Projet Open OPSE

Entrepôts de données : objectifs et principes de fonctionnement

Luc DECKER
IRD (service IST – MCST)
data@ird.fr

CC BY SA

Projet Open OPSE

Entrepôts de données de la Recherche

Définition - Principales fonctions

Service en ligne qui permet de…

- déposer
- décrire
- conserver
- référencer
- diffuser
- rechercher

des **jeux de données**

4TU.Centre for Research Data

Portail des données marines
Institut français de recherche pour l'exploitation de la mer

UK • DATA ARCHIVE 1967-2017

zenodo

DRYAD

opernicus Europe's eyes on Earth

UNC Dataverse

PDBe Protein Data Bank in Europe

ads ARCHAEOLOGY DATA SERVICE

CSIRO Data Access Portal

ⓘ La diversité des entrepôts de données est l'objet de la prochaine partie de la formation

Données, articles : convergence ?

Les journaux scientifiques
publient des articles…

Les entrepôts
publient des données !

- Une telle approche donne une place centrale à la **qualité** des données et de leurs descriptions
  – en lien avec la **réputation** d'un entrepôt

Projet Open OPSE

A quoi sert un entrepôt de données ?

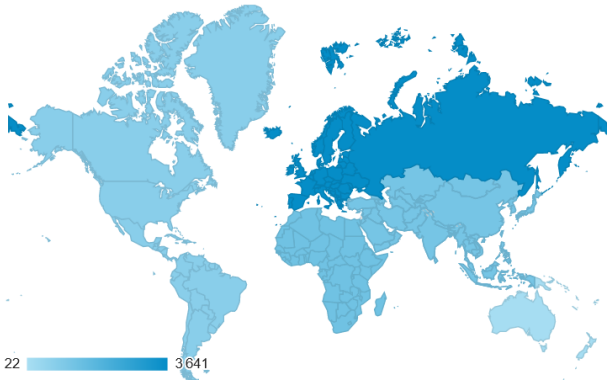
(rappel des enjeux)

- ✓ **Préservation** des données sur le long terme
- ✓ **Visibilité, partage** et **accès** aux données des laboratoires et projets de recherche
- ✓ **Valorisation**, en particulier susciter des collaborations entre recherche publique et secteur privé
- ✓ **Maîtrise de la diffusion** des données : licences et niveaux d'accès
- ✓ **Ethique** : rendre les données plus facilement accessibles aux scientifiques du Sud

Projet Open OPSE

Rendre accessible les données… partout dans le monde

Exemple : utilisation de l'entrepôt DataSuds



Statistiques Janvier 2020 – Août 2021

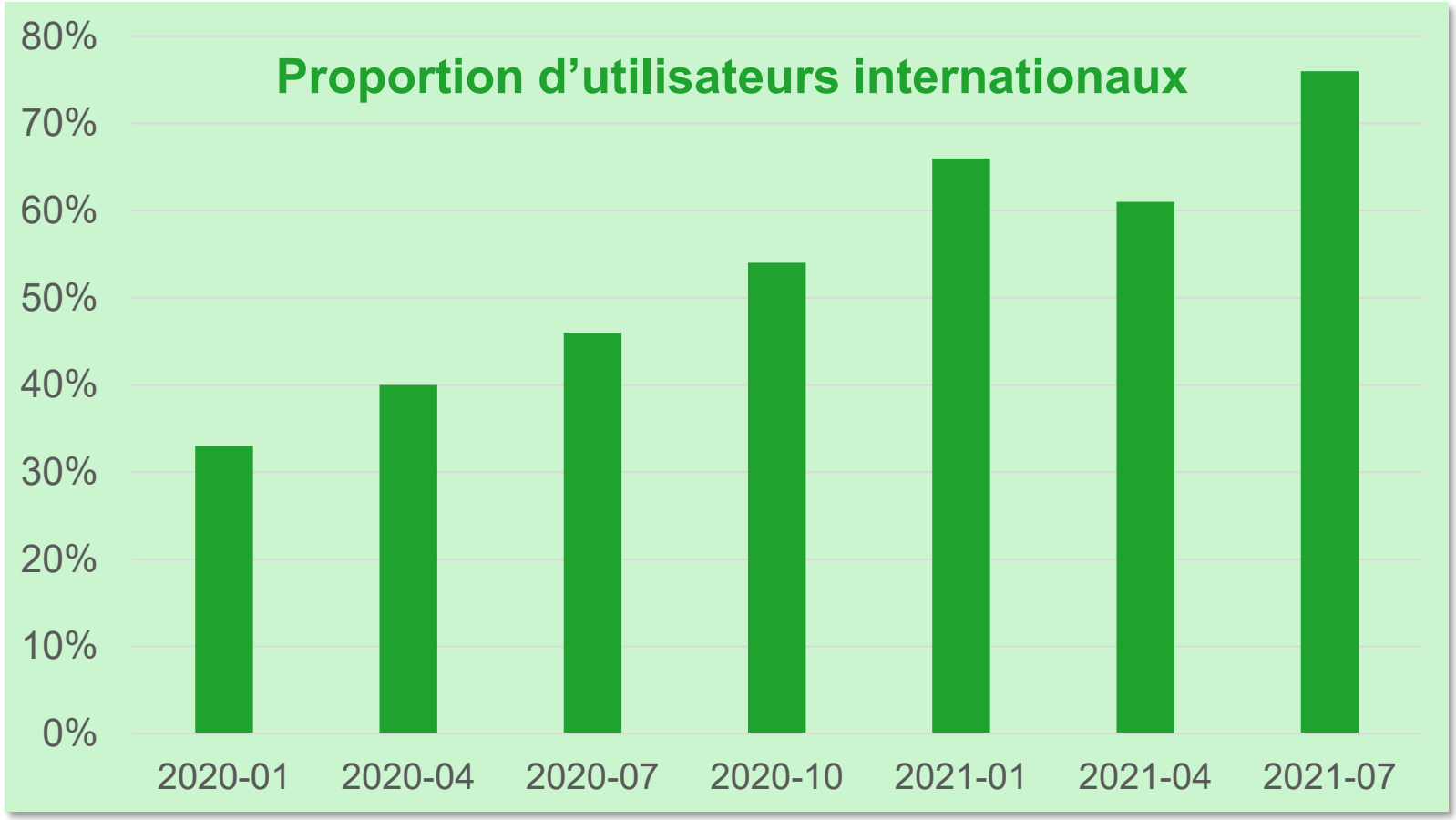
1.	Europe	**3 641** (52,16 %)
2.	Africa	**1 300** (18,62 %)
3.	Asia	**1 152** (16,50 %)
4.	Americas	**754** (10,80 %)
5.	Oceania	**112** (1,60 %)

1.	France	**2 734** (38,96 %)
2.	United States	**372** (5,30 %)
3.	Vietnam	**259** (3,69 %)
4.	India	**239** (3,41 %)
5.	Côte d'Ivoire	**222** (3,16 %)
6.	United Kingdom	**179** (2,55 %)
7.	China	**158** (2,25 %)
8.	Senegal	**128** (1,82 %)
9.	Germany	**126** (1,80 %)
10.	Brazil	**110** (1,57 %)
11.	Morocco	**108** (1,54 %)
12.	Cameroon	**102** (1,45 %)
13.	Canada	**100** (1,43 %)
14.	Netherlands	**94** (1,34 %)
15.	Italy	**93** (1,33 %)

Projet Open OPSE

Rendre accessible les données… partout dans le monde

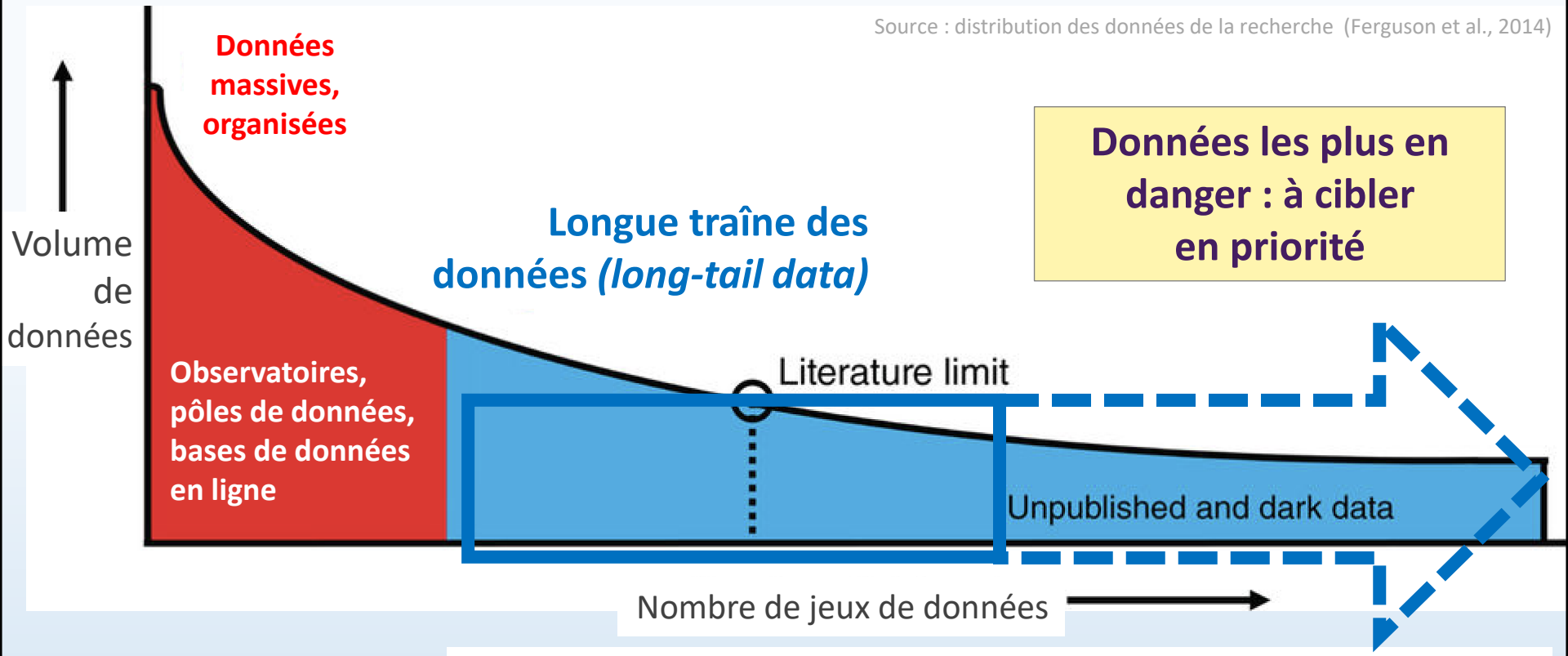
Exemple : utilisation de l'entrepôt DataSuds



Positionnement des entrepôts de données

Quelle complémentarité avec les grandes infrastructures de données ?

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



Données non structurées, non numérisées, enfouies dans les PCs, historiques, non rattachées à des bases accessibles en interne ou externe, …

Projet Open OPSE

Déposer des données personnelles sensibles ?

Plusieurs approches sont possibles :

1) **Uniquement la description des données** (métadonnées) et les **documentations** ⇨ **Visibilité**
2) **Données pseudonymisées** ou **dé-identifiées**, en **accès restreint** : accès sous conditions, engagement signé par l'utilisateur des données (confidentialité), …
3) **Données complétement anonymisées en accès libre – difficile à réaliser** et **risqué**

« FAIRiser » les données :
Aussi ouvert que possible, aussi fermé que nécessaire

Tenir également compte des dispositions prévues par le **formulaire de consentement**

Projet Open OPSE

Elément de base d'un entrepôt : le jeu de données (*dataset*)

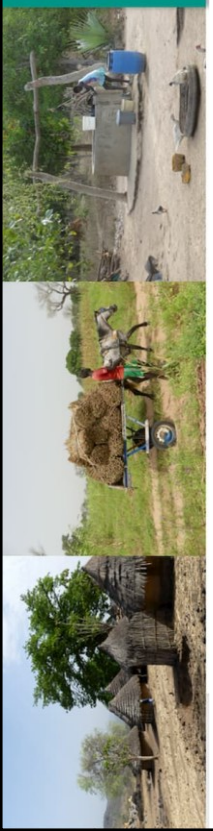
Regroupement cohérent de « données » :

- Fiche descriptive : **métadonnées**

  TITRE **AUTEURS**
  RESUME MOTS-CLES
  etc…
  IDENTIFIANT UNIQUE / REFERENCE

- **Fichiers** de données *(éventuellement aucun)*
- **Documentations** associées

+ conditions d'utilisation
+ historique des versions, en cas de mises à jour



Elément de base d'un entrepôt : le jeu de données (*dataset*)

Regroupement cohérent de « données »

➔ Quels critères de cohérence ?

Quelques exemples :

- ❑ Données associées à un article
- ❑ Données d'un projet de recherche
- ❑ Données d'un composant d'un projet
- ❑ Données d'une station de mesure

Regroupements géographiques, temporels, par type de données, …

Comment choisir le bon niveau de granularité d'un dépôt de données ?

Jeu de données simple ?	Découpage en multiples jeux de données
Citation (titre, liste des auteurs) et identifiant pérenne unique. Licence d'utilisation souvent unique.	**Citation (titre, liste des auteurs) et identifiant pérenne propres à chaque partie.** **Pertinence scientifique ?**
Permet de **télécharger toutes les données en une seule étape** : plus pratique pour les utilisateurs	Téléchargements individualisés
Préparation simplifiée	**Préparation plus complexe**, nécessité d'une **réflexion initiale** **sur le mode de découpage (géographique, thématique…)** 
Gestion simplifiée et centralisée	**Gestion indépendante des différentes parties** • publications et mises à jour différenciées • restrictions d'accès spécifiques
	Statistiques de consultation et de téléchargement plus détaillées

Une question complexe : pas de réponse unique

Projet Open OPSE

Préparer un jeu de données : questions à se poser

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

ⓘ La bonne préparation des jeux de données est l'objet d'un chapitre suivant de la formation

Où citer un jeu de données ?

- Articles scientifiques
  (citent les données et réciproquement)
- Rapports d'avancement de projet
- Rapports d'activité
- Dossiers de candidature

…

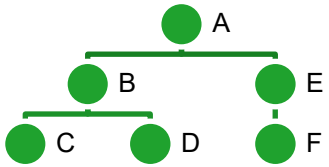
ⓘ Pour aller plus loin : *Principes Data Citation*

Projet Open OPSE

Comment organiser les jeux de données dans un entrepôt ?

Pour structurer les données, certains entrepôts permettent :

- de regrouper les jeux de données dans des **collections** (ou dossiers)
- de créer des sous-collections

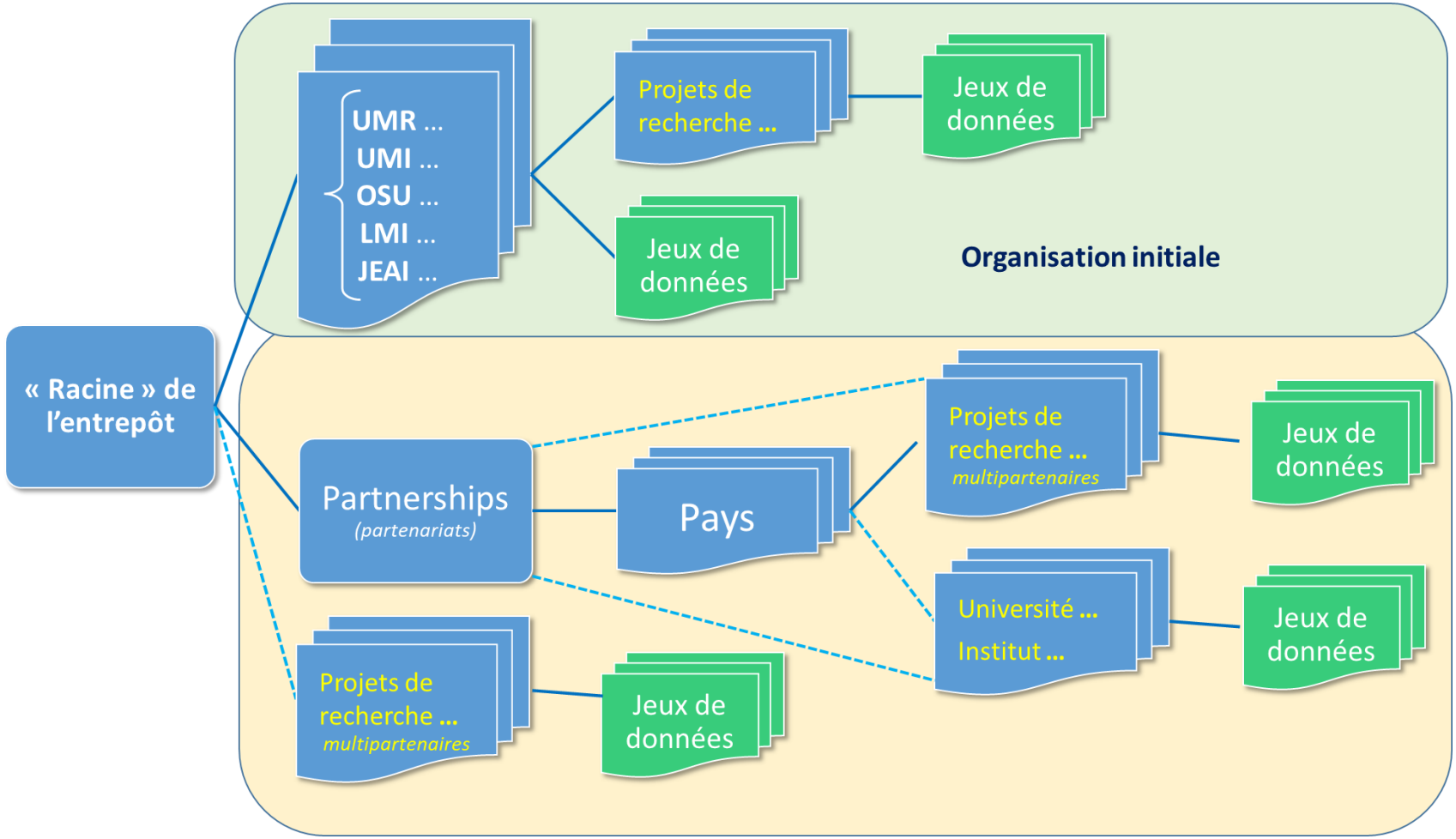


Une collection peut par exemple être dédiée

- à une thématique scientifique
- à un projet de recherche
- à un laboratoire de recherche

Projet Open OPSE

Exemple : structure de l'entrepôt DataSuds



Quizz

https://myquiz.org code 783475