

Retour d'expérience : l'entrepôt de données DataSuds de l'IRD

Luc Decker

Service Information Scientifique et Technique (IST),
Mission Culture Scientifique et Technique - IRD



dataverse.ird.fr
data@ird.fr

C@fé Renatis, 28 avril 2022



DataSuds, entrepôt de données scientifiques de l'IRD

DataSuds (IRD) [En savoir plus...](#)

Localisation des jeux de données

DataSuds
L'entrepôt de données DataSuds propose aux scientifiques de l'IRD et à leurs partenaires un service pour diffuser, préserver et valoriser leurs données de recherche en facilitant leur identification et leur citation. Il constitue l'un des éléments du dispositif de la science ouverte pour le Sud que promeut l'IRD.
[Mode d'emploi](#) [? assistance](#)

Rechercher
Chercher dans tous les dataverses... [Explorer](#)

Par Sujet

- Earth and Environmental Sciences 89
- Agricultural Sciences 43
- Medicine, Health and Life Sciences 24
- Social Sciences 19
- Computer and Information Science 7
- Mathematical Sciences 2
- Other 2
- Arts and Humanities 2
- Law 1

Derniers jeux de données

- France-CGIAR BRIDGE collaboration: integrated model-based... [1:4](#)
Equipe CERE8 [Agricultural Sciences](#) [Thu Sep 16 2021](#)
- Anthropogenic fibers (quantity, size, colour) observed in whole... [1:4](#)
Microplastic in Aquatic Environments [Earth and Environmental Sciences](#) [Fri Sep 10 2021](#)
- User centric metadata model [1:1](#)
UCMM [Computer and Information Science](#) [Earth and Environmental Sciences](#) [Tue Sep 07 2021](#)
- Replication Data for the manuscript Selection on ancient gene... [1:6](#)
Popolice genetic and genomic [Medicine, Health and Life Sciences](#) [Thu Sep 02 2021](#)
- Paysages vivants du versant Ouest de l'île de La Réunion, e... [1:16](#)
UMR ESPACE-DEV [Arts and Humanities](#) [Earth and Environmental Sciences](#) [Social Sciences](#) [Wed Jul 21 2021](#)
- Passive case detection of Human African Trypanosomiasis i... [1:6](#)
DITECT-HAT-WP2 [Medicine, Health and Life Sciences](#) [Tue Jul 20 2021](#)

Plateforme de
publication de
données

Principes
FAIR

Pérenne

Institutionnel

Thèmes de
recherche des
structures IRD

Gratuit

Dépôt ouvert aux
scientifiques des
structures IRD et aux
partenaires du Sud

Accompagnement :
prise en main, édition,
conseil, guides...

dataverse.ird.fr

DataSuds : coordination, développement et gestion

un projet collectif, transversal



Mission Science Ouverte (MSO)

Jean-Christophe Desconnets, François Sabot

Direction du Développement des Usages Numériques Innovants (DDUNI)

Thibault Coupin, Sylvain Dienst, Stéphane Banon

Mission Culture Scientifique et Technologique (MCST), Service IST

Pascal Aventurier, Caroline Doucouré, Hanka Hensens, Luc Decker

avec l'appui de **DAJ** (Affaires Juridiques), **DMOB** (Service Innovation et Valorisation)
et **DRH** (Service Développement des talents)

DataSuds : un entrepôt de données, 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

DataSuds : phase de préparation, orientations initiales

Exigences fonctionnelles

Accueil de données pluridisciplinaires
Attribution de DOIs et de citations aux jeux de données
Interface web permettant aux chercheurs de déposer
Gestion des droits au niveau des collections de données
Interopérabilité, standards de métadonnées et référencement par les moteurs de recherche
Aide au respect des principes FAIR

Préférences techniques

Solution *open-source* : adaptable, communauté de développeurs et d'utilisateurs

Partenaires, coopérations potentielles

Entrepôts existants, en particulier à l'INRAE et au CIRAD

Choix technique



www.dataverse.org

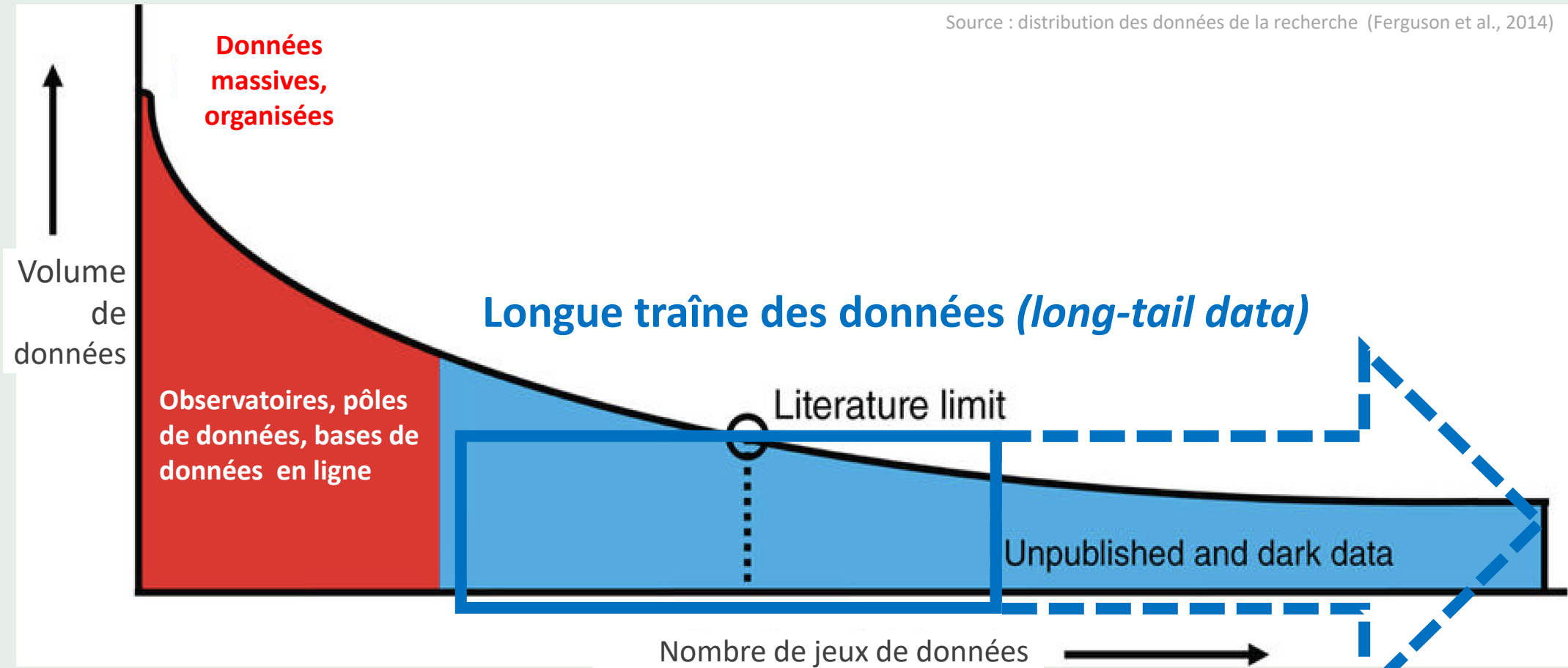
Organisation

- Comité de Pilotage
- Equipe projet
- Réseau de référents dans les Unités IRD
- Support, formations

Lancement : 09/2019

Quelles données ? positionnement de l'entrepôt DataSuds

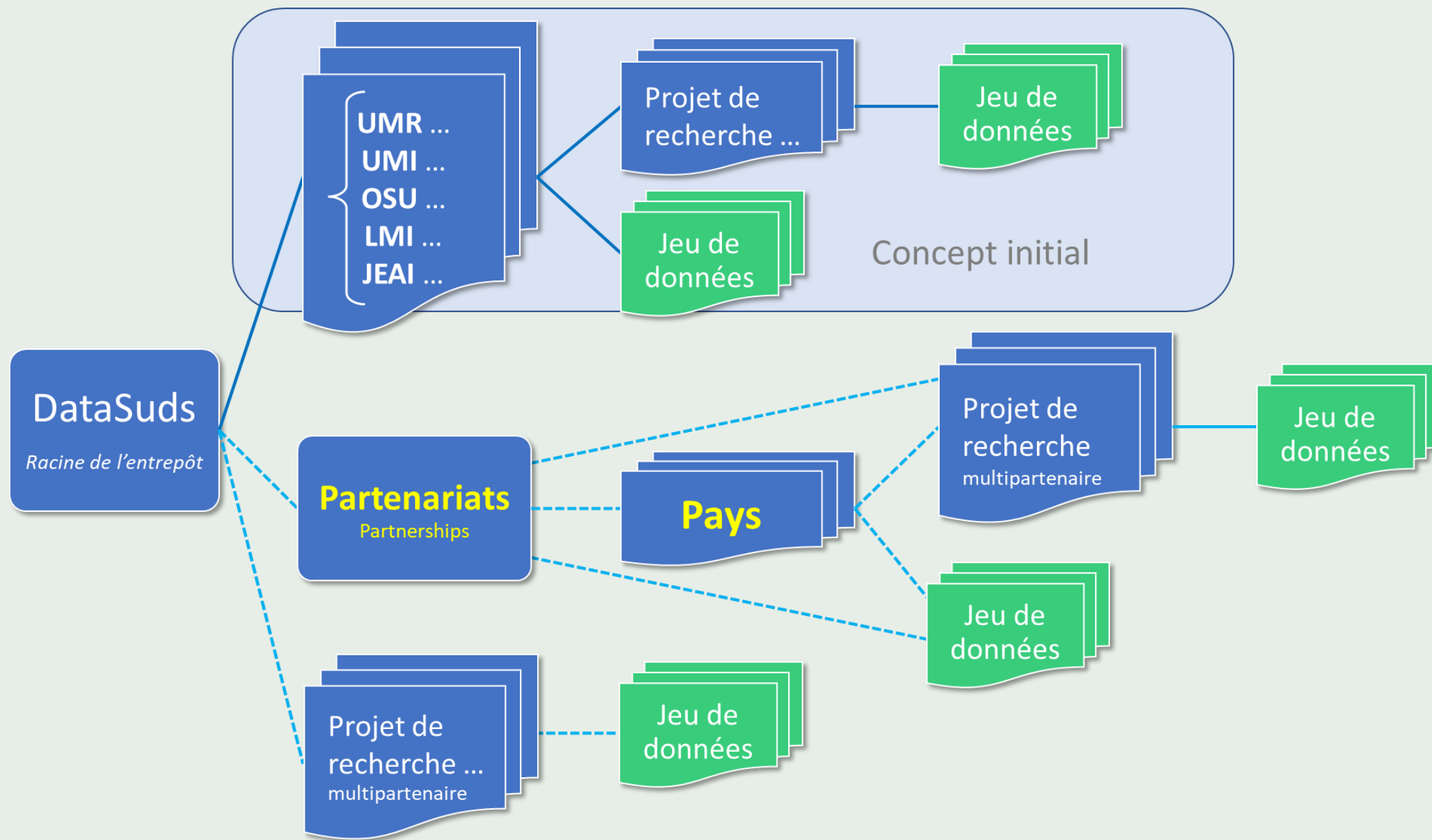
Complémentarité avec les grandes infrastructures de données



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

Données les plus en danger : à cibler en priorité

Organisation des collections de données dans l'entrepôt DataSuds

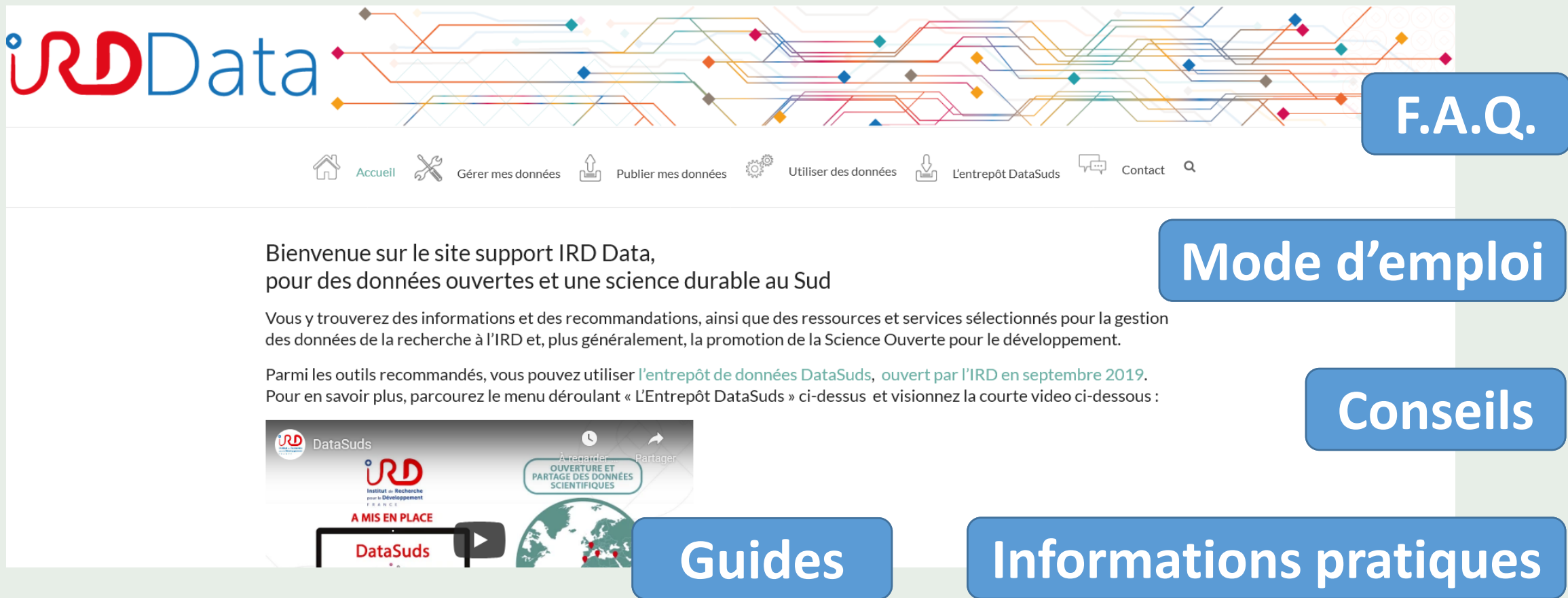


Développement d'un réseau de référents des données scientifiques

- Appels à la nomination de référents « Données scientifiques » envoyés à toutes les Unités de recherche IRD : correspondants DataSuds
- 59 référents nommés par 45 Unités de recherche
- Formations depuis 2019 : 32 ateliers pratiques
 - ✓ Généralités sur les entrepôts de données, enjeux et objectifs
 - ✓ Présentation de DataSuds, ses fonctionnalités et son organisation
 - ✓ Travaux pratiques : apprendre à administrer des collections « *dataverse* »
 - ✓ Introduction aux Plans de Gestion de Données
 - ✓ Points d'attention juridiques avant publication de données
- Séminaires des référents : 2020, 2022 (prévu)

DataSuds : accompagnement des déposants

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



The screenshot shows the IRD Data website interface. At the top, the logo 'IRD Data' is displayed next to a decorative graphic of colorful circuit lines. Below the logo is a navigation bar with icons and labels: Accueil, Gérer mes données, Publier mes données, Utiliser des données, L'entrepôt DataSuds, Contact, and a search icon. On the right side of the page, there is a welcome message in French, followed by a paragraph about the site's purpose and a video player. Overlaid on the right side of the screenshot are five blue buttons with white text: 'F.A.Q.', 'Mode d'emploi', 'Conseils', 'Guides', and 'Informations pratiques'.

F.A.Q.

Bienvenue sur le site support IRD Data,
pour des données ouvertes et une science durable au Sud

Vous y trouverez des informations et des recommandations, ainsi que des ressources et services sélectionnés pour la gestion des données de la recherche à l'IRD et, plus généralement, la promotion de la Science Ouverte pour le développement.

Parmi les outils recommandés, vous pouvez utiliser l'entrepôt de données DataSuds, ouvert par l'IRD en septembre 2019. Pour en savoir plus, parcourez le menu déroulant « L'Entrepôt DataSuds » ci-dessus et visionnez la courte video ci-dessous :

Mode d'emploi

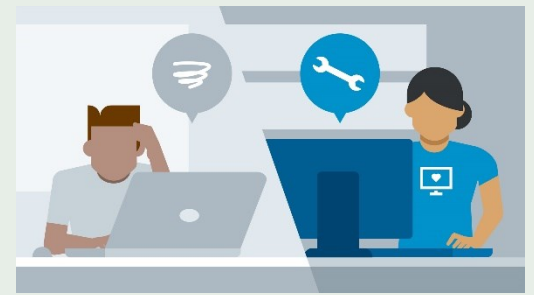
Conseils

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

DataSuds – Bilan : dépôts de données

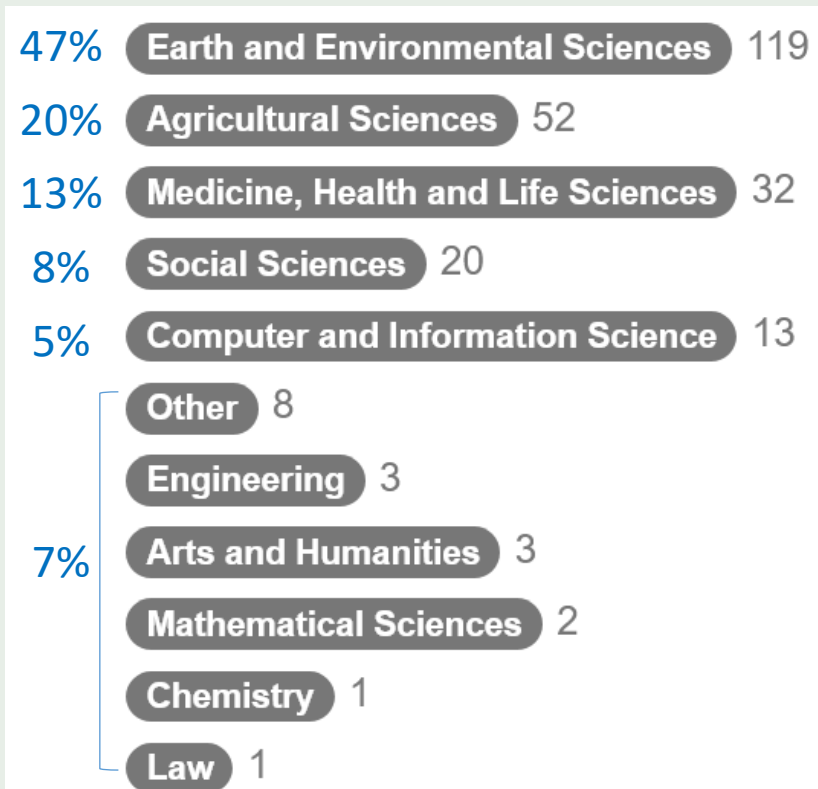
175 jeux de données publiés

+ 40 jeux de données *moissonnés* à partir d'autres entrepôts

+ environ 80 jeux de données en préparation

89 collections de données publiques
dont 30 collections principales « UMR »

Classement par
sujets des jeux de
données publiés

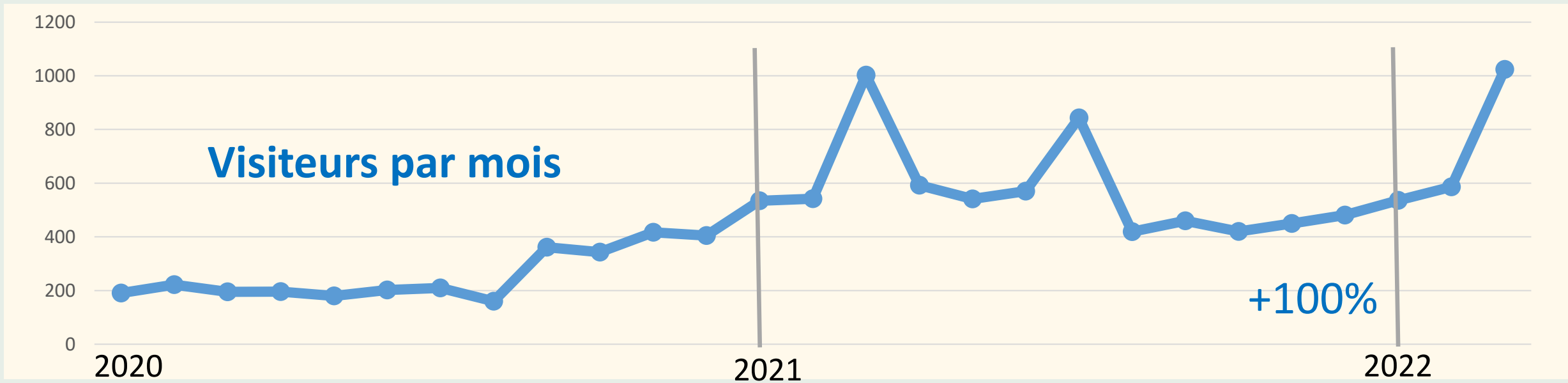


UMR ESPACE-DEV (IRD, Univ.Montpellier, Univ.La Réunion, Univ.Guyane, Univ.Antilles) 11 juil. 2019  L'UMR ESPACE-DEV développe des recherches sur les dynamiques spatiales caractérisant les éco-sociosystèmes. Ses objectifs concernent la définition d'indicateurs de ces dynamiques : bio-géophysiques, évolutions des sociétés, risques liés aux maladies émergentes en fonction de para...	11
UMR AMAP (CIRAD, CNRS, INRAE, IRD, Univ.Montpellier) 29 juil. 2019  UMR) is an interdisciplinary laboratory that conducts basic research on plants and plant communities	8
UMR iEES-Paris (Sorbonne Univ., CNRS, INRAE, IRD, Univ. de Paris, UPEC) 6 août 2019  Institut d'écologie et des sciences de l'environnement de Paris Cet espace rassemble les données de l'UMR iEES Paris. Contactez-nous à l'adresse iees.data@ird.fr si vous voulez publier des données sur cette plate-forme.	8
UMR LISAH (INRAE, IRD, Institut Agro) 2 sept. 2019  Laboratory for the study of Soil-Agrosystem-Hydrosystem interactions LISAH researches deals with hydrology, contaminants transfer and erosion in cultivated landscapes, and on the spatial properties and temporal changes of the landscapes, in order to address following an integrate...	8
UMR MERIT (IRD, Université Paris Descartes) 6 sept. 2019  Mère et enfant en milieu tropical : pathogènes, système de santé et transition épidémiologique L'unité de recherche MERIT est une unité mixte de recherche entre l'Institut de Recherche pour le Développement (IRD) et l'Université Paris Descartes. Elle conduit des programmes de rec...	8
UMR CBGP (INRAE, IRD, CIRAD, Montpellier SupAgro) 15 oct. 2019  Centre de Biologie pour la Gestion des Populations Le CBGP développe des recherches en systématique, génétique, écologie et évolution des populations et des communautés d'organismes d'intérêt en agronomie, santé humaine et biodiversité. Les petits mammifères, rongeurs en particul...	8
UMR PRODIG (CNRS, Univ. Paris 1 Panthéon-Sorbonne, Univ. Paris Diderot, IRD, AgroParisTech) 31 oct. 2019  Development, Global Change and Territorial Dynamics	8
UMR Eco&Sols (Univ.Montpellier, SupAgro, CIRAD, INRAE, IRD) 13 nov. 2019  Functional ecology and biogeochemistry of soils and agro-ecosystems Eco&Sols is a Joint Research Unit with 70 permanent staff from INRAE, IRD, Montpellier SupAgro and CIRAD. The research undertaken by Eco&Sols sets out to improve the understanding of the role of soil organisms an...	8
UMR DIADE (IRD, University of Montpellier, Cirad) 15 nov. 2019  Diversity - Adaptation - plant DEvelopment The DIADE Research Unit aims to understand the diversification of tropical plants, one of the main original reservoirs of biodiversity, and for which conservation, management and exploitation are an important issue for Sustainable Develo...	8
UMR INTERTRYP (IRD, CIRAD) 11 déc. 2019  Interactions Hôte-Vecteur-Parasite-Environnement dans les maladies tropicales négligées dues aux Trypanosomatidae	8

DataSuds – Bilan : dépôts de données

- Volume des dépôts inférieur aux attentes
- **Principales motivations des dépôts**
 1. **Publications de données associées à des articles :**
demandes des éditeurs (*journal data policies, déclarations de disponibilité des données*)
 2. **Obligations contractuelles vis-à-vis des bailleurs**
 3. Préservation de données avant départs à la retraite
 4. Adhésion aux principes de la Science Ouverte
 5. Dans le futur ? Plans de Gestion de Données
- Challenge : convaincre les scientifiques des bénéfices de publier des données dans un entrepôt

DataSuds – Progression des consultations



janvier 2020 – mars 2022

10,400 visiteurs

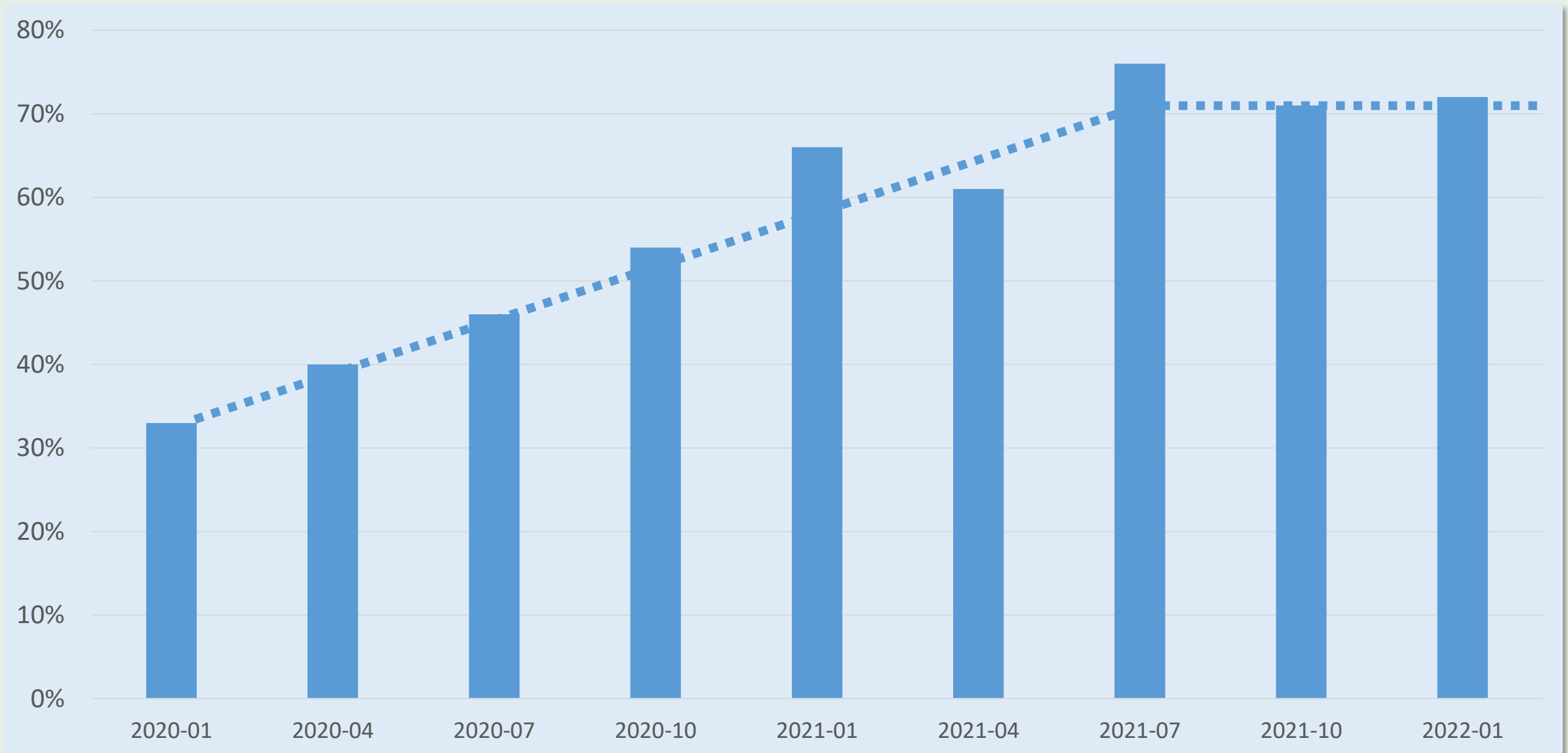
21,500 visites

98,000 pages vues

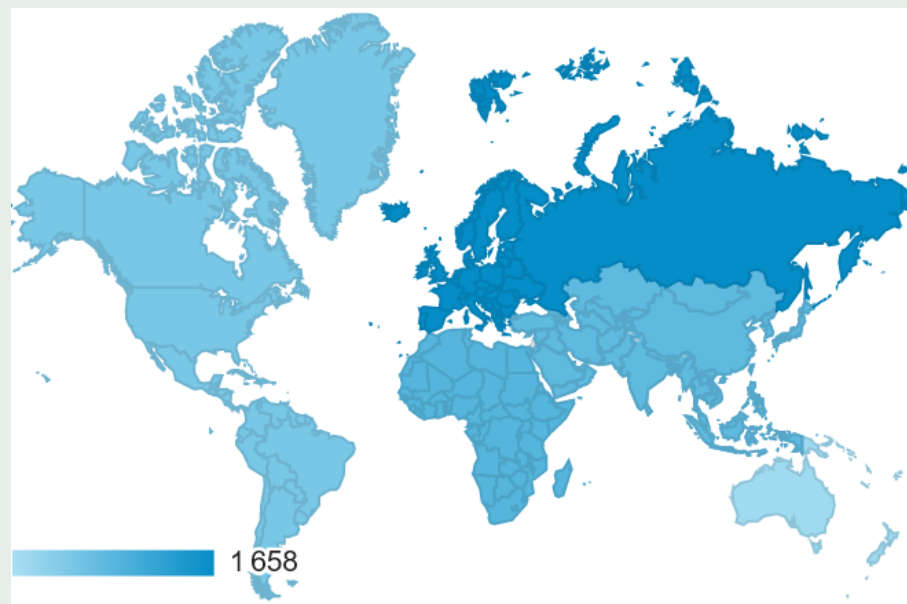
16,500 téléchargements de fichiers

6 minutes/visite, en moyenne

DataSuds : proportion croissante des utilisateurs internationaux



DataSuds : des « utilisateurs de données » partout dans le monde



1. Europe	1 658 (42,66 %)
2. Africa	861 (22,15 %)
3. Asia	752 (19,35 %)
4. Americas	498 (12,81 %)
5. Oceania	112 (2,88 %)

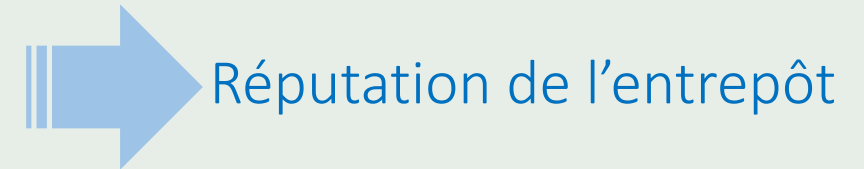
Statistiques
juillet 2021 – mars 2022

1.  France	1 031 (26,45 %)
2.  Côte d'Ivoire	244 (6,26 %)
3.  United States	231 (5,93 %)
4.  China	155 (3,98 %)
5.  Germany	114 (2,92 %)
6.  India	113 (2,90 %)
7.  Vietnam	103 (2,64 %)
8.  United Kingdom	98 (2,51 %)
9.  Brazil	77 (1,98 %)
10.  Senegal	69 (1,77 %)
11.  Cameroon	65 (1,67 %)
12.  Canada	59 (1,51 %)
13.  Spain	58 (1,49 %)
14.  Italy	49 (1,26 %)
15.  Morocco	49 (1,26 %)

DataSuds : une approche qui a évolué au fil des utilisations

...davantage considéré comme une **plateforme de publication de données**, plus en parallèle avec les publications d'articles scientifiques :

- ➔ **Qualité des métadonnées et qualité éditoriale**
- ➔ **Curation (édition) et modération des dépôts**
- ➔ **Guide de relecture avant publication**



[doi:10.5281/zenodo.5884671](https://doi.org/10.5281/zenodo.5884671)

« *Lessons learned...* »

Alimenté par la curation
de 150 jeux de données

Principes **FAIR**

Principes **Data Citation**

Qualité éditoriale

Bonnes pratiques

Guide de relecture d'un jeu de données avant publication (Qualité de l'entrepôt de données **DataSuds**)

Version 2.4, janvier 2022 – Luc Decker (administrateur de DataSuds <https://dataverse.ird.fr/>),
Hanka Hensens, Caroline Doucouré et Pascal Aventurier – Service IST/MCST, IRD, France



Métadonnées

Champs	Réf.	Préconisations, recommandations - Conseils pratiques	Finalité et commentaires
tous	L1	De préférence, saisir toutes les informations en anglais. Dans ce cas, l'interface en anglais facilite la saisie : avant d'ouvrir les formulaires, changer de langue grâce au menu en haut à droite de l'écran.	Visibilité et valorisation du jeu de données : recommandé, comme pour les articles scientifiques.
	L2	Ne pas mélanger différentes langues, sauf éventuellement pour le champ « Description » (ajouter une séparation entre les langues avec le code <hr>) ainsi que pour le champ « Autre titre » (possibilité de renseigner une traduction : un titre dans une langue locale peut améliorer la visibilité du jeu de données.) et pour les mots-clés. La langue sélectionnée dans le menu s'applique uniquement à l'interface utilisateur.	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 des informations traduites dans différentes langues.
	L3	Orthographe et grammaire : vérifier tous les textes saisis dans les métadonnées. Lorsque les saisies sont terminées, se rendre dans l'onglet « Métadonnées », copier/coller tout le texte dans un logiciel de traitement de texte (tel que <i>Word</i>), sélectionner la langue utilisée et lancer la vérification automatique de l'orthographe.	Prouve que les informations ont été relues avant publication, témoigne du soin apporté dans la gestion des données. Réputation de l'entrepôt, des auteurs et de leur laboratoire.
titre	T1	Caractérisation des données : type de données, contexte, période de collecte, localisation géographique - si applicable et pertinent. Si un autre jeu de données risque d'avoir le même titre, alors le titre devrait être plus spécifique. Autres possibilités : « <i>Replication data for...</i> » / « <i>Supplementary 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 : 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. Le titre doit être compréhensible par lui-même.
	T2	Longueur appropriée, approximativement entre 3 et 20 mots	Suivre les usages, comme pour un article scientifique.
	T3	Retirer les éléments qui n'ont en général pas leur place dans un titre : noms de fichiers, noms d'auteurs, citation complète d'un article, codes ou références internes au projet, parenthèses inutiles, caractères spéciaux. Le nom du projet peut être inclus mais ne constitue pas un titre à lui seul.	

DataSuds : une approche qui a évolué au fil des utilisations

- **Plus de flexibilité** : adaptation à la variété des besoins des déposants et des projets de recherche, « au cas par cas » avec parfois des modalités d'utilisation non prévues initialement et des règles remises en question.
Exemple : publication de données à intervalles réguliers au cours de projets de recherche de type « observatoire ».
- **Plus de réactivité** dans la démarche de dépôts de données : Comment aller plus rapidement au but ? Anticiper pour simplifier et optimiser les échanges.
- Prise en compte de l'hétérogénéité du réseau des référents des données scientifiques à l'IRD : profils, compétences, expériences, disponibilité, domaines scientifiques... Le droit de publier n'est plus accordé systématiquement et immédiatement à la fin de la session (3h) de formation.

DataSuds : quoi de neuf depuis 2019 ?

- Page d'accueil : carte de localisation des jeux de données, liste des collections, statistiques
- Aides à la saisie des métadonnées : identifiants *ORCID* ; noms de pays, provinces et villes : appel à *GeoNames*
- Ouverture possible aux données de Santé, sous conditions (procédure en cours de rédaction)
- Nouvelles versions de l'application *Dataverse* : ergonomie et apparence améliorées

Des données ouvertes pour une science durable au Sud

DataSuds (FR) En savoir plus...

Localisation des jeux de données



DataSuds

Rechercher

Par Sujet

Carte et Environmental Sciences 10
Agricultural Sciences 14
Medicine, Health and Life Sciences 24
Social Sciences 13
Computer and Information Sciences 7
Environmental Sciences 7
Other 2
Arts and Humanities 2
LSD 1

Demiers jeux de données

France CGARR BRIDGE collaboration: integrated model-based app...
Exploit CERN 5
Thu Sep 16 2021

Antropogenic fibers (quantity, size, color) observed in whole bod...
Mouvement in Aquatic Environment
Fri Sep 10 2021

User-centric metadata model
ISMM
Tue Sep 07 2021

Replication Data for the manuscript Selection an ancient genetic va...
Populism genetics and genetics
Thu Sep 02 2021

Paragraphe viants du versant Ouest de l'île de La Réunion, entre M...
UMR ESPACE-DEV
Wed Jul 27 2021

Positive Case detection of Human African Trypanosomiasis in Côte...
DITECT-1667-8072
Tue Jul 20 2021

Les Unités de recherche et Projets

Dataverse
Mission pour la Science ouverte (MISO)

AMAP
UMR AMAP

Hydro
UMR HydroSciences Montpellier

Intertryp
UMR INTERTRYP

MERIT
UMR MERIT

NUTRIPASS
UMR NUTRIPASS

Dataverse
The Development Fabric Knowledge Deck

EcoSols
UMR EcoSols

CBGP
UMR CBGP

ESPACE-DEV
UMR ESPACE-DEV

DIADÉ
UMR DIADÉ

Paddle
Paddle project

LMV
UMR LMV

ENTROPIE
UMR Entropie

MIVEGEC
UMR MIVEGEC

PHM
UMR PHM

Partnerships

UM Pathobios

EGCE
UMR EGCE

Rain Cell projects

LEDa
UMR LEDa

metra
UMR Metra

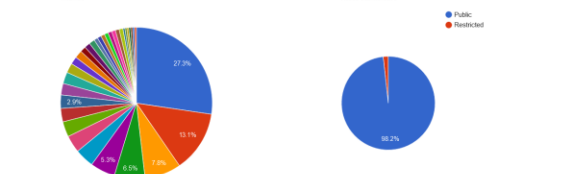
AMMA-CATCH
AMMA-CATCH associated projects

M-TROPICS
Data collected as part of M-TROPICS based research projects

Thèmes

Accès aux fichiers

Public
Restricted



DataSuds : perspectives

- Mise en service d'un outil de moissonnage de données
 - ➔ Référencer tout jeu de données déposé ailleurs à partir de son identifiant DOI
 - ➔ Utiliser DataSuds comme un catalogue/annuaire de données
- Aide à la saisie de mots-clés, connexion à un thesaurus (prévu 2022)
- Dossier de certification *Core Trust Seal*
- Migration du fond cartographique de l'IRD « *SPHAERA* »
- Renforcement des partenariats avec d'autres entrepôts Dataverse