

**Des données ouvertes  
pour une science durable au Sud**



# Présentation de l'entrepôt de données DataSuds

Luc Decker

Administrateur de DataSuds - Service IST, MCST, IRD



[dataverse.ird.fr](http://dataverse.ird.fr)

[data@ird.fr](mailto:data@ird.fr)

Séminaire DataSuds 2021 - 23 septembre 2021



# DataSuds

Entrepôt de données institutionnel de l'IRD, plateforme de publication de données scientifiques

Lancement en sept. 2019

## Solution technique



Open source research data repository software

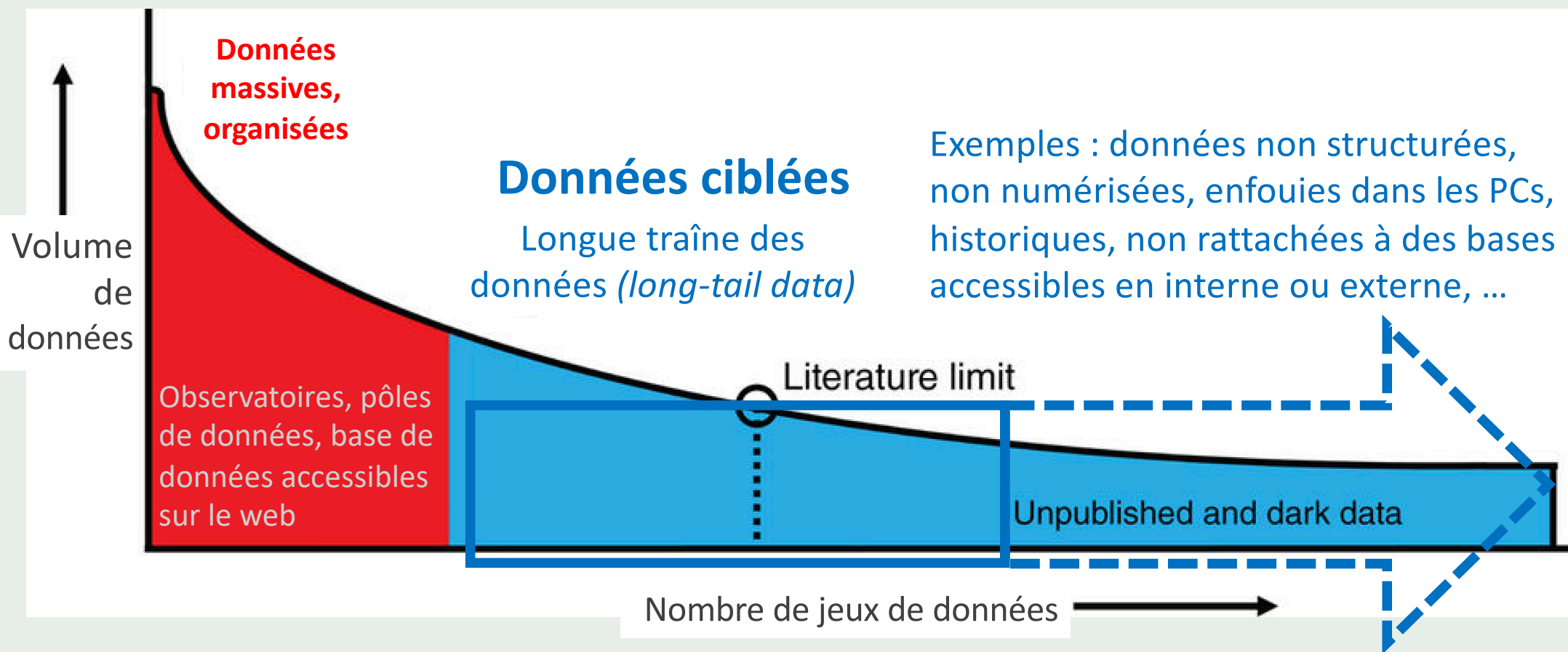
The screenshot shows the Dataverse website interface. At the top, there is a navigation bar with the Dataverse logo, a search bar, and links for 'À propos', 'Guide d'utilisation', 'Support', 'Français', and 'Se connecter'. Below the navigation bar, there is a banner with the text 'Des données ouvertes pour une science durable au Sud' and a large blue box containing the URL 'https://dataverse.ird.fr'. The main content area features a 'DataSuds (IRD)' entry with a download count of 13,041. Below this, there is a section titled 'Comment déposer des données ?' with a link to 'service support'. A carousel of logos for various institutions is displayed, including Partnerships, UMR IIES-Paris, UMR HydroSciences Montpellier, and UMR NUTRIPASS. At the bottom, there is a search bar and a list of search results. The first result is 'France-CGIAR BRIDGE collaboration: integrated model-based approaches for climate and water smart decisions' by Vadez, Vincent; Bossuet, Jérôme, dated 16 sept. 2021. The second result is 'Anthropogenic fibers (quantity, size, colour) observed in whole body, gills, digestive glands and remaining tissues of Asiatic Clam (Meretrix Lyrata) sampled in Can Gio mangrove, Ho Chi Minh City, Viet Nam in 2016' by Kieu-Le, Thuy-Chung; Tran, Quoc-Viet; Truong, Tran Nguyen Sang; Strady, Emilie, dated 10 sept. 2021.

# DataSuds : un entrepôt pour quels besoins ?

## Enjeux et objectifs

- Sur le long terme, préserver les données scientifiques *en danger*
- Valoriser le « patrimoine des données »
  - Visibilité et partage avec la communauté scientifique
  - Augmenter l'impact (global) des projets de recherche
- Transparence et reproductibilité des expériences
- Aider à répondre aux demandes des financeurs et des éditeurs

## Positionnement de l'entrepôt DataSuds



Source : distribution des données de la recherche (Ferguson et al., 2014)

## DataSuds : principales fonctionnalités

Attribution de DOI

Gestion des  
dépôts par le  
scientifique

Métadonnées  
adaptées à la  
discipline

Autorisation  
Restriction d'accès  
si nécessaire

Organisation en  
arborescence

Lien provisoire  
sécurisé pour les  
reviewers

Fédération  
d'identité  
API

Moissonnage

Gestion fine des  
droits utilisateur

# Présentation d'un jeu de données dans l'entrepôt DataSuds

**Identifiant** (DOI: doi)

**Métadonnées**

**Licence** (CC BY)

**Citation d'un jeu de données**

- Articles
- Rapports de projet
- Rapports d'activité
- Dossiers de candidature

**Accès fichiers**

**Base de données insulaires mondiale (BIM)**

DEPNATIERE, Christian, 2019, "Base de données insulaires mondiale (BIM)", <https://doi.org/10.25706/131760K>, DataSuds, V1

La base de données SIG (GIS en Anglais) veilleurs BIM concerne toutes les îles océaniques de plus de 0,06 km<sup>2</sup> (6 hectares, 116500 îles) avec des informations géographiques, toponymiques, altimétriques, climatiques, géologiques "par île". A ces données "par île" s'ajoutent des données thématiques par point de jonction bathymétrique (ETOPOS), îles/îlots 15000 SR, climat actuel et scénarii 2050, Méta-sources (îles, îlots, marais, lagunes, lieux habités, ...) relatif visible en mer atlas ligne d'horizon, etc. (2019)

Earth and Environmental Sciences

Neotologie, Etudes insulaires, Système d'Informations Géographiques (SIG)

Depierre, C. (2005a). The Challenge of Neotology: A Global Outlook on the World Archipelago. Part I: Scene Setting the World Archipelago. *Island Studies Journal*, Vol. 3, No. 1, 1-16.

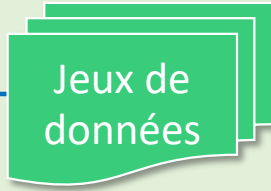
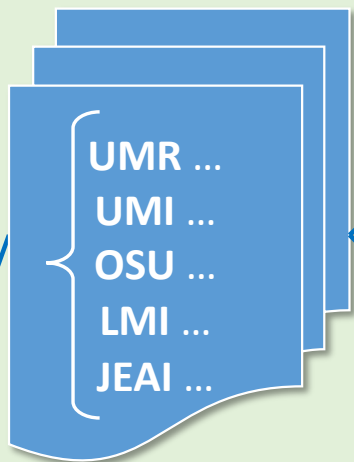
La neotologie (neotology en Anglais) fait référence à la science des îles. La base BIM est un outil d'étude général pour les Etudes insulaires (Island studies en Anglais).

Fichiers | Métadonnées | Conditions | Versions

1 à 6 de 6 Fichiers

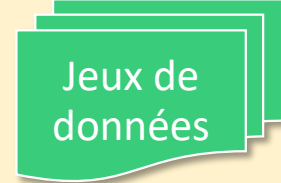
Fichier	Download
BIM_îles_cpg	Télécharger
BIM_îles_dbf	Télécharger

**DataSuds :**  
structuration  
des collections  
de données  
(dataverses)



**Organisation initiale**

**DataSuds**  
« Racine » de  
l'entrepôt



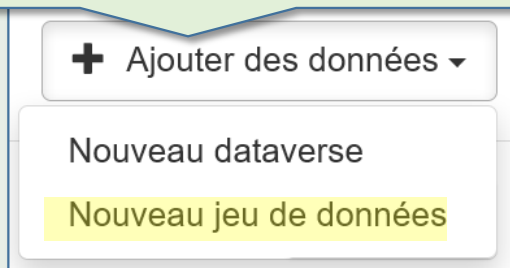
**Autres cas (à discuter)**

# DataSuds : facilité d'utilisation

1. Créer un compte utilisateur (login *Renater*)



2. Contacter [data@ird.fr](mailto:data@ird.fr) (ou le référent du laboratoire) : demander l'accès à une collection ou sa création



3. Saisir un nouveau jeu de données (formulaire en ligne)

The form includes the following fields and sections:

- Titre \***: Title field.
- Auteur \***: Author information, including:
  - Nom \***: DECKER, Luc
  - Affiliation \***: IRD - French National Research Institute
  - Schéma de l'identifiant \***: Sélectionner...
  - Identifiant \***: [Empty field]
- Personne-ressource \***: Resource person information, including:
  - Nom \***: DECKER, Luc
  - Affiliation \***: IRD - French National Research Institute
  - Courriel \***: luc.decker@ird.fr
- Description \***: Description field with a note: "Ce champ n'autorise que certains balises HTML." and a "Texte" sub-field.
- Date de description \***: Date field in format jj/mm/aaaa.
- Sujet \***: Subject dropdown menu.
- Mot-clé \***: Keyword field, including:
  - Terme \***: [Empty field]
  - Vocabulaire \***: [Empty field]
  - Adresse URL du vocabulaire \***: Entrez l'adresse URL complète commen...
- Thématique scientifique \***: Scientific theme dropdown menu.
- Publication connexe \***: Related publication field.

4. Relecture et édition (accompagnement)  
➔ Publication



## DataSuds : préparer la publication de données

Droits de diffuser les données ?

Sélection et granularité des données ?

Description précise (métadonnées) ?

Documentations associées ?

Convertir les fichiers dans des **formats ouverts** ?

Quelle **licence** (ou conditions d'utilisation) attribuer ?

Faciles à trouver

Accessibles

Réutilisables

Interopérables

... principes « **FAIR** » & « *Data Citation* », qualité éditoriale, bonnes pratiques ...

# DataSuds : accompagnement des déposants

Site web support <https://data.ird.fr/>



F.A.Q.

Mode d'emploi

Conseils

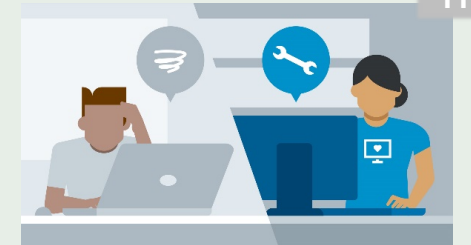
Guides

Informations pratiques

Gérer, publier et utiliser des données de la Recherche

## DataSuds : accompagnement des déposants


Service : assistance, formation et communication



- Rendez-vous en présentiel ou par visioconférence (1h)
  - conseils personnalisés
  - aide à la préparation des métadonnées et des fichiers
  - en lien avec les référents du laboratoire (le cas échéant)
- Par email : [data@ird.fr](mailto:data@ird.fr)
- Organisation d'ateliers « *datathons* » et de formations
- Présentation du service, par exemple au cours d'une session d'un séminaire interne d'un laboratoire

# Qualité de l'entrepôt de données DataSuds

## Guide de dépôt d'un jeu de données

Métadonnées			
Champs	Réf.	Préconisations ou/et recommandations - Conseils pratiques	Finalité et commentaires
tous	L1	Saisir les informations en anglais de préférence. Afficher l'interface de DataSuds (les formulaires) en anglais peut faciliter la saisie : changer de langue dans le menu principal en haut de l'écran.	Visibilité et valorisation du jeu de données : recommandé mais pas obligatoire, comme pour les articles scientifiques
	L2	Ne pas mélanger différentes langues, sauf éventuellement pour le champ « Description » (dans ce cas, ajouter une séparation entre les langues avec le code <hr>) ainsi que les mots-clés. Possibilité de saisir une traduction du titre dans le champ « Autre titre » car un titre dans la langue du pays améliore la visibilité au niveau local.	Clarté de la présentation du jeu de données. Dans l'entrepôt, les champs de métadonnées ne sont pas multilingues, sauf exception : il n'est pas possible de saisir les informations traduites dans différentes langues. La langue sélectionnée dans le menu supérieur s'applique uniquement à l'interface utilisateur.
Titre	T1	Spécificité et caractérisation des données : type de données, contexte, période de collecte ou/et localisation géographique - si applicable et pertinent. Autre possibilité de titre : « <i>Replication data for...</i> (insérer le titre de l'article scientifique associé aux données)... » Exemples : consulter les jeux de données publiés récemment dans DataSuds. Pour davantage de conseils : <a href="https://coop-ist.cirad.fr/rediger/article-scientifique/le-titre/1-le-titre-premier-niveau-de-selection-sur-le-web">https://coop-ist.cirad.fr/rediger/article-scientifique/le-titre/1-le-titre-premier-niveau-de-selection-sur-le-web</a>	Selon la formulation et la précision du titre, un utilisateur de données potentiel ira - ou non - consulter plus en détails le jeu de données.
	T2	Longueur appropriée, approximativement entre 3 et 20 mots	Suivre les usages, comme pour un article scientifique.
	T3	Retirer les informations trop détaillées qui n'ont en général pas leur place dans un titre : noms de fichiers, noms d'auteurs, citation complète de l'article associé, parenthèses inutiles, caractères spéciaux.	
Auteurs	A1	Personnes qui ont contribué à la production des données : rôle scientifique ou technique : conception, collecte, traitement, analyse. Le responsable du projet valide la liste des auteurs. Conseils : 1) Utiliser le bouton  pour ajouter des lignes au formulaire. 2) Attention, ce champ est prérempli par DataSuds avec le nom de la personne qui dépose « techniquement » les données : cette personne n'est pas nécessairement 1 <sup>er</sup> auteur, parfois pas même auteur ... à corriger si nécessaire. 3) Une méthode consiste à reprendre la liste des auteurs d'un article associé aux données - ou encore de modifier cette liste pour ajouter ou mettre en avant des intervenants ayant joué un rôle important dans la collecte ou le traitement des données.	Procéder comme pour un article scientifique dans le choix et l'ordre des auteurs.  <a href="#">Data Citation Principles</a>
	A2	Format : « Nom, Prénom », avec les noms et prénoms en lettres minuscules. Exemple : Dupont, Jean	Suivre les usages, comme pour un article scientifique. Le format des auteurs se retrouve dans la citation du jeu de données.

Conçu initialement pour le service d'accompagnement et les référents des données.

- Connaître les critères qui seront appliqués au moment de la relecture, en comprendre les motivations
- Possibilité d'anticiper et d'améliorer pour une publication plus rapide

## En conclusion

### Des objectifs raisonnables

- ◆ Accueillir les données que les chercheurs souhaitent préserver et diffuser
- ◆ Rendre les données Faciles à trouver, Accessibles, Interopérables, Réutilisables
- ◆ Complémentarité avec les entrepôts thématiques, les pôles de données, ...

### ...pour répondre à la diversité des besoins

- ◆ Valorisation, attribution de DOIs, *Data papers*, se conformer aux demandes des financeurs et des éditeurs, ...

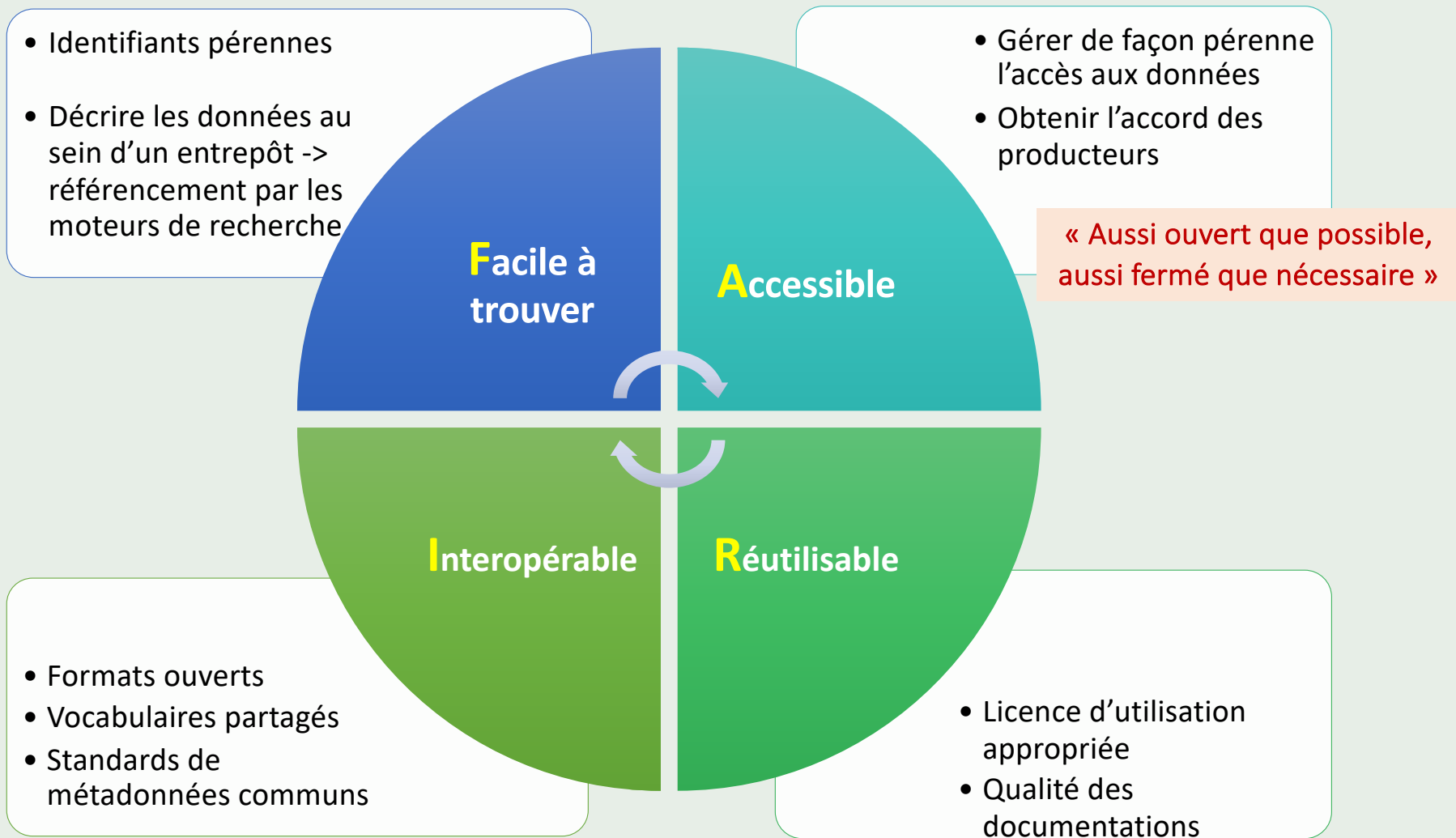
### ...au service de l'IRD

- ◆ Améliorer la connaissance et la gestion du patrimoine de données
- ◆ Contribuer de manière concrète à la Science Ouverte





# Principes FAIR : augmenter le potentiel des données





## Données de la recherche

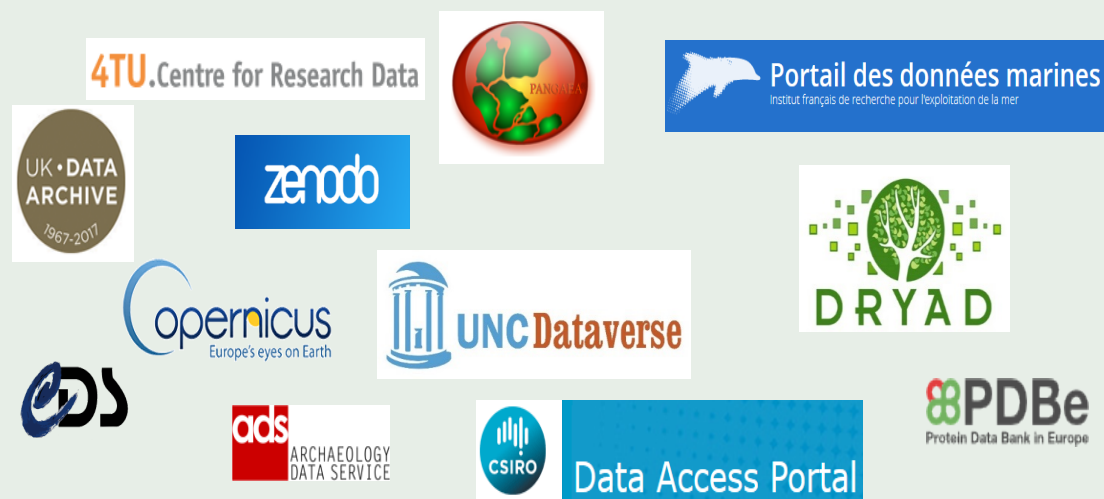
**Données primaires ou brutes** : enregistrements factuels (chiffres, textes, images et sons), sources principales pour la recherche scientifique, reconnus comme nécessaires pour valider des résultats de recherche [*définition OCDE, 2007*]

**Données dérivées** : élaborées à partir de données primaires.

**Données d'intérêt** : réutilisables afin d'améliorer les connaissances par l'enrichissement, la combinaison à d'autres jeux de données.

## Entrepôts de données de la recherche

Service en ligne (outil) : dépôt, description, conservation, recherche et diffusion des jeux de données.



Types d'entrepôts : disciplinaires / institutionnels / généralistes / infrastructures nationales, internationales / commerciaux / créés par des éditeurs scientifiques / ....

## A quoi sert un entrepôt de données ?

- ✓ **Visibilité, partage et accès** aux données des Unités et projets de recherche
- ✓ **Maîtrise de la diffusion** des données : licences et niveaux d'accès
- ✓ **Valorisation** : susciter des collaborations entre recherche publique et secteur privé
- ✓ **Ethique** : rendre les données plus facilement accessible à vos partenaires du Sud, obtenir leur accord pour la diffusion

# Préservation des données scientifiques ?

**20 ans après publication, 80% des données ont été perdues**

## Causes

- ✓ Destruction des supports, virus
- ✓ Obsolescence matérielle ou logicielle
- ✓ Lieu de stockage indéfini
- ✓ Erreurs humaines

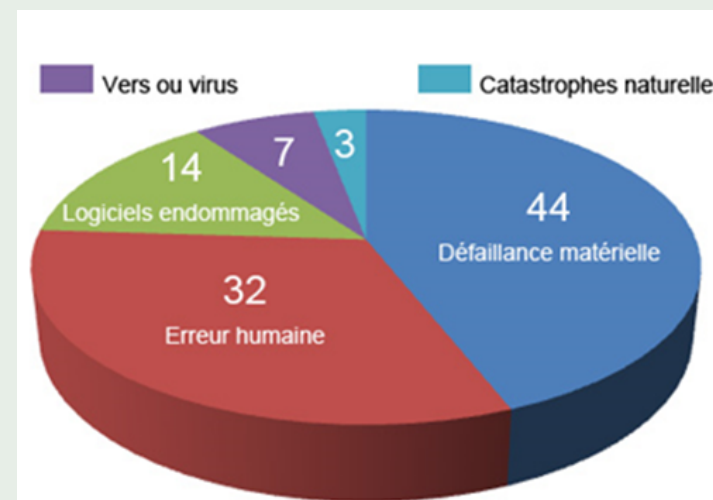
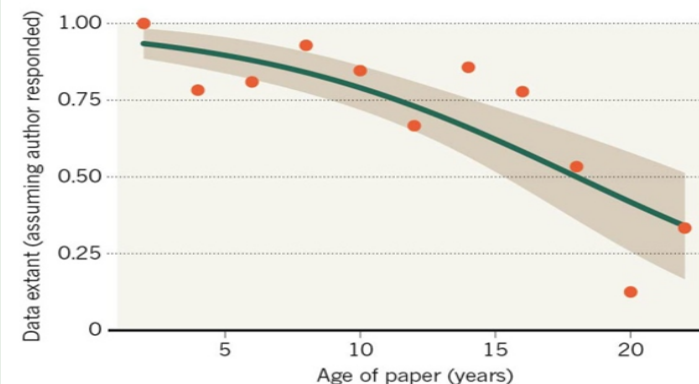
## Conséquences

- ✓ Perte de temps, d'argent, de fonds publics
- ✓ Moins de possibilités de vérification des résultats
- ✓ Pas de comparaison des résultats dans le temps ou l'espace
- ✓ Pas de réutilisations par d'autres publics ou pour d'autres fins

VINES Timothy H., et al. [The Availability of Research Data Declines Rapidly with Article Age](#), *Current Biology*, 2014.

### MISSING DATA

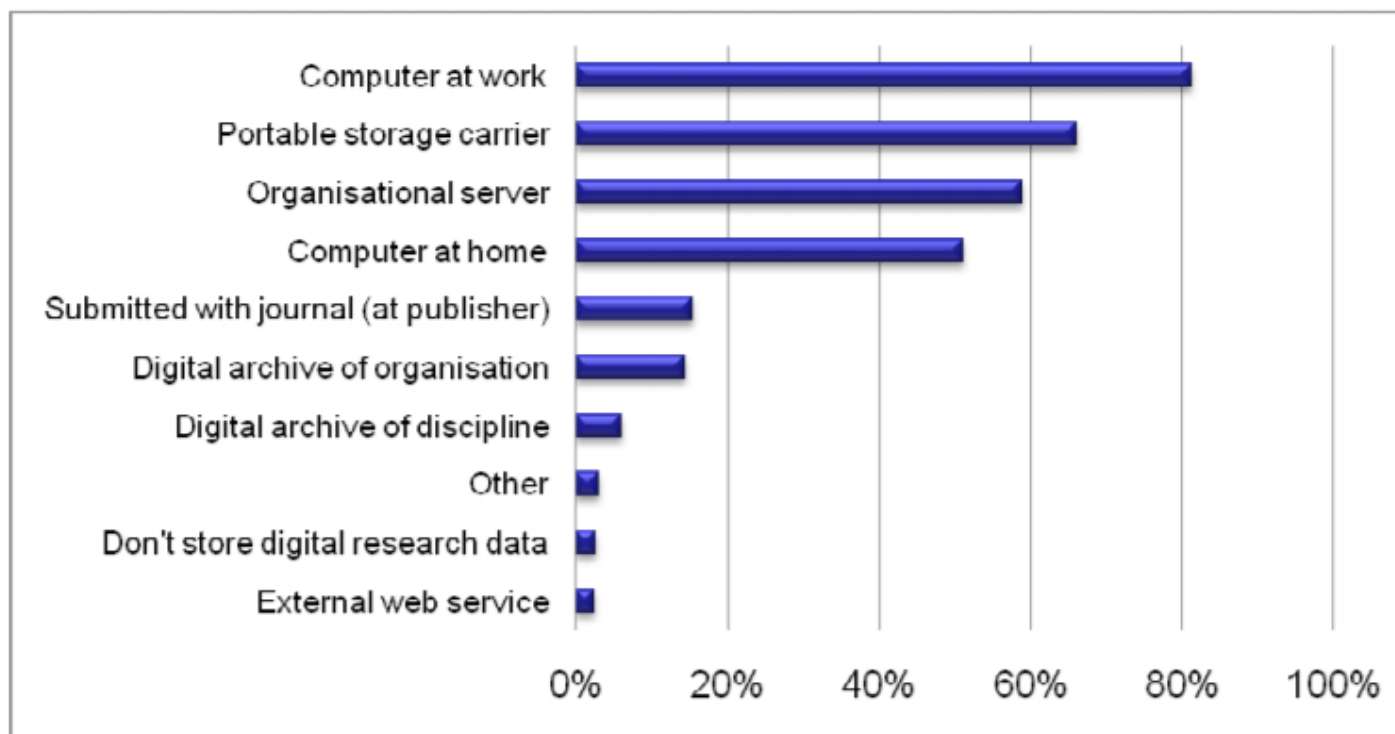
As research articles age, the odds of their raw data being extant drop dramatically.



# Préservation : des données en danger ?

Perte de 17 % par an (Pierre Corvol, Collège de France)

Where do you currently store your research data ? (multiple answers possible)

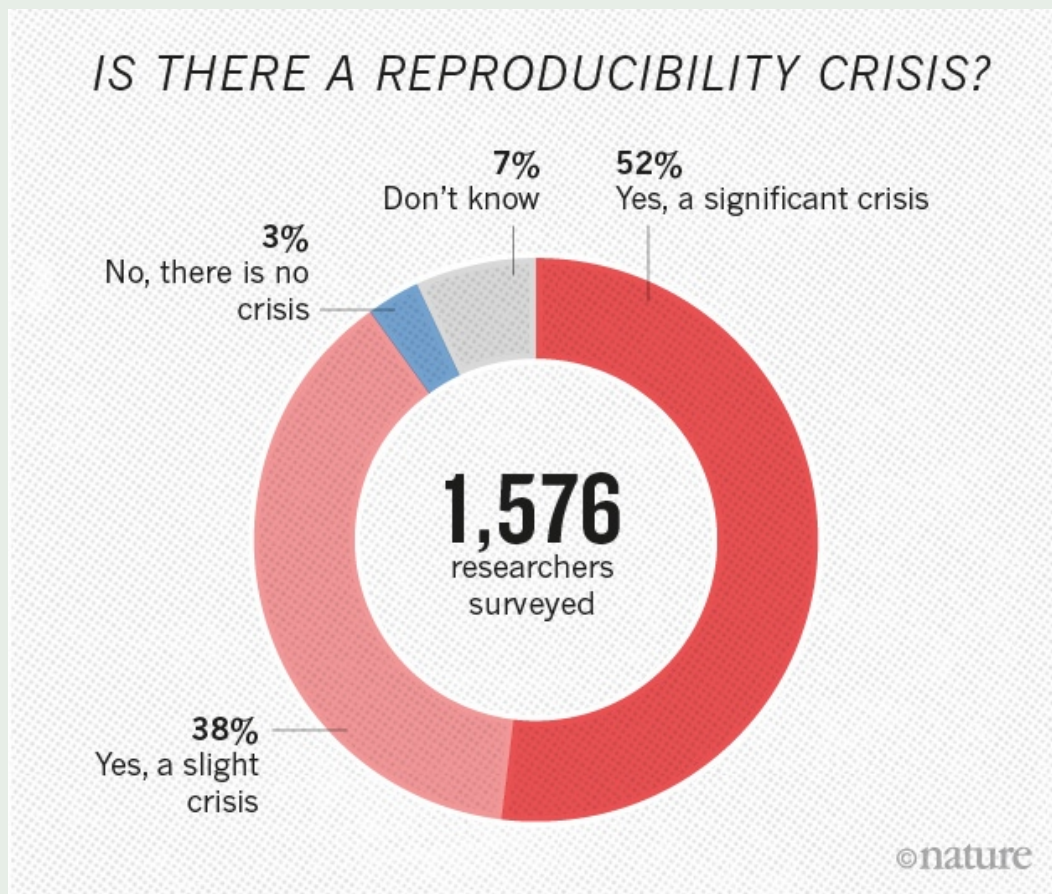


Graph 2: Source: [PARSE.Insight<sup>2</sup> survey](#), held among researchers internationally, N = 1202 researchers

[https://libereurope.eu/wp-content/uploads/PARSE-Insight\\_D3-5\\_InterimInsightReport\\_final.pdf](https://libereurope.eu/wp-content/uploads/PARSE-Insight_D3-5_InterimInsightReport_final.pdf)

## Reproductibilité des expériences ?

1500 chercheurs répondent à *Nature*

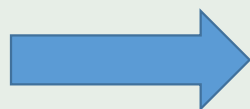


*“More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments”*

# Pourquoi *partager* les données ? entre injonctions et bénéfices



Incitations,  
obligations  
de partage



- ✓ Reproductibilité, Preuve
- ✓ Résultats accessibles à tous public
- ✓ Préservation
- ✓ Description et visibilité
- ✓ **Valorisation**

**Augmenter ses citations**  
Le partage des données augmente l'attractivité des articles (30%-70%)

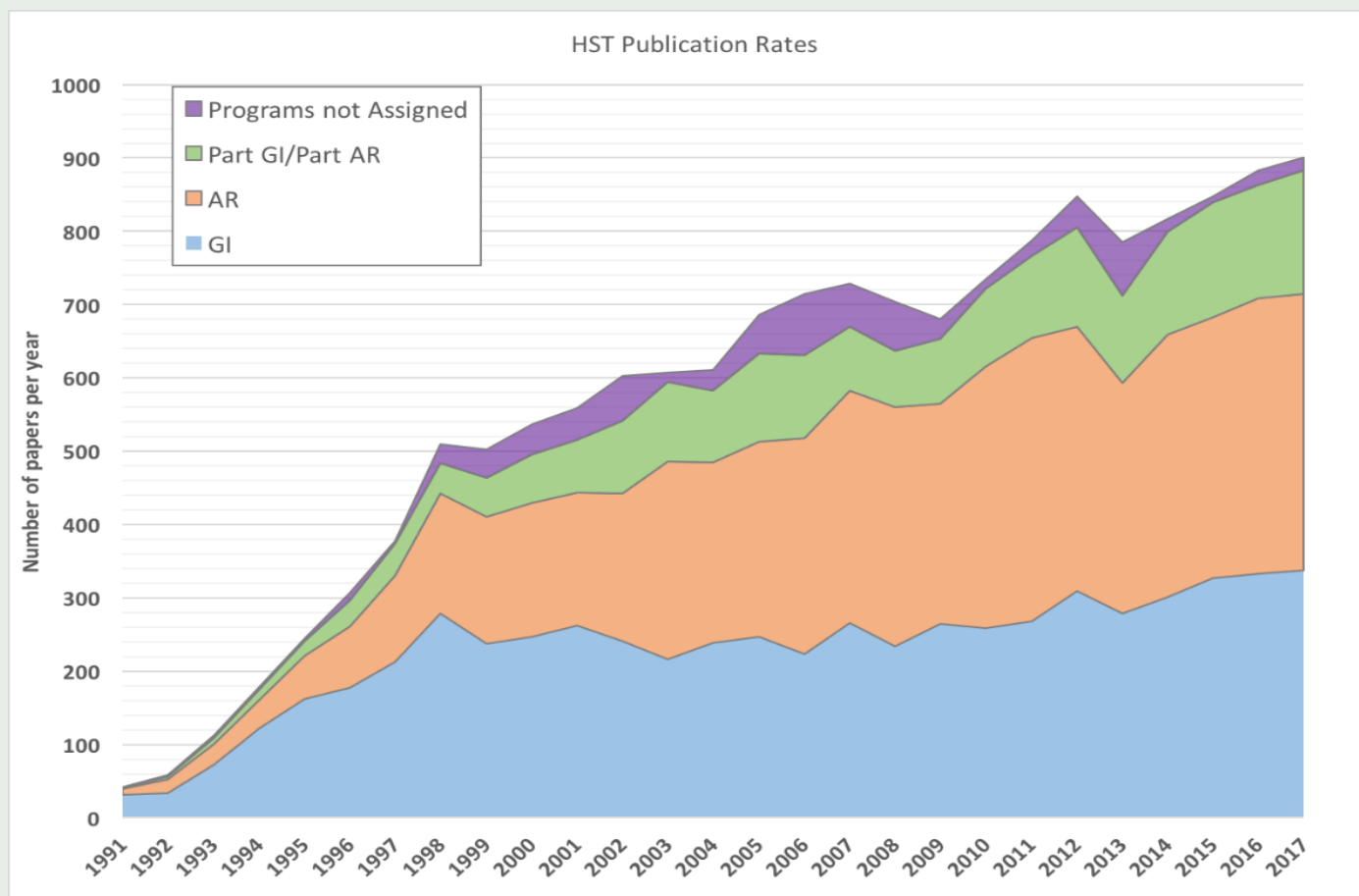
**Étendre son réseau**  
Nouvelles opportunités pour échanger/collaborer

**Attirer sur son profil chercheur**  
La demande pour réutiliser les données est en forte hausse

**Renforcer ses possibilités de recevoir des financements**  
Adéquation avec les exigences des financeurs

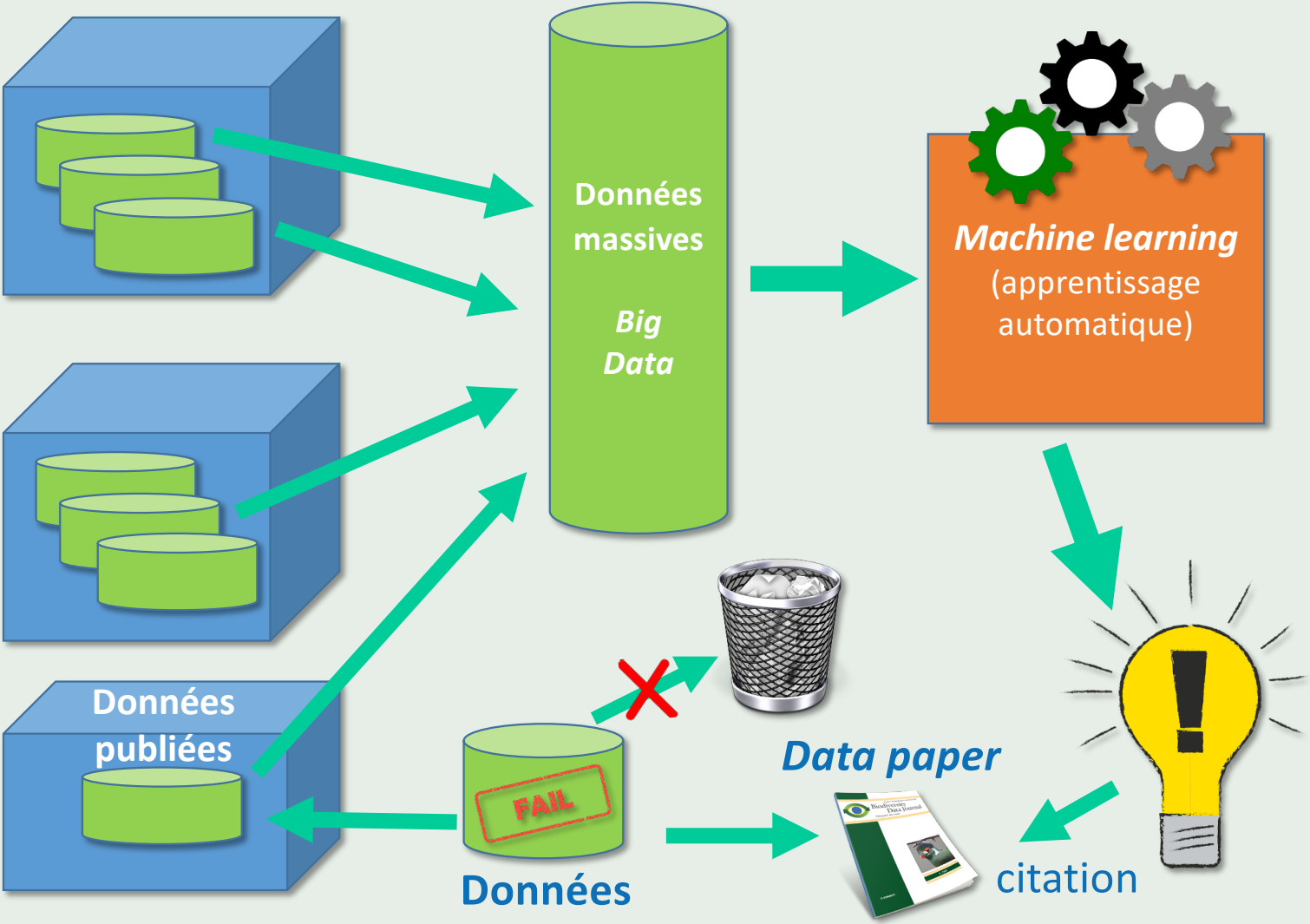
## Données archivées : davantage utilisées et citées que les données récentes

### Exemple : réutilisation des données en Astronomie

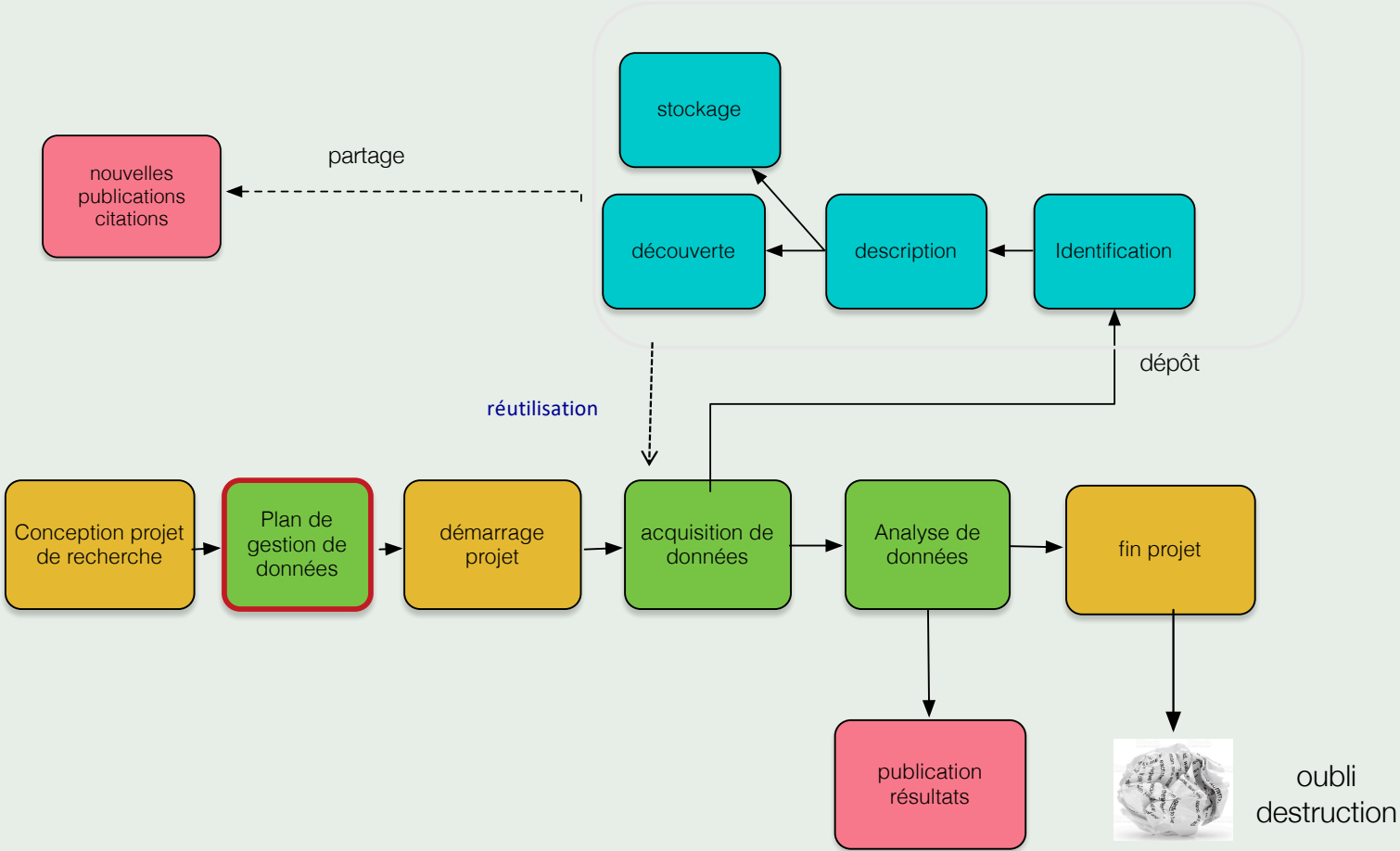




# Partage de données : alimenter les systèmes d'intelligence artificielle



# Changer les pratiques de gestion des données



# Plan national pour la science ouverte (2018-2021)

## 3 axes concrets

Axe 1 : Généraliser l'accès ouvert aux publications

### **Axe 2 : Structurer et ouvrir les données de la recherche**

Inciter à la diffusion ouverte des données

Données ouvertes associées aux articles scientifiques

Créer un réseau d'administrateurs de données

Axe 3 : s'inscrire dans une dynamique durable, européenne et internationale

## Questions soulevées par l'ouverture des données

### 1. Compliqué ?

→ Outils et assistance disponibles, infrastructures de données, Plans de gestion de données

### 2. Coûteux ?

→ Frais de structuration et d'ouverture éligibles dans les appels à projets

3. **Risqué ?** pour les données sensibles (personnelles, sécurité publique, secret professionnel, industriel et commercial)

→ Principes FAIR

« Aussi ouvert que possible, aussi fermé que nécessaire »