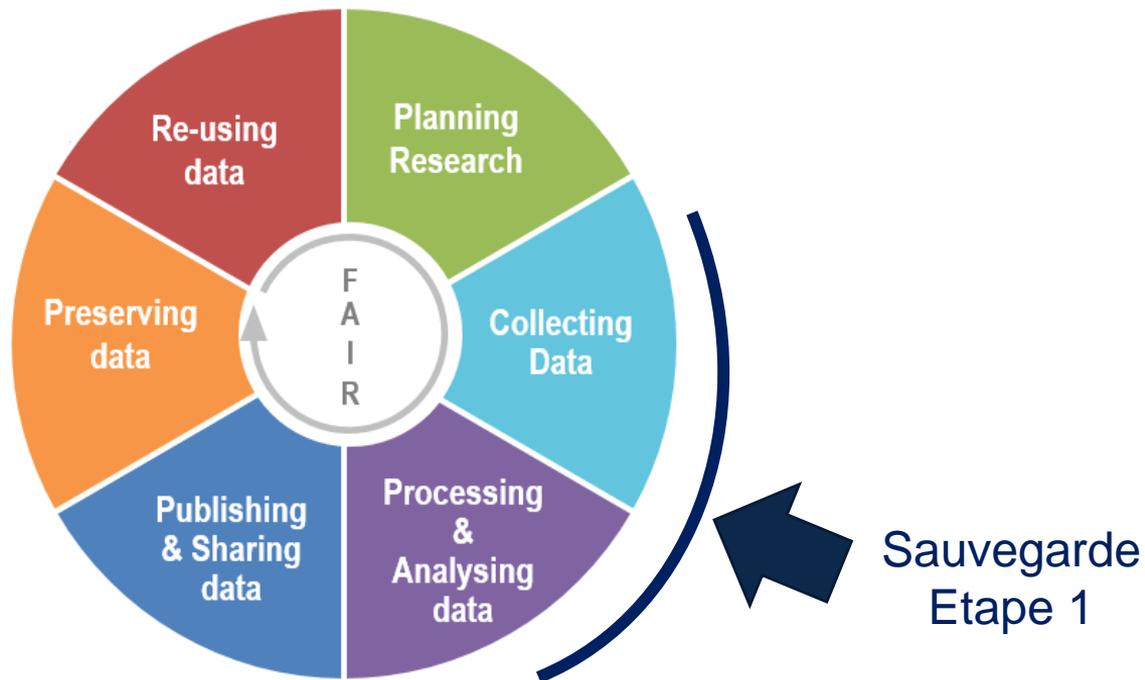
STOCKAGE SÉCURISÉ DURANT LE PROJET

3 ÉTAPES DE SAUVEGARDE DES DONNÉES



CYCLE DE VIE DES DONNÉES DE RECHERCHE

SAUVEGARDE DES DONNÉES



D'après Research data lifecycle – UK Data Service
<https://www.ukdataservice.ac.uk/manage-data/lifecycle>

STOCKAGE SÉCURISÉ DURANT LE PROJET

MESURES DE SAUVEGARDE (STOCKAGE)

- Dans l'idéal, dupliquer et stocker les données à différents endroits sur différents supports

Règle du 3-2-1 :

- garder 3 exemplaires des données,
 - sur 2 supports ou technologies différents,
 - dont 1 se trouve hors site
-
- Organiser et planifier ces sauvegardes
 - Définir la volumétrie des données
 - Définir l'hébergement
 - Gérer les versions

STOCKAGE SÉCURISÉ DURANT LE PROJET

NOMMAGE DES FICHIERS

La fiabilité d'accès passe par un nommage unique et précis des fichiers de données :



Bonnes pratiques

- 30 caractères maximum
- Noms de partenaires insérables si leur graphie est harmonisée entre les fichiers
- Numéros de versions le cas échéant
- Dates au format ISO : AAAA-MM-JJ



A éviter

- Pas de caractères spéciaux ou accentués du type `ùéàç+'@°[] :</>*/ »& !$...`
- Séparateurs : pas d'espace, pas de mots vides, éventuellement Majuscules ou underscore `_`
- Pas de dénomination vague : divers, autres, à classer...

STOCKAGE SÉCURISÉ DURANT LE PROJET

FORMATS DE FICHIERS

Formats ouverts et non propriétaires

Opter pour des formats de fichiers les plus ouverts possible (non propriétaires), standardisés et pérennes

Exemples :

- Privilégier .csv à .xls
- Privilégier .odt à .doc
- Privilégier .jpg à .tif

Choix du format

Le choix d'un format peut être guidé par :

- les recommandations de son institution
- les usages de la communauté scientifique de la discipline
- les logiciels ou équipements utilisés

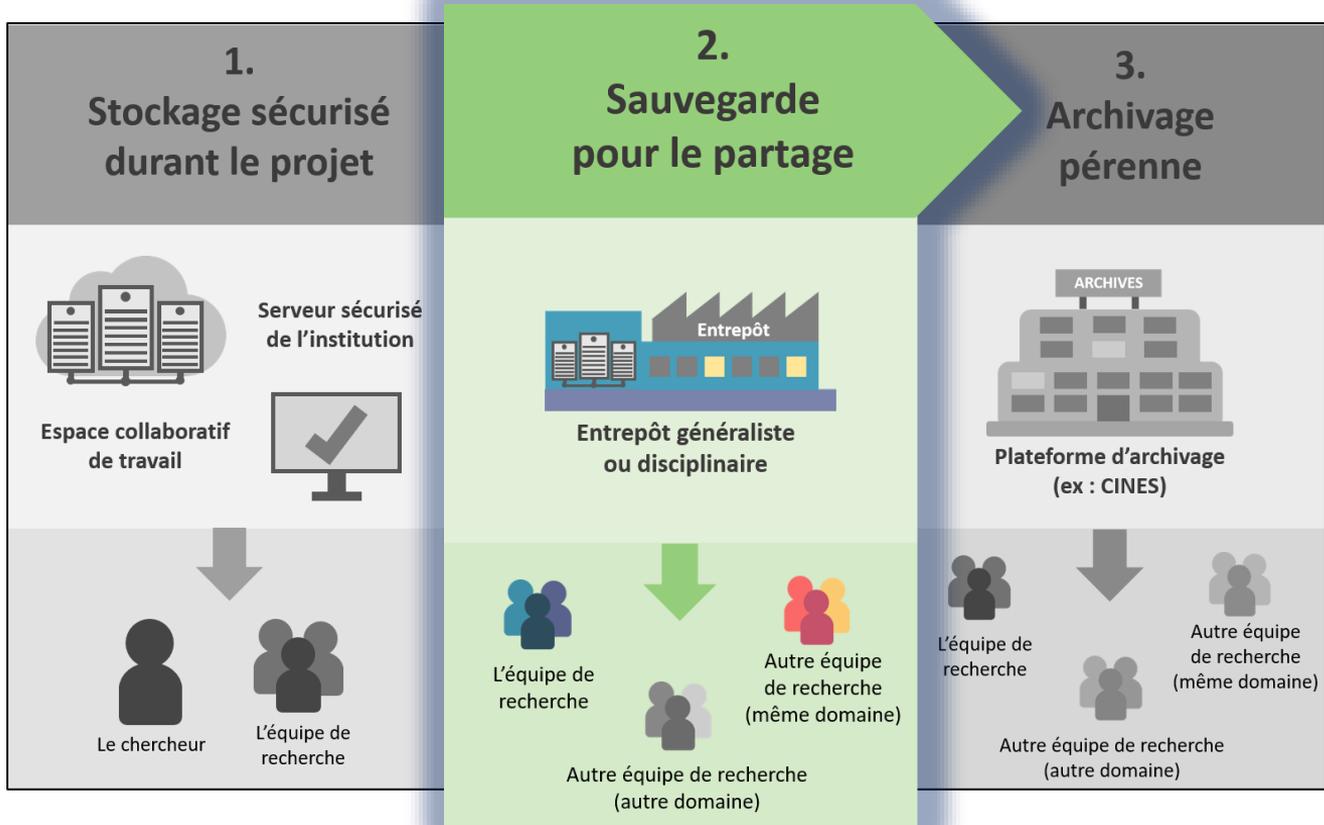
Il faudra le justifier dans le DMP

ETAPE 2

**PARTAGE, DIFFUSION DES
DONNÉES**

**DÉPÔT DANS UN
ENTREPÔT**

3 ÉTAPES DE SAUVEGARDE DES DONNÉES



CYCLE DE VIE DES DONNÉES DE RECHERCHE

SAUVEGARDE DES DONNÉES



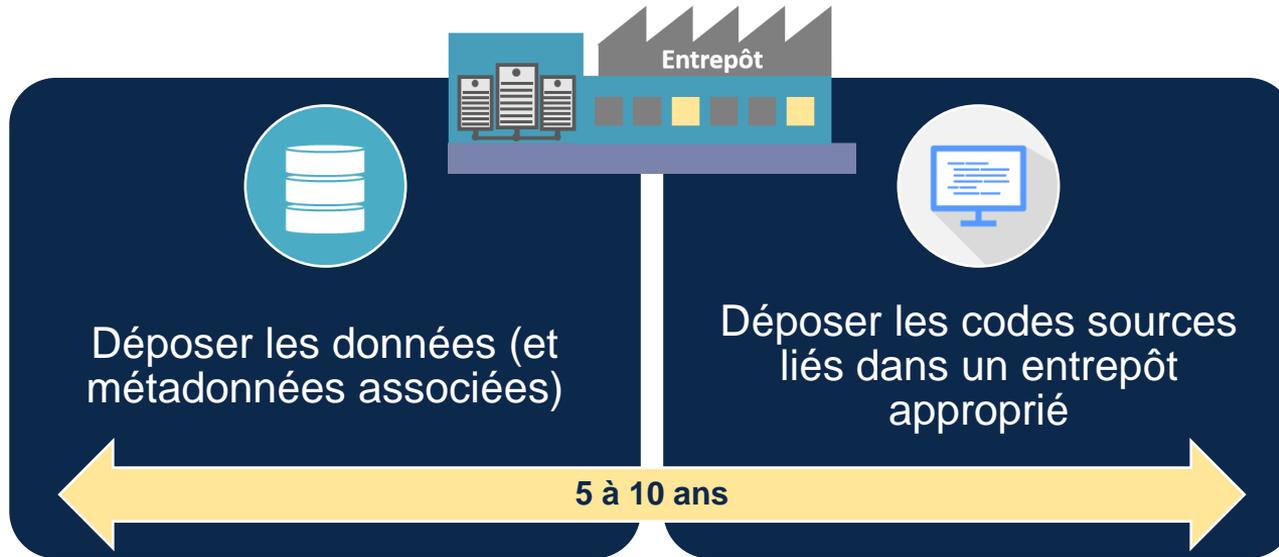
Sauvegarde
Etape 2



D'après Research data lifecycle – UK Data Service
<https://www.ukdataservice.ac.uk/manage-data/lifecycle>

SAUVEGARDE POUR LE PARTAGE DÉPÔT DANS UN ENTREPÔT

- Permet le partage et la réutilisation optimale des données sur le court et le moyen terme (5 à 10 ans)



SAUVEGARDE POUR LE PARTAGE PRÉPARATION DES DONNÉES



Check-list

- Sélectionner les données à partager
- Vérifier la compatibilité et l'interopérabilité des formats de données
- Migrer si nécessaire vers un format adapté, le plus ouvert possible
- Préparer si nécessaire les codes sources (ex : scripts) qui permettront de lire et traiter les données
- Compléter et enrichir les métadonnées (en fonction de l'entrepôt choisi)
 - Si ce n'est pas déjà fait, choisir un standard de métadonnées
 - S'il n'en existe pas d'adapté, créer un schéma de métadonnées
 - Compléter les champs pour chaque jeu de données, suivant le standard adopté

SAUVEGARDE POUR LE PARTAGE

CHOIX DE L'ENTREPÔT

1

- Il est souvent recommandé par son **institution**, son **financeur** ou sa **communauté scientifique**
- Exemples : Data Inra, Nakala

2

- Il est parfois imposé par un **éditeur**
- Exemple : Gene Expression Omnibus

3

- S'il n'y a pas de recommandation, le choisir dans un **annuaire**, en fonction de ses besoins
- Annuaire d'entrepôts : re3data, OAD, OpenDOAR

SAUVEGARDE POUR LE PARTAGE

PRINCIPAUX CRITÈRES DE CHOIX D'UN ENTREPÔT



Discipline / Institution

Type de données acceptées

Qualité des métadonnées

Entrepôt de confiance - Certification

Pérennité des métadonnées et des données

Génération d'un identifiant unique pérenne

Gestion des versions

Gestion des licences

SAUVEGARDE POUR LE PARTAGE EXEMPLES D'ENTREPÔTS



- Entrepôt en Sciences de la Vie, Agronomie, Géosciences, Anthropologie et Sciences comportementales



- Entrepôt en Sciences Humaines et Sociales



- Entrepôt généraliste

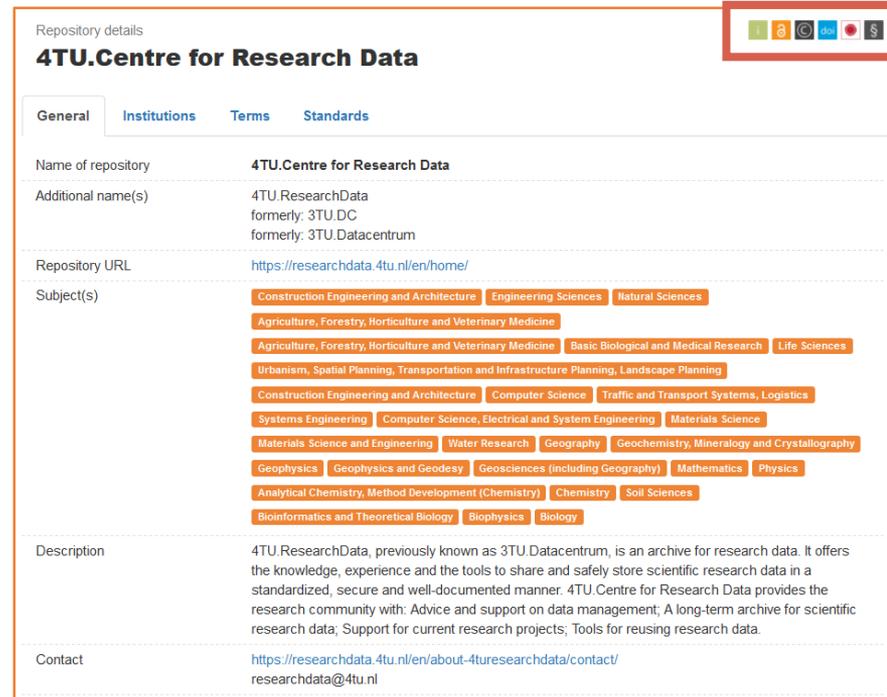
SAUVEGARDE POUR LE PARTAGE RECHERCHE D'ENTREPÔTS

Dans **re3data**, exemple de recherche à partir des critères suivants :

-  Informations complémentaires fournies
-  Libre accès aux données
-  Conditions d'utilisation et licence fournis
-  Génération d'un DOI
-  Certification
-  Politique de l'entrepôt fournie

Plusieurs entrepôts répondent à ces critères :

- [4TU.Centre for Research Data](#)
- [CLARIN repository at the University of Tübingen](#)
- [NASA Socioeconomic Data and Applications Center](#)
- [PANGAEA](#)



Repository details

4TU.Centre for Research Data

General Institutions Terms Standards

Name of repository **4TU.Centre for Research Data**

Additional name(s) 4TU.ResearchData
formerly: 3TU.DC
formerly: 3TU.Datacentrum

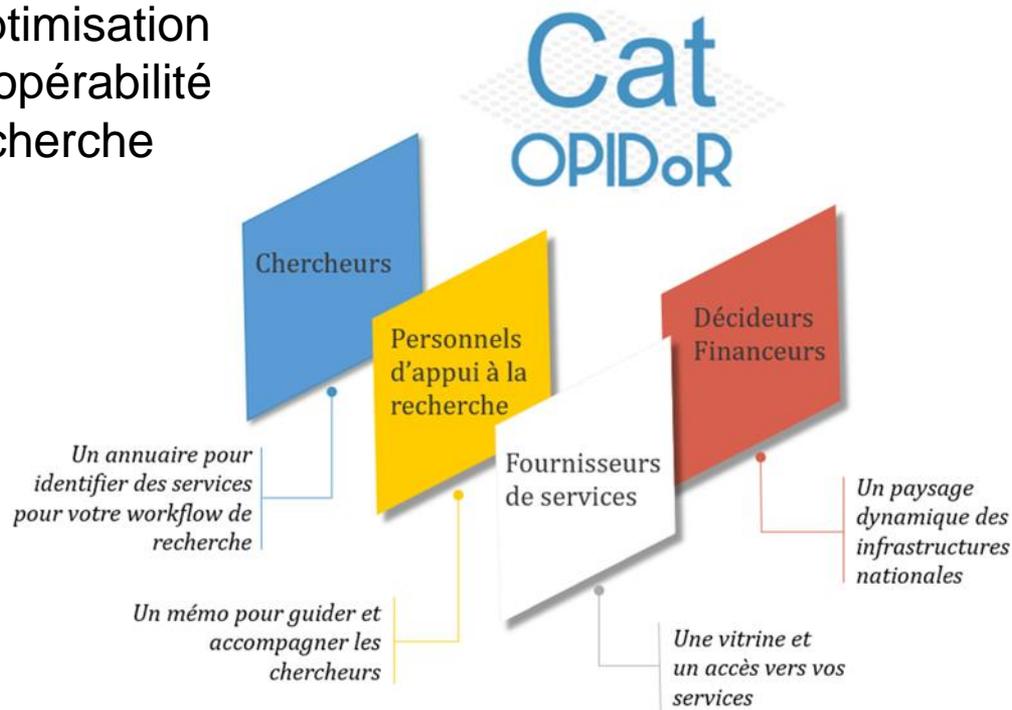
Repository URL <https://researchdata.4tu.nl/en/home/>

Subject(s) **Construction Engineering and Architecture** **Engineering Sciences** **Natural Sciences**
Agriculture, Forestry, Horticulture and Veterinary Medicine
Agriculture, Forestry, Horticulture and Veterinary Medicine **Basic Biological and Medical Research** **Life Sciences**
Urbanism, Spatial Planning, Transportation and Infrastructure Planning, Landscape Planning
Construction Engineering and Architecture **Computer Science** **Traffic and Transport Systems, Logistics**
Systems Engineering **Computer Science, Electrical and System Engineering** **Materials Science**
Materials Science and Engineering **Water Research** **Geography** **Geochemistry, Mineralogy and Crystallography**
Geophysics **Geophysics and Geodesy** **Geosciences (including Geography)** **Mathematics** **Physics**
Analytical Chemistry, Method Development (Chemistry) **Chemistry** **Soil Sciences**
Bioinformatics and Theoretical Biology **Biophysics** **Biology**

Description 4TU.ResearchData, previously known as 3TU.Datacentrum, is an archive for research data. It offers the knowledge, experience and the tools to share and safely store scientific research data in a standardized, secure and well-documented manner. 4TU.Centre for Research Data provides the research community with: Advice and support on data management; A long-term archive for scientific research data; Support for current research projects; Tools for reusing research data.

Contact <https://researchdata.4tu.nl/en/about-4turedsearchdata/contact/>
researchdata@4tu.nl

Catalogue pour une Optimisation
du Partage et de l'Interopérabilité
des Données de la Recherche



<https://cat.opidor.fr/>

- Recense et décrit les **services français** dédiés aux données scientifiques
- Proposé sous forme d'un wiki, cet **outil collaboratif** ouvert à tous permet de repérer et ajouter des services utiles dans le cadre d'un projet de recherche
- Cat OPIDoR présente par **domaine scientifique** :
 - des sites d'information,
 - de formation,
 - des outils de gestion,
 - des plateformes,pour accompagner les chercheurs sur l'ensemble des étapes clés de la gestion, collecte, stockage, conservation et ouverture des données

SAUVEGARDE POUR LE PARTAGE

OUTIL CAT OPIDoR



The screenshot shows the Cat OPIDoR website interface. At the top, there is a navigation bar with 'Accueil', 'Discussion', and a search bar. The main content area is titled 'Cat OPIDoR, wiki des services dédiés aux données de la recherche'. It features three main sections:

- Quel type de service ?** [modifier] - A list of service categories: INFORMATION, FORMATION, ACCOMPAGNEMENT, OUTILS DE GESTION DES DONNÉES, PLATEFORME D'ACQUISITION, PLATEFORME DE CALCUL, **ENTREPÔT DE DONNÉES** (highlighted with a red box), PLATEFORME D'ACCÈS, and PLATEFORME D'ARCHIVAGE.
- A quel stade du cycle de vie des données ?** [modifier] - A circular diagram showing the data lifecycle stages: Planification, Collecte, Analyse, Documentation, Stockage, Conservation, Exposition, and Réutilisation.
- Dans quel domaine scientifique ?** [modifier] - A list of scientific domains: SCIENCES HUMAINES & SOCIALES [Afficher], SCIENCES & TECHNOLOGIES [Afficher], and VIE & SANTÉ [Afficher].

At the bottom right, there is a map of Europe with various locations marked by blue and yellow pins, indicating the geographical distribution of services.

SAUVEGARDE POUR LE PARTAGE

OUTIL CAT OPIDOR





Non connecté [Discussion](#) [Contributions](#) [Créer un compte](#) [Se connecter](#)

Page [Discussion](#)
[Lire](#) [Voir le texte source](#) [Afficher l'historique](#)

Rechercher dans Cat OPIDoR Q

Entrepôt de données

- Sur quelles plateformes puis-je déposer et partager les données que j'ai produites au cours de mes recherches ?
- Existe-t-il un entrepôt français dans ma discipline de recherche ?

Afficher les entrées Rechercher:

| Services | Domaine scientifique | Mots clés | Localisation | Stade du cycle de vie |
|-----------|------------------------------|--|--------------------|--|
| ANPERSANA | Sciences Humaines & Sociales | Linguistique Langue basque Textes basques Musicologie Chants basques Licence Creative Commons | Bayonne | Documentation Conservation Exposition Réutilisation |
| ARCHITOU | Sciences Humaines & Sociales | Histoire géographie sociétés. | Toulouse | Exposition Réutilisation |
| ArkeoGIS | Sciences Humaines & Sociales | Archéologie Histoire Géographie Système d'information Géographique Métadonnées Dublin Core Identifiant pérenne Handle | Strasbourg | Conservation Exposition Réutilisation |
| AVISO+ | Sciences & Technologies | Allimétrie spatiale Océanographie Climatologie Hydrologie Glaciologie Météorologie Hauteur de mer Hauteur des vagues Vitesse du vent | Ramonville St-Agne | Conservation Exposition Réutilisation |
| BASS2000 | Sciences & Technologies | Astronomie Astrophysique Soleil | Meudon | Conservation Exposition Réutilisation |

Accueil
A propos
Modifications récentes

Naviguer par
Type de service
Stade du cycle de vie
Domaine scientifique
Service
Structure
d'appartenance

Contribuer
Ajouter un service
Ajouter une structure
d'appartenance

Aide
Description d'un
service
Description d'une
structure
FAQ
Glossaire
Cat OPIDoR en 2mn

Outils
Pages liées
Suivi des pages liées
Pages spéciales
Version imprimable
Lien permanent
Information sur la
page
Chercher les propriétés

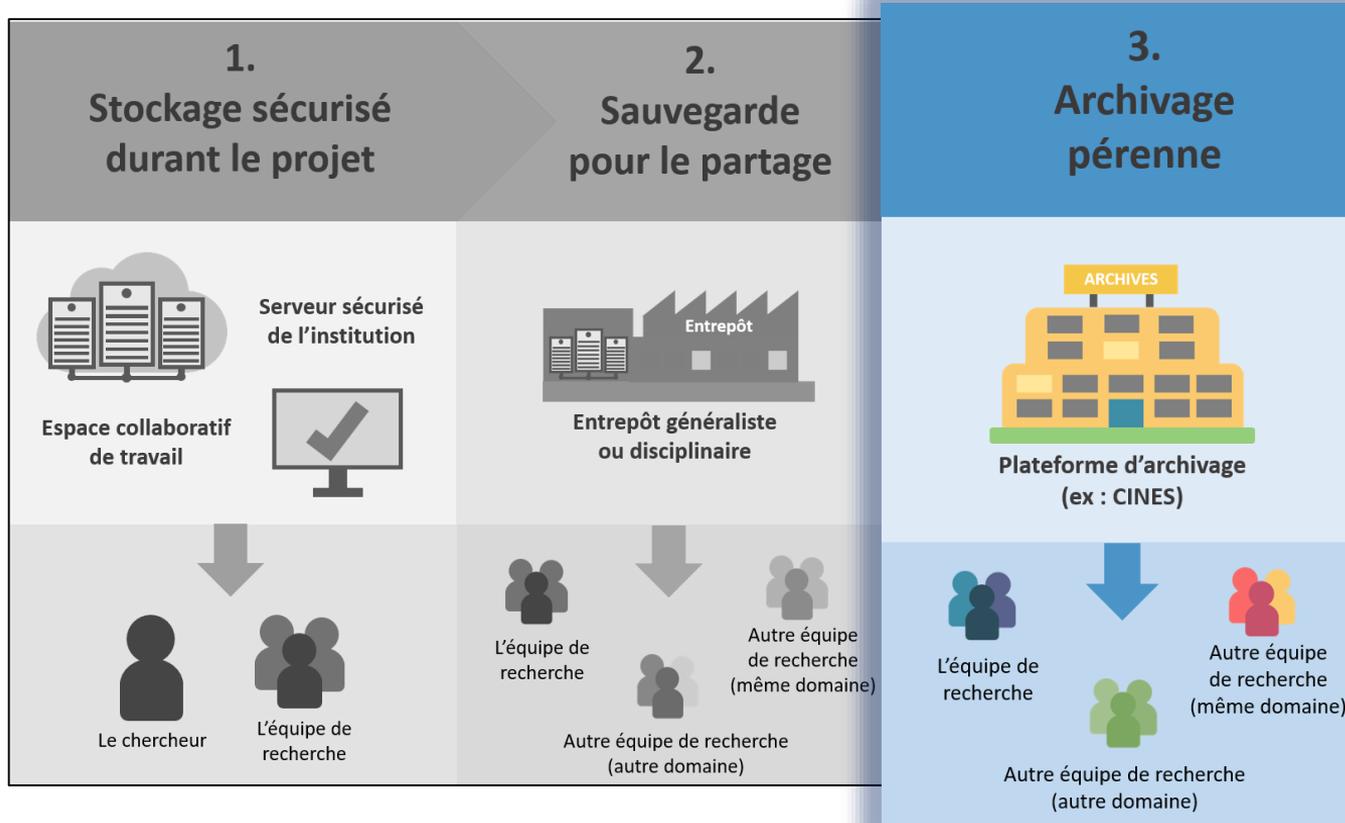




ETAPE 3

ARCHIVAGE PÉRENNE

3 ÉTAPES DE SAUVEGARDE DES DONNÉES



CYCLE DE VIE DES DONNÉES DE RECHERCHE

SAUVEGARDE DES DONNÉES

Sauvegarde
Etape 3



D'après Research data lifecycle – UK Data Service
<https://www.ukdataservice.ac.uk/manage-data/lifecycle>



- Le **CINES** est l'opérateur mandaté par le Ministère pour opérer la mission d'archivage pérenne pour l'Enseignement Supérieur et la Recherche. Il développe différentes solutions, en particulier **PAC**, la **Plateforme d'Archivage au CINES**
- Selon son institution, sa discipline ou l'entrepôt choisi, il existe déjà des partenariats avec le CINES, proposant un accompagnement pour l'archivage.
Ex : Huma-Num en SHS



Les données à archiver doivent présenter une **valeur scientifique reconnue** par la communauté dont elles proviennent

L'archivage garantit une conservation des données pour **plus de 30 ans**

ARCHIVAGE DES DONNÉES

VALEUR DES DONNÉES A CONSIDÉRER (TRI / SELECTION DES DONNÉES)

VALEUR SCIENTIFIQUE
DES DONNÉES

MESURES DE
CONTRÔLE DE LA
QUALITÉ DES
DONNÉES

CONSIDERATIONS
POLITIQUES /
INSTITUTIONNELLES

CONSIDERATIONS
JURIDIQUES /
STATUTAIRES

CONSIDERATIONS
FINANCIÈRES

RÈGLES DE TRI ET DE
CONSERVATION DES
ARCHIVES

- L'archivage numérique pérenne des **documents électroniques** consiste à conserver le document et l'information qu'il contient :
 - Dans son aspect physique comme dans son aspect intellectuel
 - Sur le très long terme
 - De manière à ce qu'il soit en permanence accessible et compréhensible

ARCHIVAGE PÉRENNE

PRÉPARATION DES DONNÉES À ARCHIVER

1 Sélectionner les jeux de données (et métadonnées associées) à conserver à long terme (peuvent être différents des jeux de données partagés)

2

Traiter les données si cela est nécessaire

- Ex : Données personnelles (nécessitent une anonymisation)

3

Vérifier la validité des formats de fichiers de données avec l'outil **FACILE** mis en place par le CINES

4

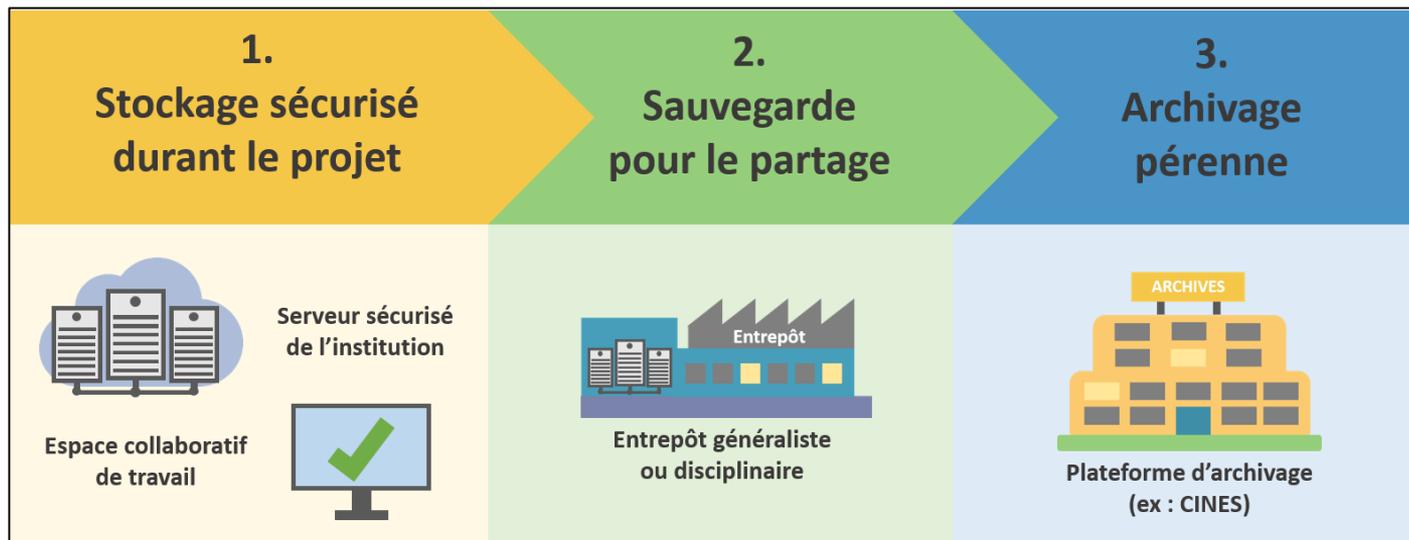
Documenter également les **logiciels** permettant l'accès aux données

5

Compléter et enrichir si besoin les **métadonnées**
(Les données doivent posséder une description minimale imposée par le CINES)

A RETENIR

Principe « Aussi ouvert que possible, aussi fermé que nécessaire »

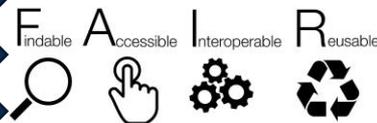


Sauvegardes

Métadonnées

Identifiants pérennes

Licences



A RETENIR

R

Provenance

Description riche

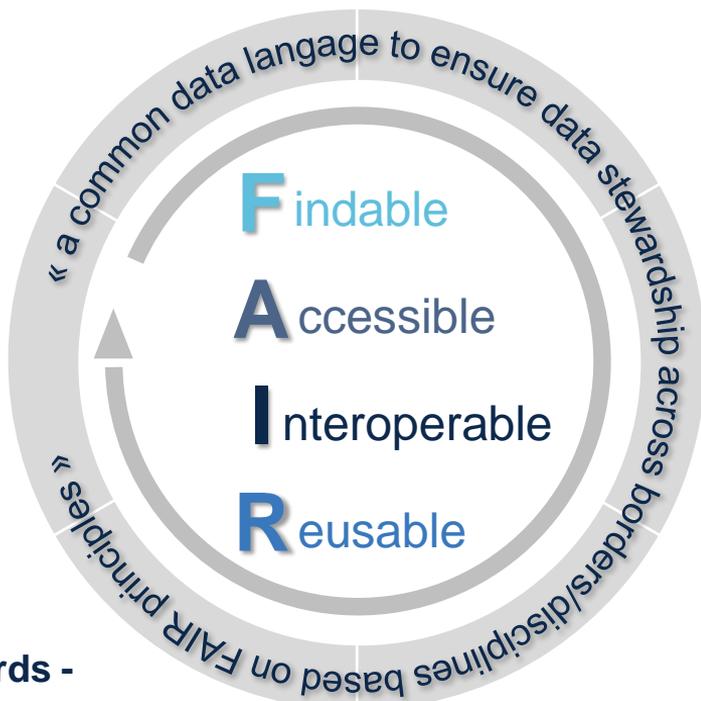
Standards
communautaires

Licences

Formats standards

Lien vers d'autres
(méta)données

Vocabulaires standards -
ontologies



Identifiants pérennes

Métadonnées riches

Catalogues-Entrepôts

Protocoles standardisés
& ouverts

Authentification &
autorisation si besoin

Accès permanent aux
métadonnées

Source: Implementation Roadmap for the European Science Cloud (Staff Working Document SWD(2018) 83), 14 March 2018

Traduction Inra <https://www6.inra.fr/datapartage/Produire-des-donnees-FAIR>

Des données FAIR plus faciles à partager et réutilisables
par les hommes et par les machines

Merci de votre attention

paolo.lai@inist.fr

www.inist.fr

Cat OPIDoR : <https://cat.opidor.fr/>

Contact : infocatopidor@inist.fr

PID OPIDoR : <https://opidor.fr/identifier/>

Contact : datasets@inist.fr

www.cnrs.fr



DMP OPIDoR : <https://dmp.opidor.fr/>

Contact : info-opidor@inist.fr

DoRANum : <https://doranum.fr/>

Contact : contact@doranum.fr

