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
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

The Dataverse® Project

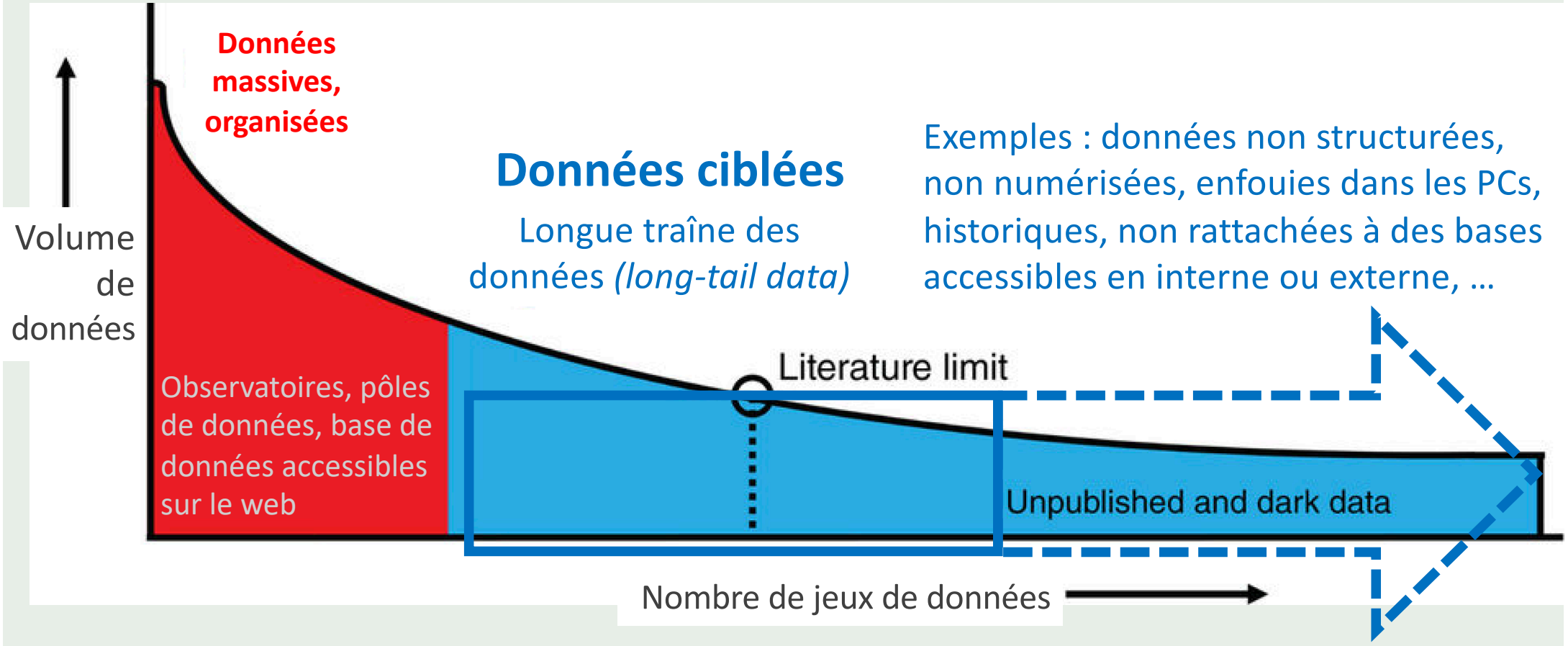
Open source research data repository software

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

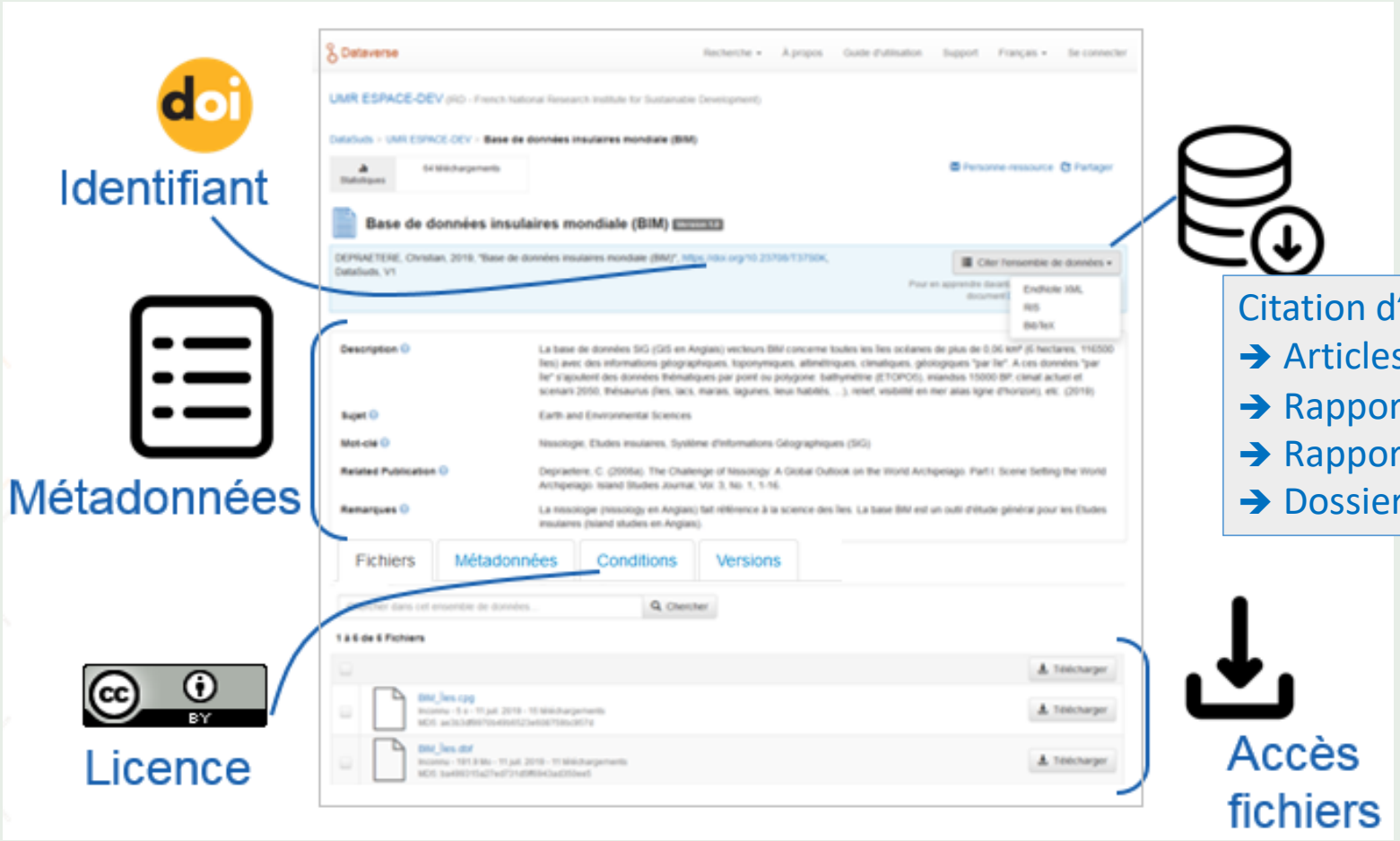
Positionnement de l'entrepôt DataSuds



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

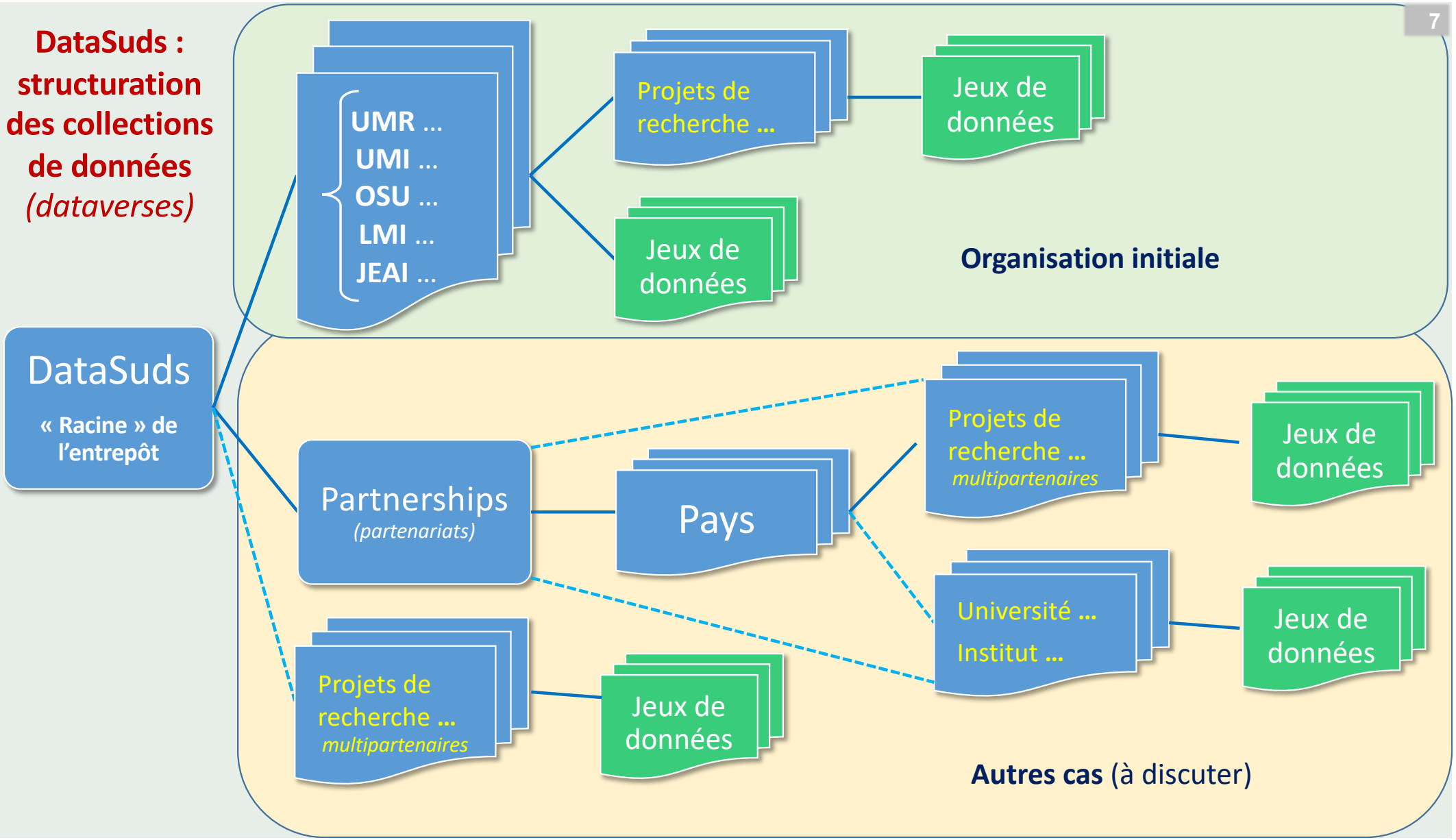
DataSuds : principales fonctionnalités

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



Citation d'un jeu de données
- ➔ Articles
- ➔ Rapports de projet
- ➔ Rapports d'activité
- ➔ Dossiers de candidature

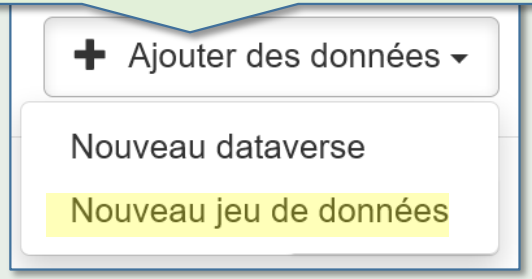
DataSuds : structuration des collections de données *(dataverses)*



DataSuds : facilité d'utilisation

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

2. Contacter 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)

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

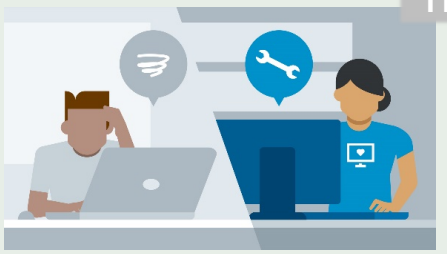
Conseils

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**
- 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 : « *Replication data for*…(insérer le titre de l'article scientifique associé aux données)… » Exemples : consulter les jeux de données publiés récemment dans DataSuds. Pour davantage de conseils : https://coop-ist.cirad.fr/rediger/article-scientifique/le-titre/1-le-titre-premier-niveau-de-selection-sur-le-web	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 1er 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. Data Citation Principles
	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.

- Connaitre 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

https://data.ird.fr/wp-content/uploads/2021/02/DataSuds_qualite_depots_202102v23_LD.pdf

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

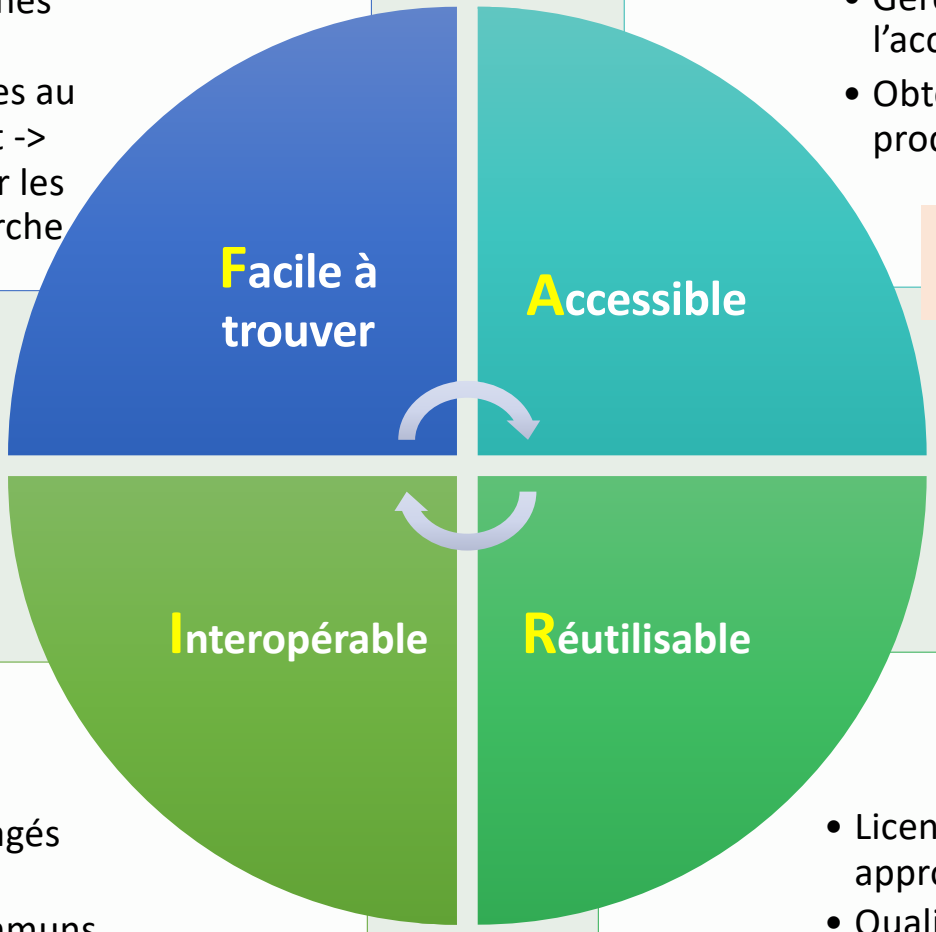
- Valorisation, attribution de DOIs, *Data papers*, se conformer aux demandes des financeurs et des éditeurs, …

…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

- Identifiants pérennes
- Décrire les données au sein d'un entrepôt -> référencement par les moteurs de recherche



- Gérer de façon pérenne l'accès aux données
- Obtenir l'accord des producteurs

« Aussi ouvert que possible, aussi fermé que nécessaire »

- Formats ouverts
- Vocabulaires partagés
- Standards de métadonnées communs

- Licence d'utilisation appropriée
- Qualité des documentations

Données de la recherche

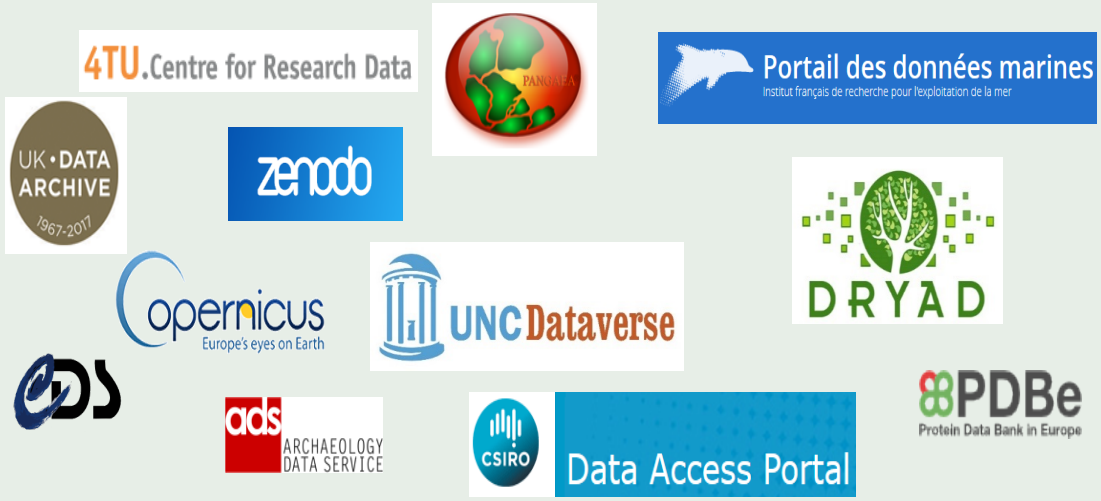
Données primaires ou brutes : enregistrements factuels (chiffres, textes, images et sons), sources principales pour la recherche scientifique, reconnus comme nécessaires pour valider des résultats de recherche *[définition OCDE, 2007]*

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 / ….

D'après Sylvie Cocaud. Cocaud et Aventurier 2017

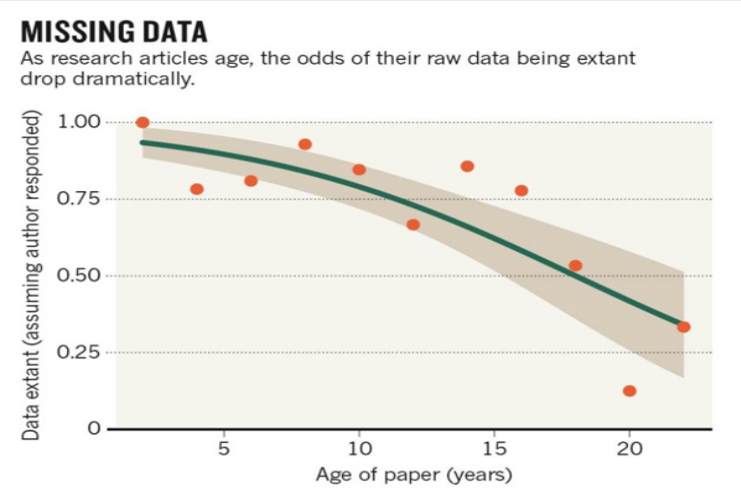
A quoi sert un entrepôt de données ?

- ✓ **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

VINES Timothy H., et al. The Availability of Research Data Declines Rapidly with Article Age, Current Biology, 2014.

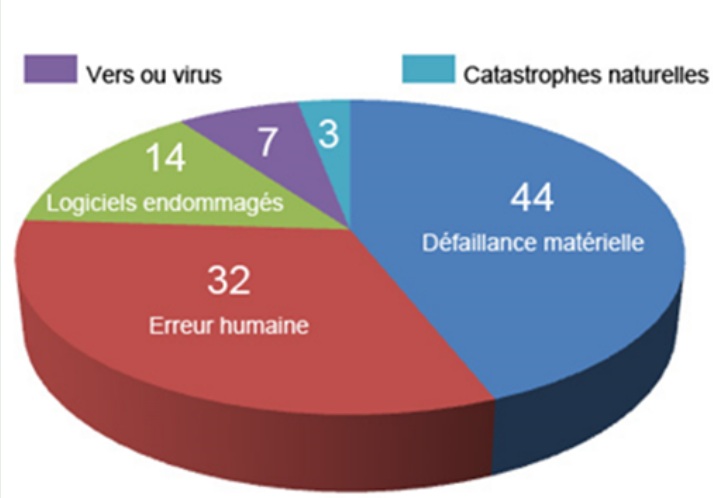


Causes

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

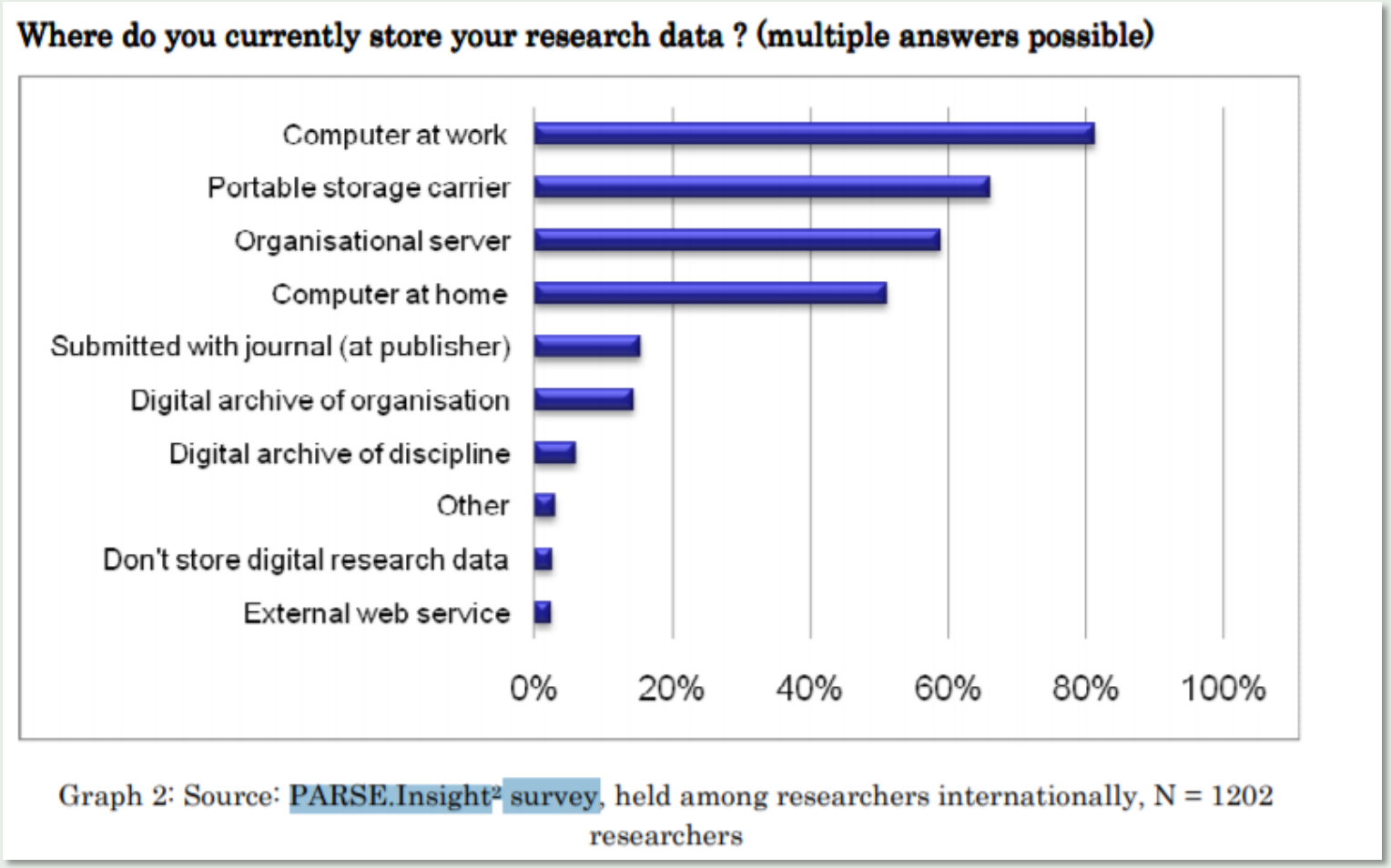
Conséquences

- ✓ 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



Préservation : des données en danger ?

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

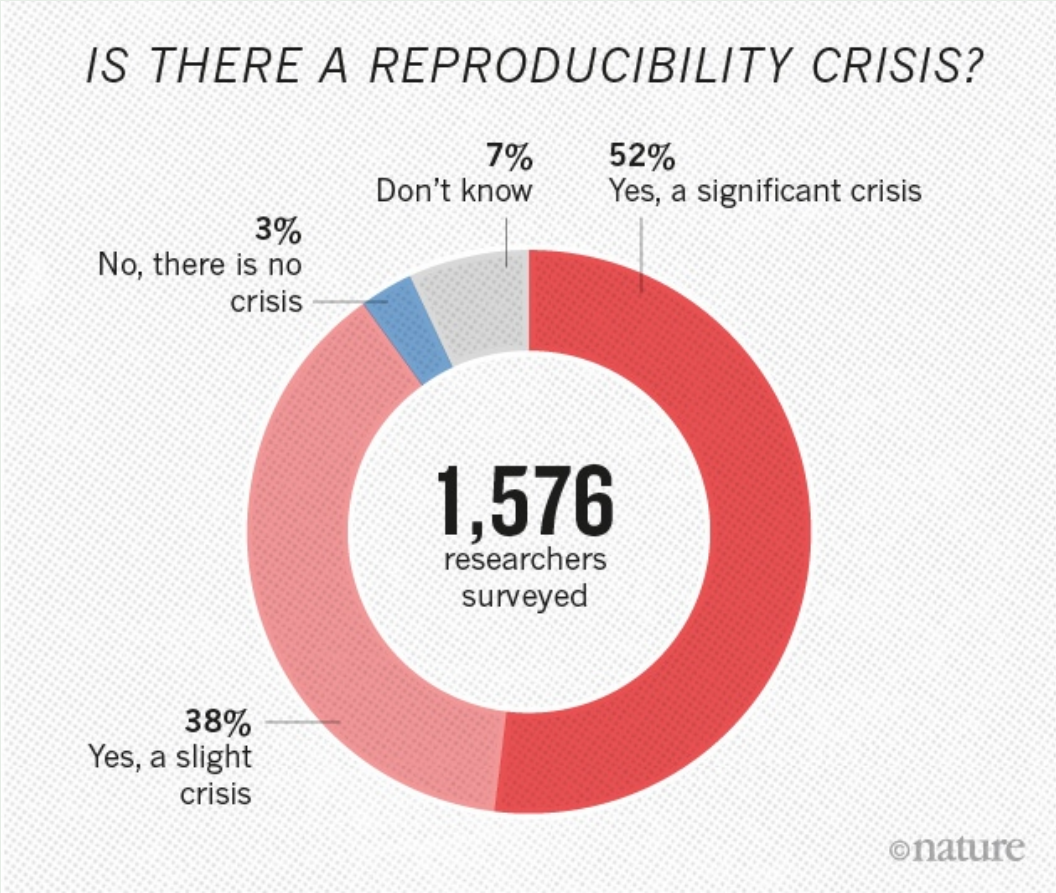


Graph 2: Source: PARSE.Insight[2] survey, held among researchers internationally, N = 1202 researchers

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"

Nature May 2016 : https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Pourquoi *partager* les données ?

entre injonctions et bénéfices

RESEARCH & INNOVATION
European Commission
Participant Portal H2020 Online Manual

ANR
Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION
NERC SCIENCE OF THE ENVIRONMENT
wellcome trust

Incitations, obligations de partage

ELSEVIER
PLOS
SPRINGER NATURE
Wiley
INFORMATION PARTICIPATION JUSTICE
AARHUS CONVENTION for our environment
Protocole de Nagoya

- ✓ 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%)

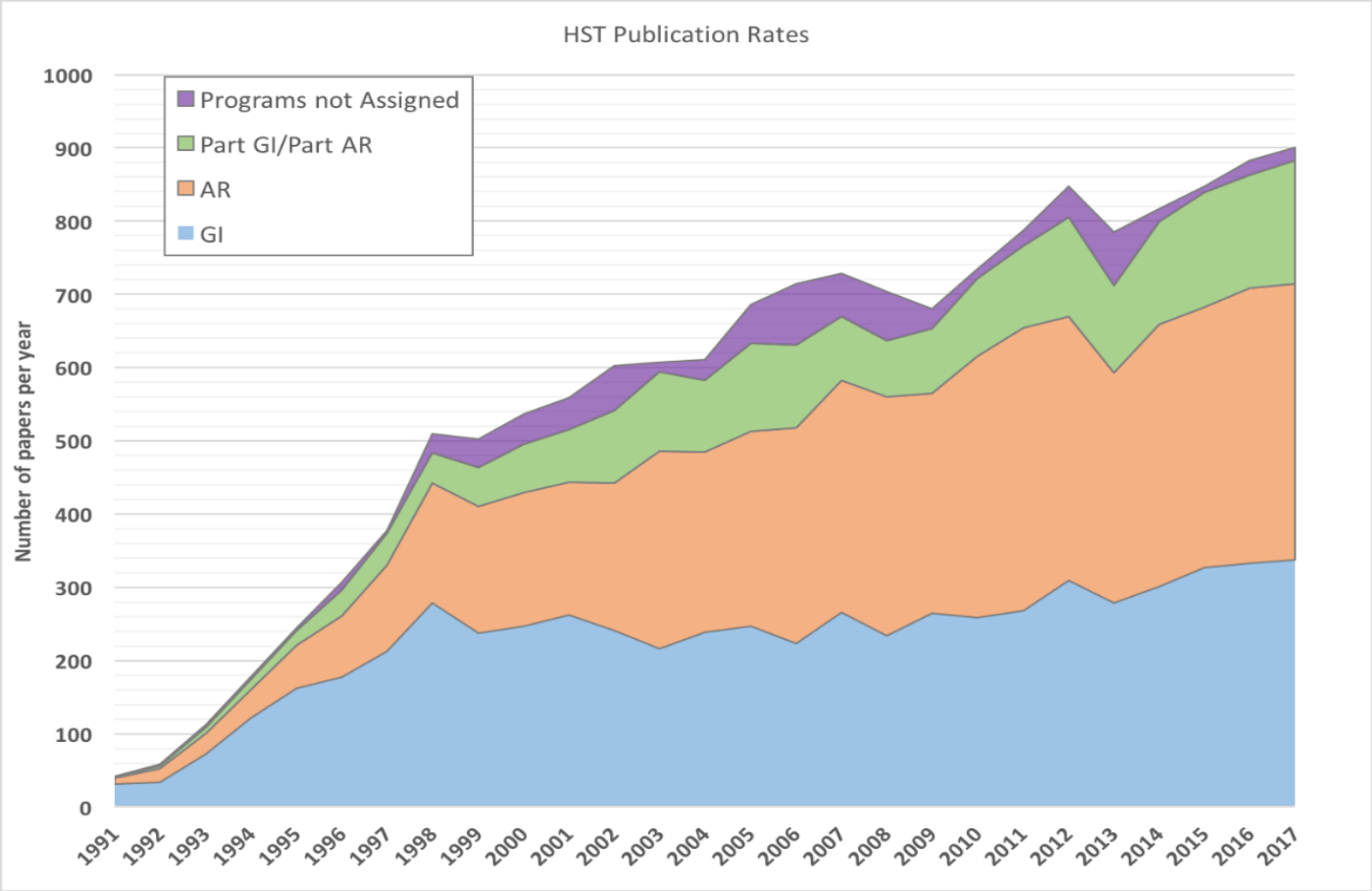
Étendre son réseau
Nouvelles opportunités pour échanger/collaborer

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

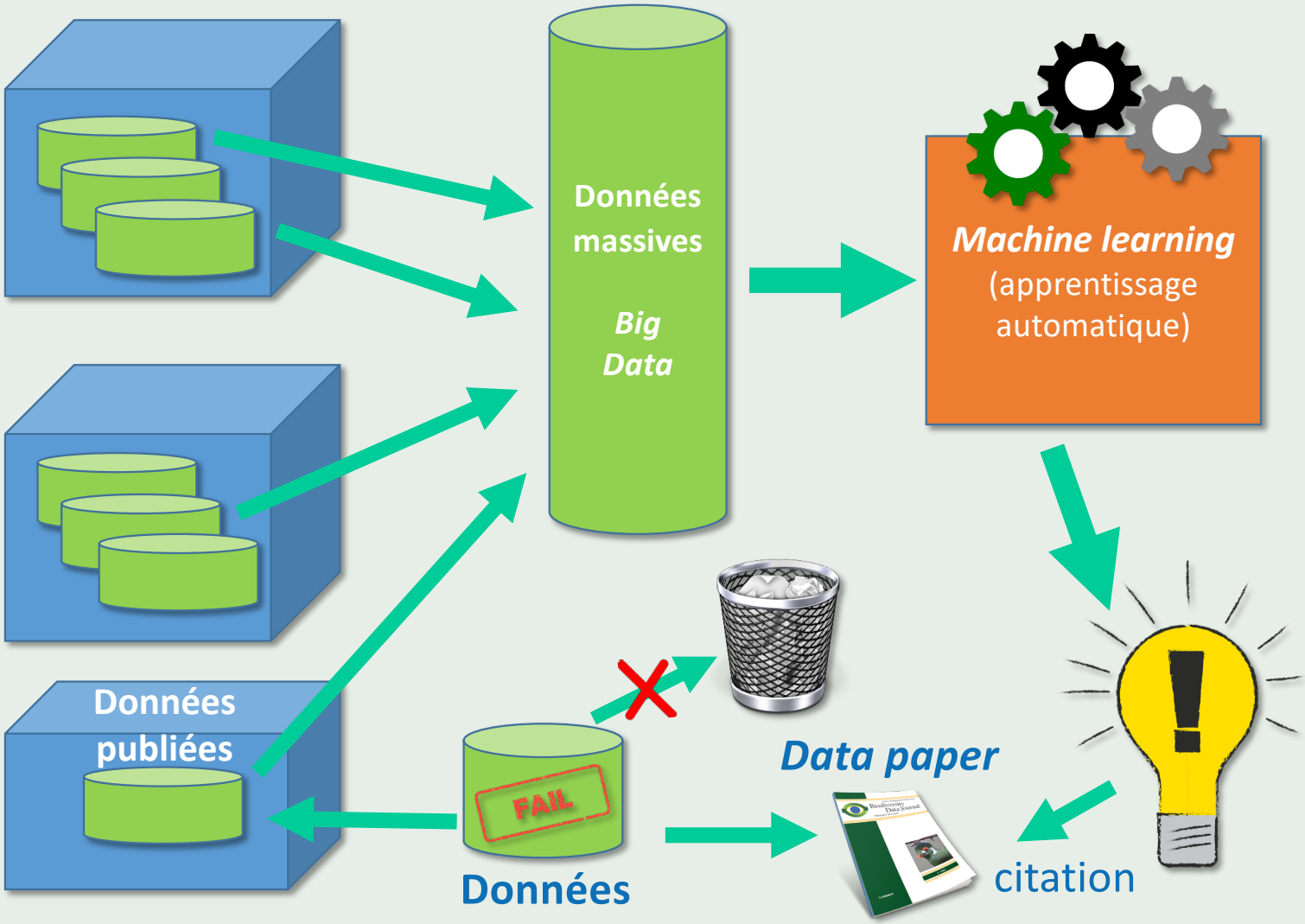
Données archivées : davantage utilisées et citées que les données récentes

Exemple : réutilisation des données en Astronomie

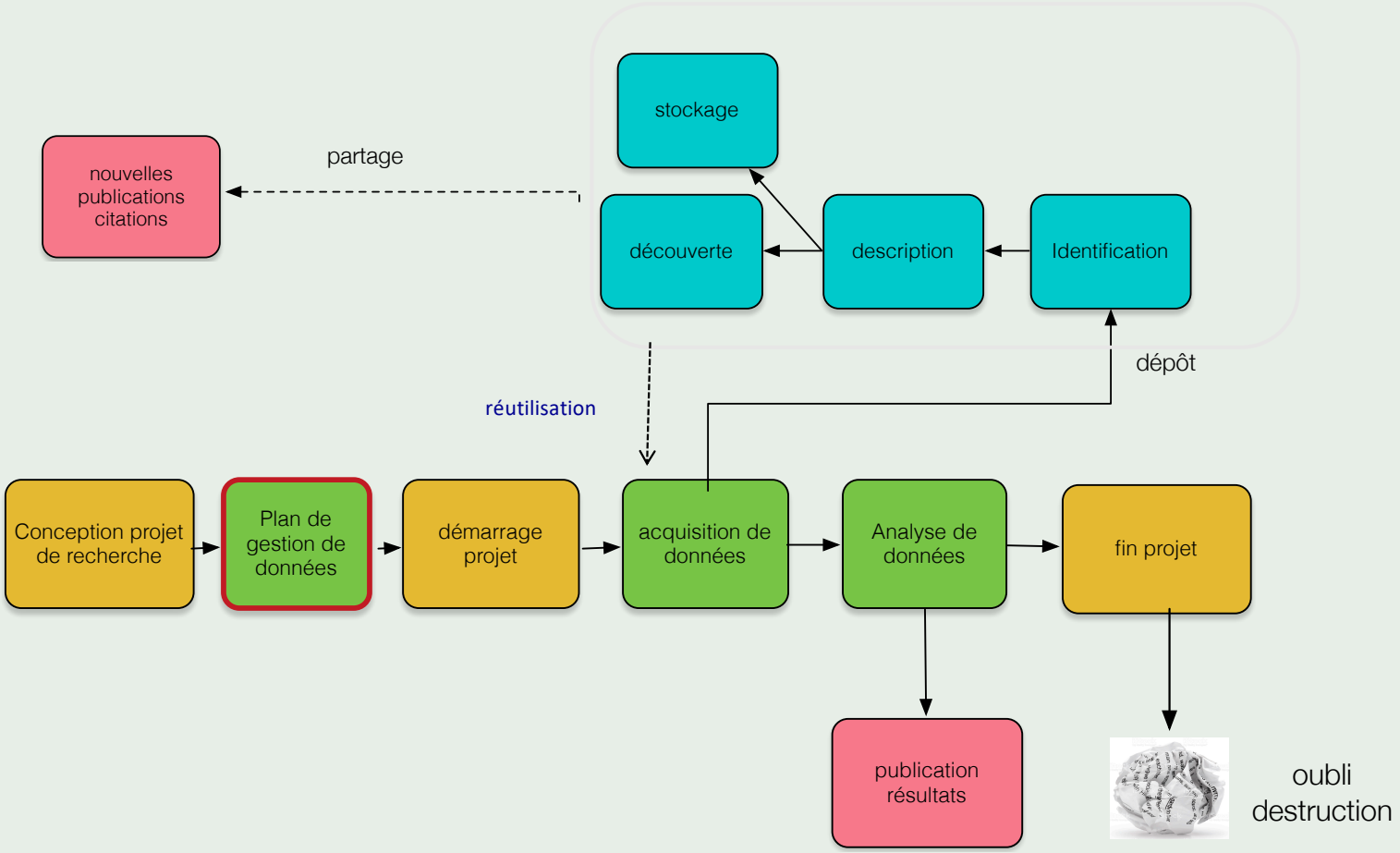


https://archive.stsci.edu/hst/bibliography/pubstat.html

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 »