

INFORMATION NUMÉRIQUE

Enjeux et Pratiques

Licence 3 / UE502 - Accès expert à l'information - 2024-2025

Monde de l'Information Scientifique et Technique



philippe.houdry@inist.fr
lucile.bourguignon@inist.fr
valerie.bonvallot@inist.fr

Open Data, Open Access



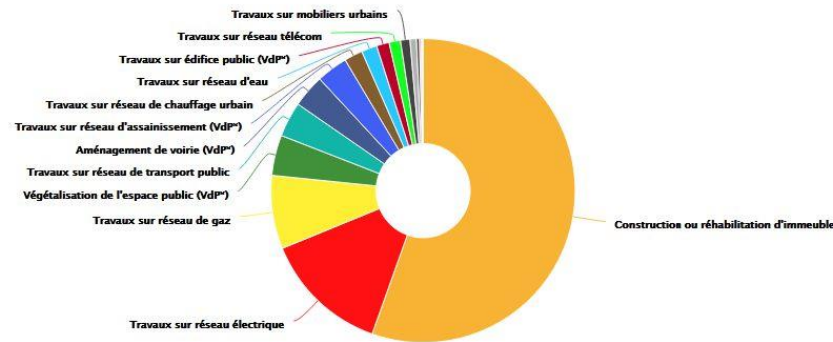
Open Research Data



Open Data, exemple avec Paris

Nombre de chantiers par natures des travaux
* Ville de Paris

<https://opendata.paris.fr/pages/home>



Accédez aux données

Aller plus loin

Thèmes utilisés

Mobilité et Espace Public

Utilisé par 68 jeux de données

Administration et Finances Publiques

Utilisé par 60 jeux de données

Citoyenneté

Utilisé par 45 jeux de données

Equipements, Services, Social

Utilisé par 37 jeux de données

Urbanisme et Logements

Utilisé par 36 jeux de données

Jeux de données les plus populaires

Vélib' - Vélos et bornes - Disponibilité temps réel

3 308 966 téléchargements

Un verger dans mon école

122 167 téléchargements

Belib' - Prises de recharge pour véhicules électriques - Disponibilité temps réel

116 810 téléchargements

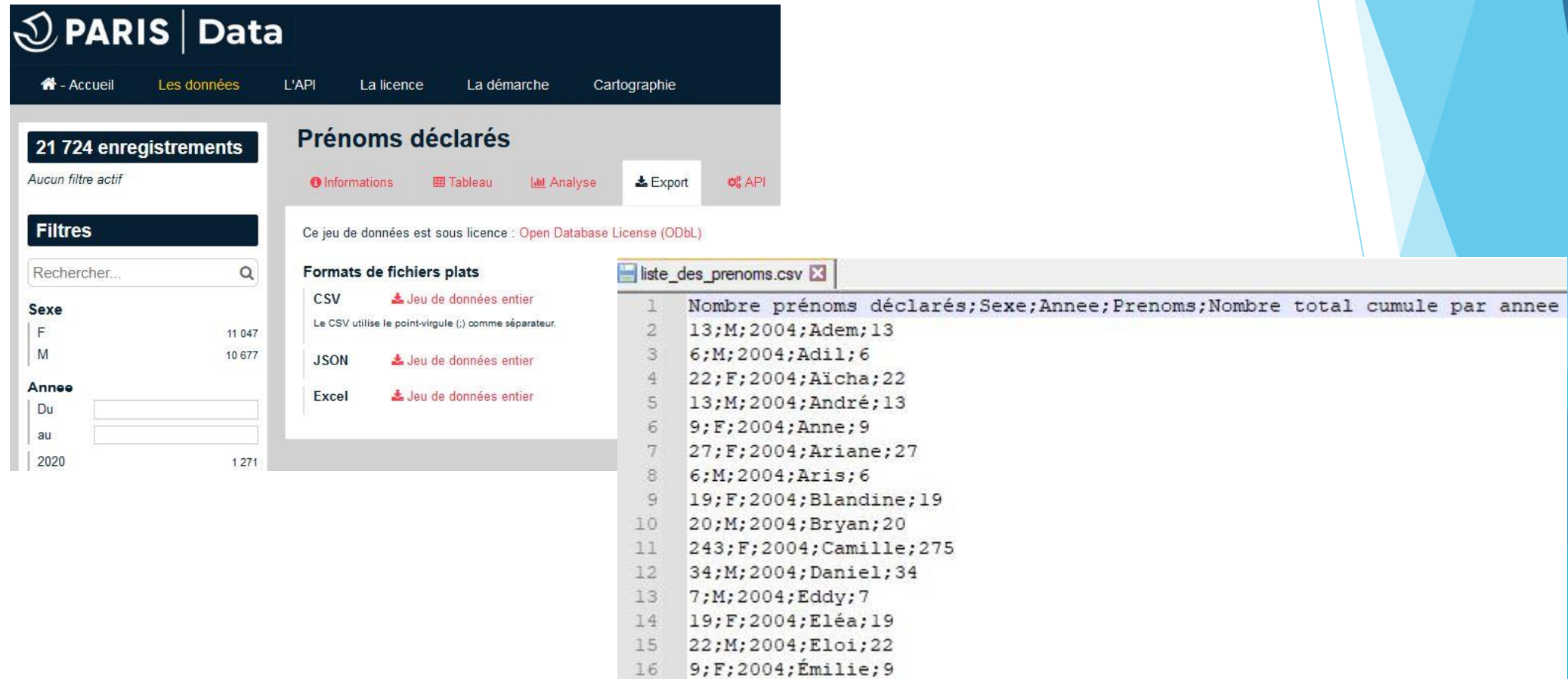
Parcs de stationnement concédés

79 981 téléchargements

Prénoms déclarés

44 535 téléchargements

Open Data, exemple avec Paris



PARIS | Data

Accueil Les données L'API La licence La démarche Cartographie

21 724 enregistrements
Aucun filtre actif

Filtres

Rechercher...

Sexe

- F 11 047
- M 10 677

Annee

Du [] au []

2020 1 271

Prénoms déclarés

Informations Tableau Analyse Export API

Ce jeu de données est sous licence : [Open Database License \(ODbL\)](#)

Formats de fichiers plats

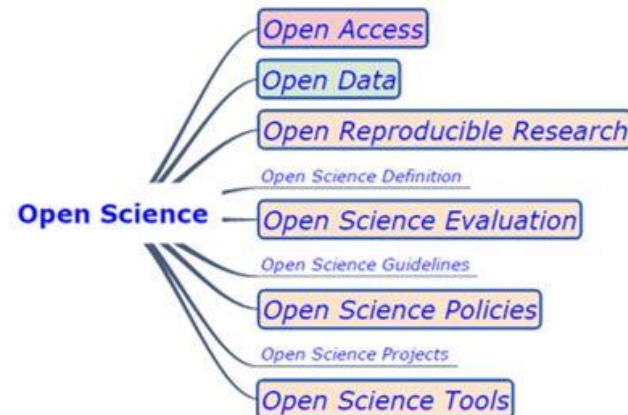
- CSV [Jeu de données entier](#)
Le CSV utilise le point-virgule (;) comme séparateur.
- JSON [Jeu de données entier](#)
- Excel [Jeu de données entier](#)

liste_des_prenoms.csv

	Nombre prénoms déclarés	Sexe	Annee	Prenoms	Nombre total cumule par annee
1	13	M	2004	Adem	13
2	6	M	2004	Adil	6
3	22	F	2004	Aïcha	22
4	13	M	2004	André	13
5	9	F	2004	Anne	9
6	27	F	2004	Ariane	27
7	6	M	2004	Aris	6
8	19	F	2004	Blandine	19
9	20	M	2004	Bryan	20
10	243	F	2004	Camille	275
11	34	M	2004	Daniel	34
12	7	M	2004	Eddy	7
13	19	F	2004	Eléa	19
14	22	M	2004	Eloi	22
15	9	F	2004	Émilie	9
16					

Vers la Science ouverte (Open Science)

Mouvement pour rendre la recherche scientifique, les données accessibles à tous les niveaux de la société (traduit de FOSTER)



D'après « The taxonomy tree » <https://www.fosteropenscience.eu/foster-taxonomy/open-science>

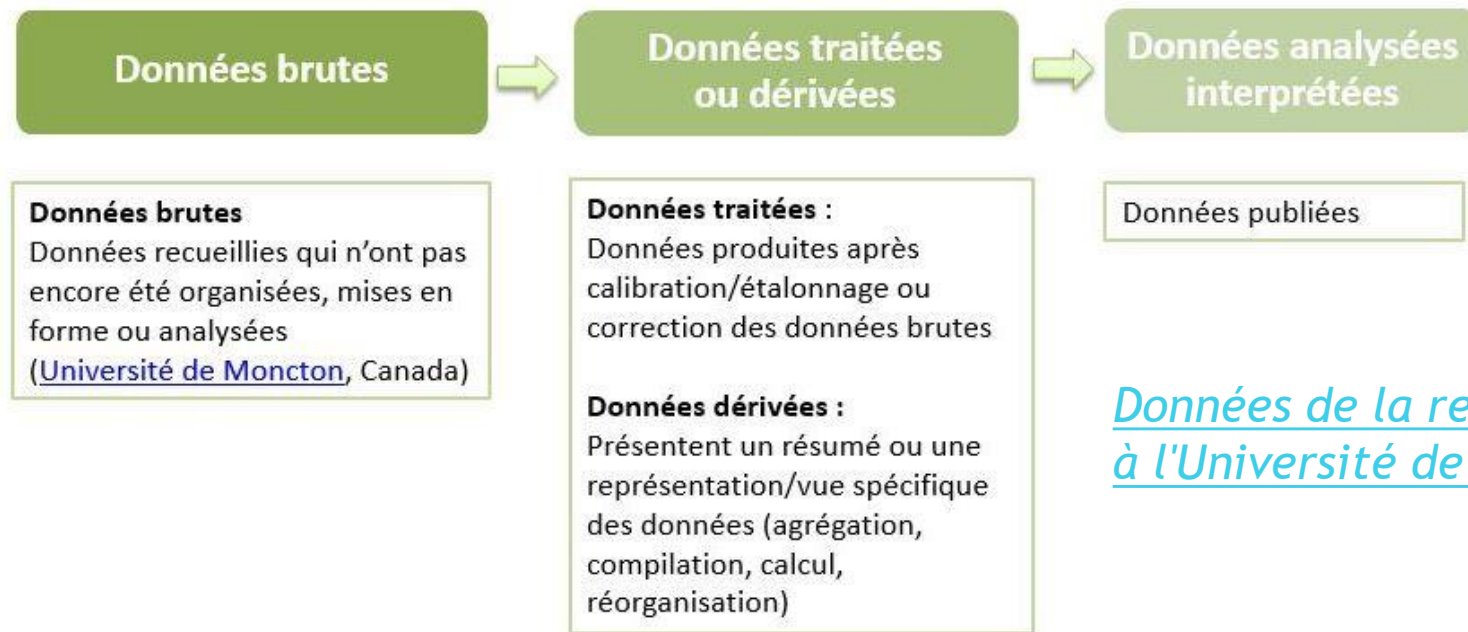
« Qu'est-ce que la Science ouverte ?

L'Open Science est une nouvelle approche transversale de l'accès au travail scientifique, des visées et du partage des résultats de la science mais aussi une nouvelle façon de FAIRE de la science, en ouvrant les processus, les codes et les méthodes. »

DIST-CNRS (2016). Livre blanc – Une Science ouverte dans une République numérique

Données de la Recherche

« Les données de la recherche sont l'ensemble des informations et matériaux produits et reçus par des équipes de recherche et des chercheurs. Elles sont collectées et documentées à des fins de recherche scientifique. A ce titre, elles constituent une partie des archives de la recherche. »



[Données de la recherche à l'Université de Lorraine](#)

Types de données de la 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*

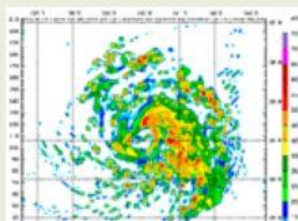


[Pixabay](#), CC0

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



Dataset ou
Jeu de données :
numérique, texte,
son, image.

Data papers et Data journals

Un data paper est une publication dont le but est de décrire un ou plusieurs jeux de données scientifiques, notamment à l'aide d'informations précises appelées métadonnées qui doivent respecter formats et standards parfois disciplinaires. Ces données doivent pouvoir être réutilisées notamment à des fins scientifiques comme la reproductibilité d'expérience.

L'accès en ligne aux données est expliqué dans le data paper (souvent via un DOI). Les jeux de données doivent être déposés dans un entrepôt reconnu parfois imposé par l'éditeur et posséder une licence type CC.

[Webinaire "Data Papers : Quand ? Comment ? Pourquoi ?" du 05/07/2022
par le GTSO Données de Couperin](#)

Data papers et Data journals

Les data papers sont publiés dans deux types de revues :

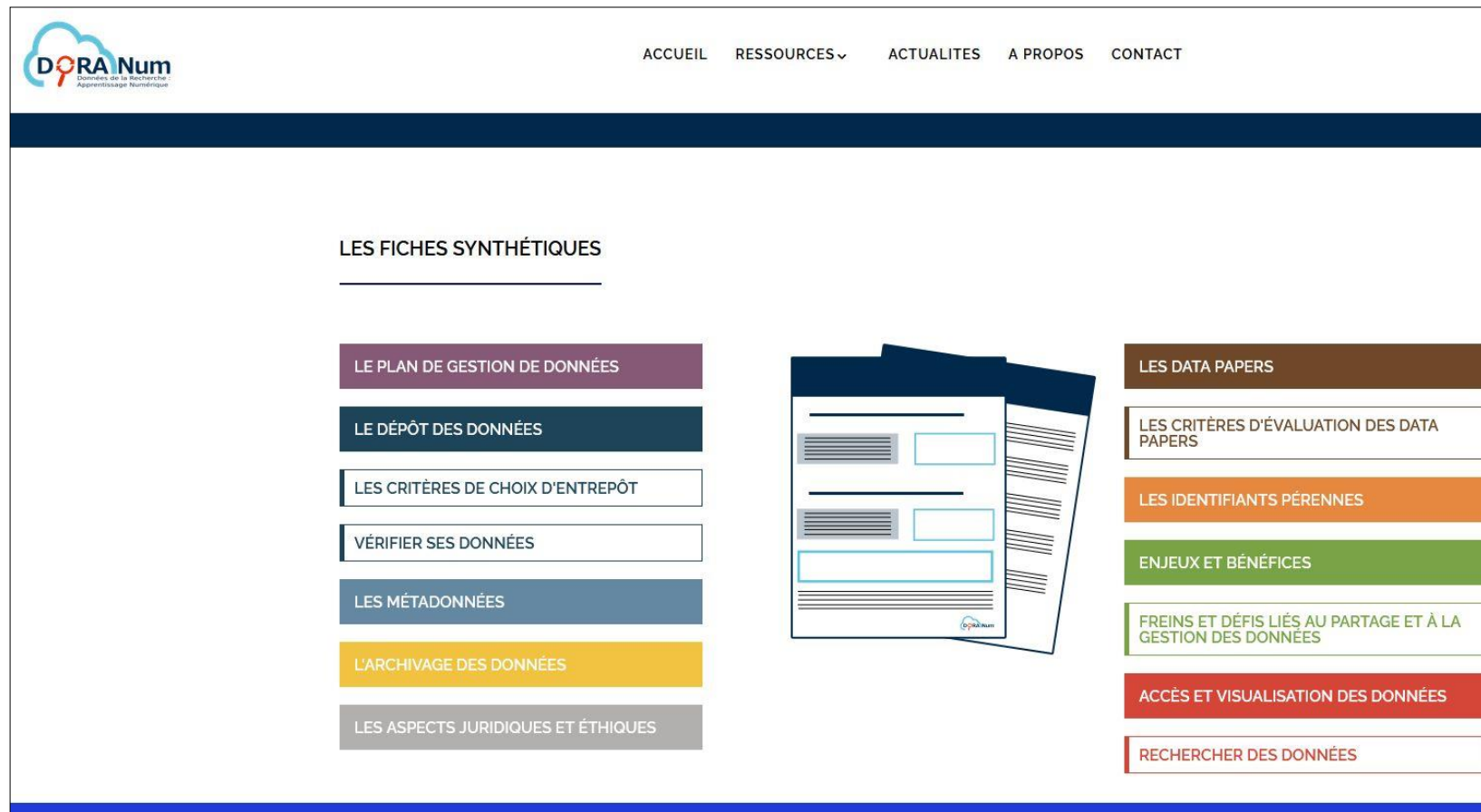
- Revues classiques (Elsevier, Springer,...)
- Data journals : ne contiennent que des data papers (ex: [Ubiquity press](#) avec les « research data »,...)

Certains journaux, notamment en SHS, peuvent accepter des articles « hybrides » où la description des données est accompagnée de résultats, analyses et discussions. Ce qui n'est généralement pas le cas car un data paper n'est pas vu comme un article scientifique classique. Et ses métadonnées peuvent utiliser un type de document spécifique.

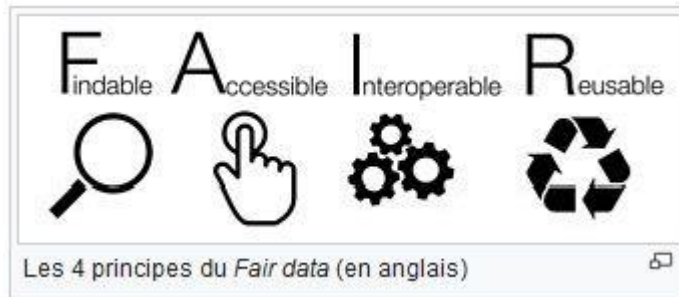
[Publier un Data paper \(CooplST\)](#)

DoRANum

DoRANum : Apprentissage Numérique à la gestion et au partage des données de la recherche



Données FAIR (FAIR data) [Principes FAIR sur go-fair.org](https://go-fair.org)



Contexte Science Ouverte et ouverture des données.

FAIR : Facile à trouver, Accessible, Interopérable et Réutilisable.

[Article fondateur en 2016 dans Nature](#)



GO FAIR FAIR Principles Implementation Networks News Events Resources About GO FAIR

FAIR Principles

Home > FAIR Principles

- > **FAIR Principles**
 - > **F1: (Meta) data are assigned globally unique and persistent identifiers**
 - > **F2: Data are described with rich metadata**
 - > **F3: Metadata clearly and explicitly include the identifier of the data they describe**
 - > **F4: (Meta)data are registered or indexed in a searchable resource**
 - > **A1: (Meta)data are**

In 2016, the **'FAIR Guiding Principles for scientific data management and stewardship'** were published in *Scientific Data*. The authors intended to provide guidelines to improve the **Findability**, **Accessibility**, **Interoperability**, and **Reuse** of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

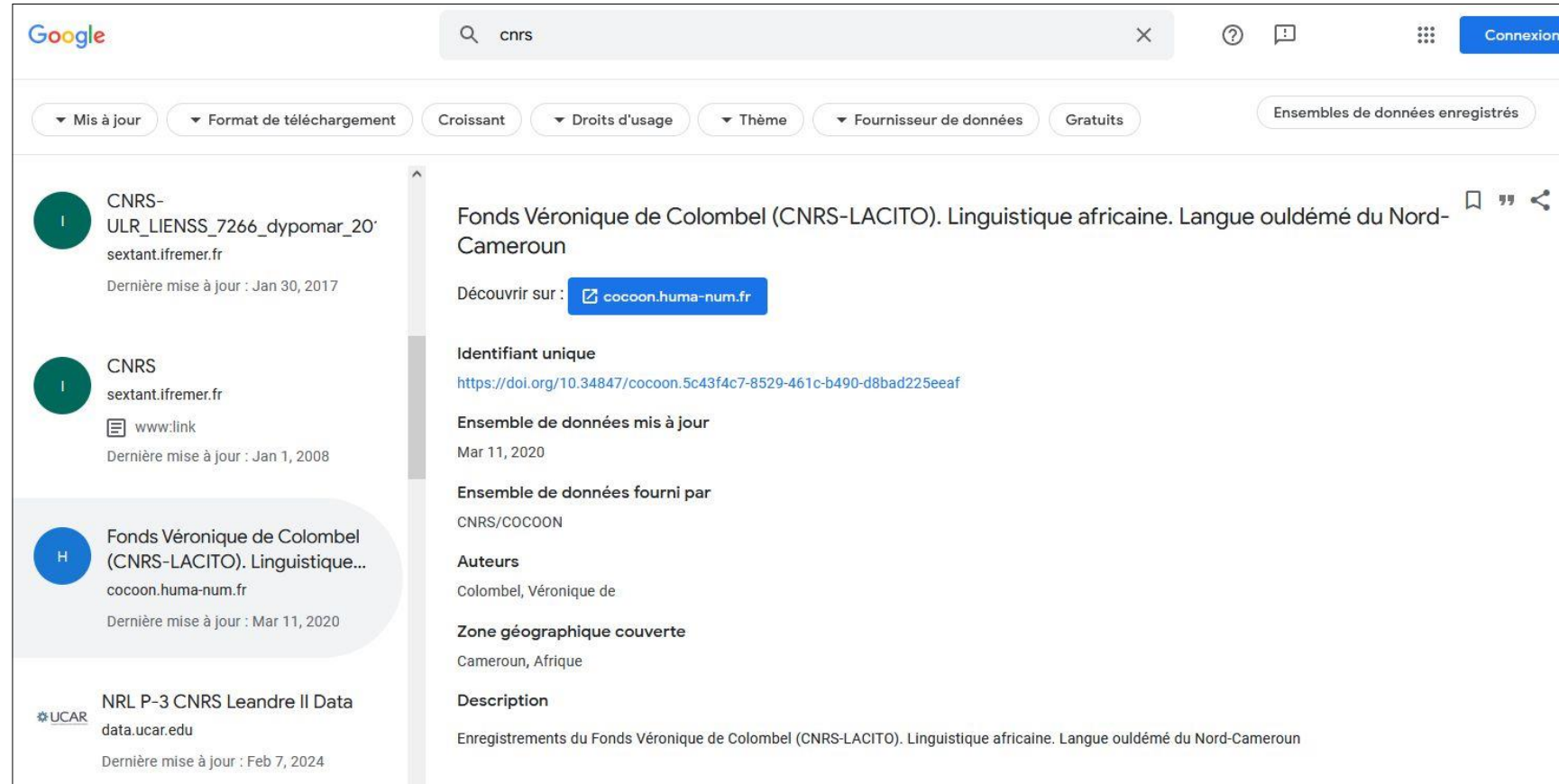
A practical "how to" guidance to go FAIR can be found in the **Three-point FAIRification Framework**.

Findable
The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the **FAIRification process**.

F1. (Meta)data are assigned a globally unique and persistent identifier

<https://datasetsearch.research.google.com/>

Google : Dataset Search



The screenshot shows the Google Dataset Search interface. The search bar at the top contains the text 'cnrs'. Below the search bar, there are several filters: 'Mis à jour', 'Format de téléchargement', 'Croissant', 'Droits d'usage', 'Thème', 'Fournisseur de données', 'Gratuits', and 'Ensembles de données enregistrés'. The search results are displayed in a list on the left and a detailed view on the right.

Search Results:

- 1** CNRS-ULR_LIENSS_7266_dypomar_20' sextant.ifremer.fr
Dernière mise à jour : Jan 30, 2017
- 1** CNRS sextant.ifremer.fr
www.link
Dernière mise à jour : Jan 1, 2008
- H** Fonds Véronique de Colombel (CNRS-LACITO). Linguistique... cocoon.huma-num.fr
Dernière mise à jour : Mar 11, 2020
- UCAR** NRL P-3 CNRS Leandre II Data data.ucar.edu
Dernière mise à jour : Feb 7, 2024

Dataset Details (Fonds Véronique de Colombel (CNRS-LACITO). Linguistique africaine. Langue ouldémé du Nord-Cameroun):

- Découvrir sur :** cocoon.huma-num.fr
- Identifiant unique:** <https://doi.org/10.34847/cocoon.5c43f4c7-8529-461c-b490-d8bad225eeaf>
- Ensemble de données mis à jour:** Mar 11, 2020
- Ensemble de données fourni par:** CNRS/COCOON
- Auteurs:** Colombel, Véronique de
- Zone géographique couverte:** Cameroun, Afrique
- Description:** Enregistrements du Fonds Véronique de Colombel (CNRS-LACITO). Linguistique africaine. Langue ouldémé du Nord-Cameroun



Article exécutable

*Publication avec article, illustrations multimédia, données de recherche, codes informatiques et graphiques... le tout interactif.
Parfois anciennement appelé document computationnel.*



- Très utile à la recherche reproductible
- Regroupe en un document toutes les informations utiles
- Codes informatiques utilisables avec d'autres données (sur une copie du document original)
- Généralement en Jupyter notebook avec les langages Python et R (mais pas seulement)

[Un exemple de revue : Journal of digital history](#)

Article exécutable

Narrative Text

Notebook title and introduction

Description of model parameters

Description of need to profile data

Code and Visualizations

Importing external packages

Implementation of parameters

Profile plotting code

Inline plot

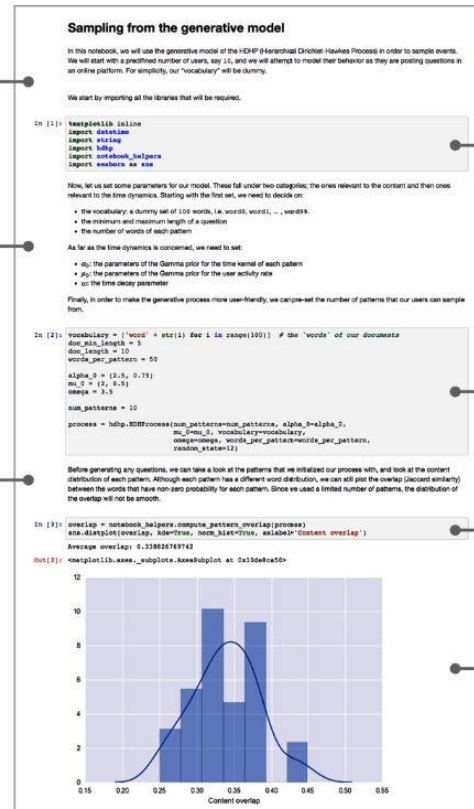


Figure 1: The first half of a computational notebook analyzed in our second study, which demonstrates a novel Python package for modeling patterns of online learning activity. The notebook combines code, visualizations, and text into a computational narrative

Article exécutable

[Why Jupyter is data scientists' computational notebook of choice?](#)

Sampling from the generative model

In this notebook, we will use the generative model of the HDHP (Hierarchical Dirichlet-Hawkes Process) in order to sample events. We will start with a predefined number of users, say 10, and we will attempt to model their behavior as they are posting questions in an online platform. For simplicity, our "vocabulary" will be dummy.

We start by importing all the libraries that will be required.

```
In [1]: %matplotlib inline
import datetime
import string
import hdhp
import notebook_helpers
import seaborn as sns
```

Now, let us set some parameters for our model. These fall under two categories; the ones relevant to the content and then ones relevant to the time dynamics. Starting with the first set, we need to decide on:

- the vocabulary: a dummy set of 100 words, i.e. word0, word1, ..., word99.
- the minimum and maximum length of a question
- the number of words of each pattern

As far as the time dynamics is concerned, we need to set:

- α_0 : the parameters of the Gamma prior for the time kernel of each pattern
- μ_0 : the parameters of the Gamma prior for the user activity rate
- ω : the time decay parameter

Finally, in order to make the generative process more user-friendly, we can pre-set the number of patterns that our users can sample from.

```
In [2]: vocabulary = ['word' + str(i) for i in range(100)] # the `words` of our documents
doc_min_length = 5
doc_length = 10
words_per_pattern = 50

alpha_0 = (2.5, 0.75)
mu_0 = (2, 0.5)
```



Software Heritage

C'est une archive ouverte pour les codes sources des logiciels, développée par INRIA en 2015, et vraiment ouvertes en 2018 suite à un accord avec l'UNESCO.

La mission de Software Heritage est de collecter, préserver et partager tous les logiciels disponibles publiquement sous forme de code source, dans le but de construire une infrastructure commune et partagée.

Les codes sources sont en particulier récupérés à partir de forges logicielles comme GitHub ou GitLab mais aussi à partir de HAL.

Software Heritage


Software Heritage
Archive

Search archived software

☒ only show origins visited at least once
 ☒ filter out origins with no archived content
 ☐ search in metadata (instead of URL)

visit type: any

Origin type	Origin url	Archiving status
git	https://github.com/Inist-CNRS/lodex-processing	✓ Archived
git	https://github.com/Inist-CNRS/lodex-v2	✓ Archived
git	https://github.com/Inist-CNRS/lodex	✓ Archived
git	https://github.com/Inist-CNRS/lodex-extended	✓ Archived
git	https://github.com/Inist-CNRS/lodex-styleguide	✓ Archived
git	https://github.com/Inist-CNRS/lodex-themes	✓ Archived
git	https://github.com/Inist-CNRS/lodex-dumps	✓ Archived
git	https://github.com/Inist-CNRS/lodex-widgets	✓ Archived
git	https://github.com/Inist-CNRS/lodex-workers	✓ Archived
git	https://github.com/Inist-CNRS/lodex-workers-pytorch	✓ Archived
git	https://github.com/Inist-CNRS/lodex-doc	✓ Archived
git	https://github.com/Inist-CNRS/lodex-webservice	✓ Archived
git	https://github.com/Inist-CNRS/lodex-v1	✓ Archived

Features

Search

Downloads

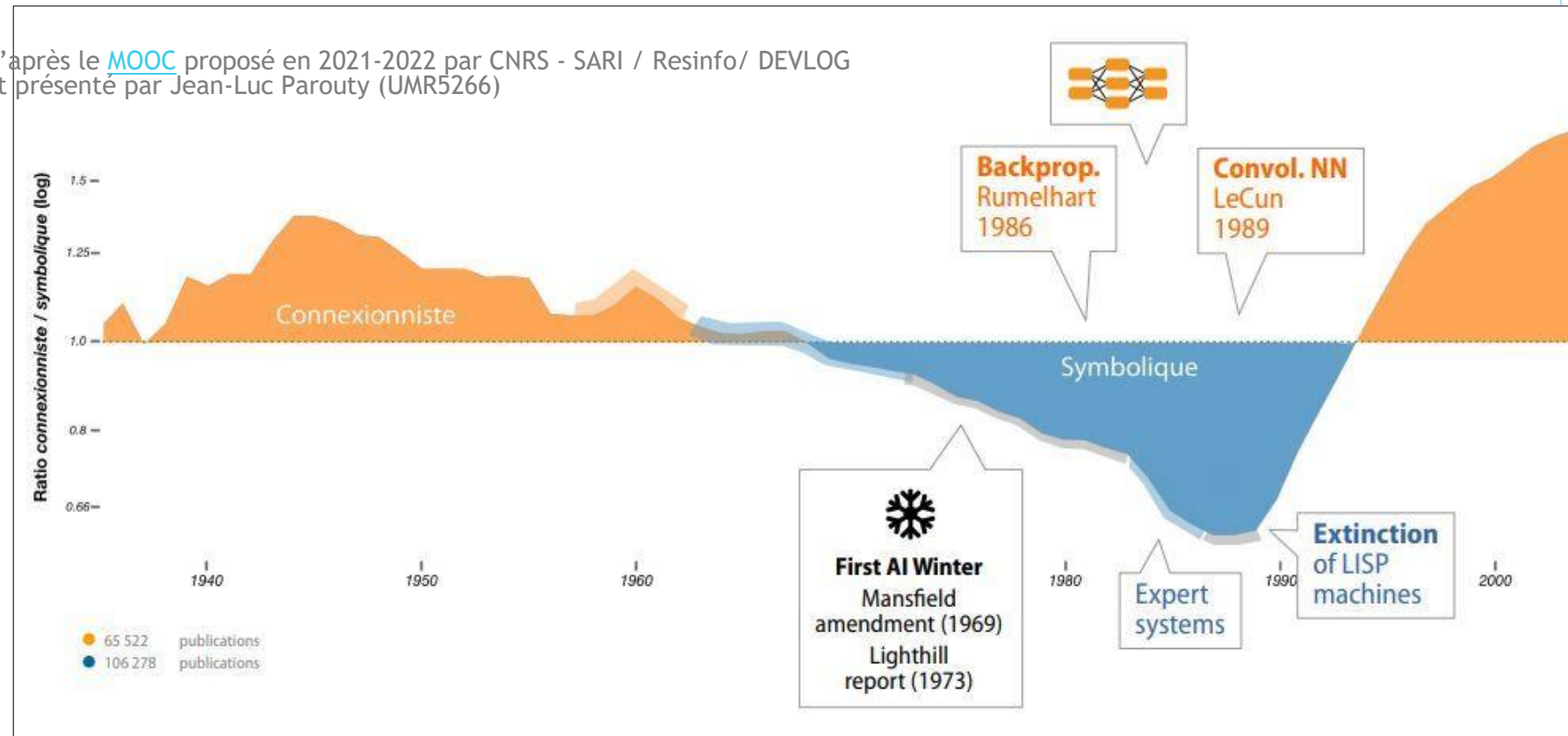
Save code now

Add forge now

Help

IA : Historique et réseaux de neurones

D'après le [MOOC](#) proposé en 2021-2022 par CNRS - SARI / Resinfo/ DEVLOG
et présenté par Jean-Luc Parouty (UMR5266)

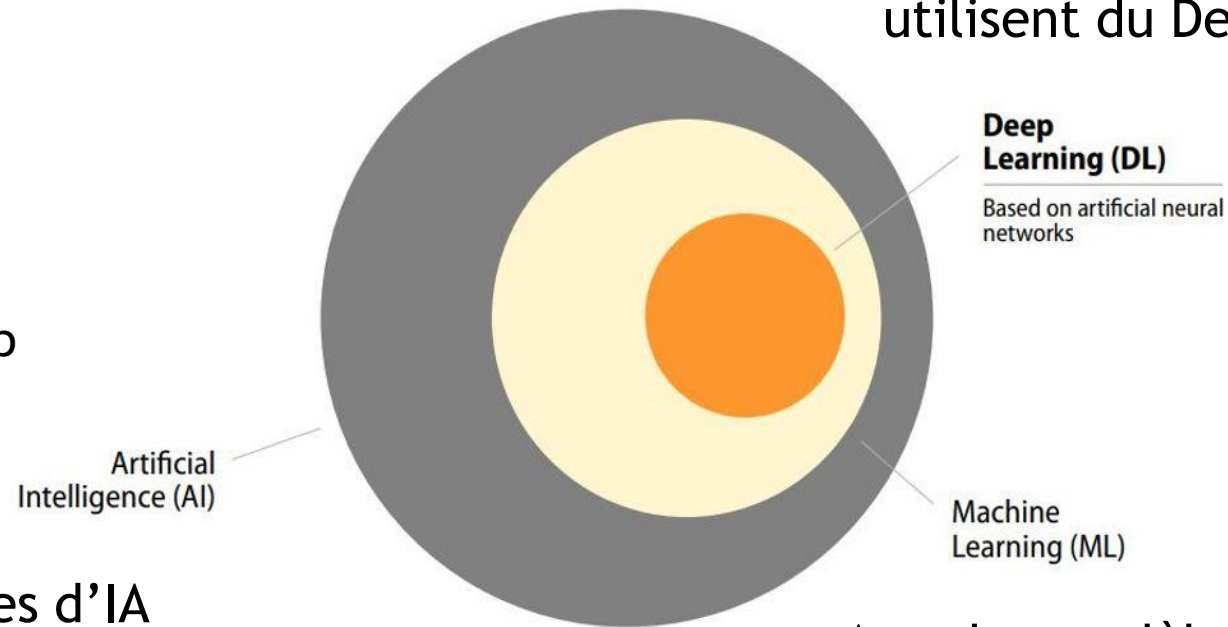


IA : Différents types d'IA

IA symbolique :
moteurs de règles
(systèmes experts).

IA connexionniste :
réseaux de neurones
(machine learning, deep
learning).

Il existe d'autres types d'IA
comme celles basées sur des
algorithmes génétiques.



Les véhicules autonomes
comme les IA génératives
utilisent du Deep learning.

Avec les modèles entraînés en
ML/DL, on réalise des
classifications et/ou prédictions.

IA : IA générative (Chat GPT, Gemini,...)

Il s'agit de Machine learning et de Deep learning. Bien entraînée, une IA générative peut aider à de nombreuses choses :

- comprendre/recommander des informations
- créer des contenus (textes, images, sons/musiques, codes logiciels)

Parmi les limites les plus classiques, on peut noter :

- sources utilisées généralement non citées (mais ça change)
- entraînement insuffisant ou biaisé (et donc erreurs dans les réponses, mais ça change aussi)

Exemple d'IA générative biaisée pour l'exemple : monadGPT. Son entraînement a été fait sur un corpus du XVIIème siècle uniquement !

[MonadGPT \(utilisable en Français\)](#)

[Conversation philosophique avec MonadGPT](#)

IA : Services de traitements de corpus documentaires

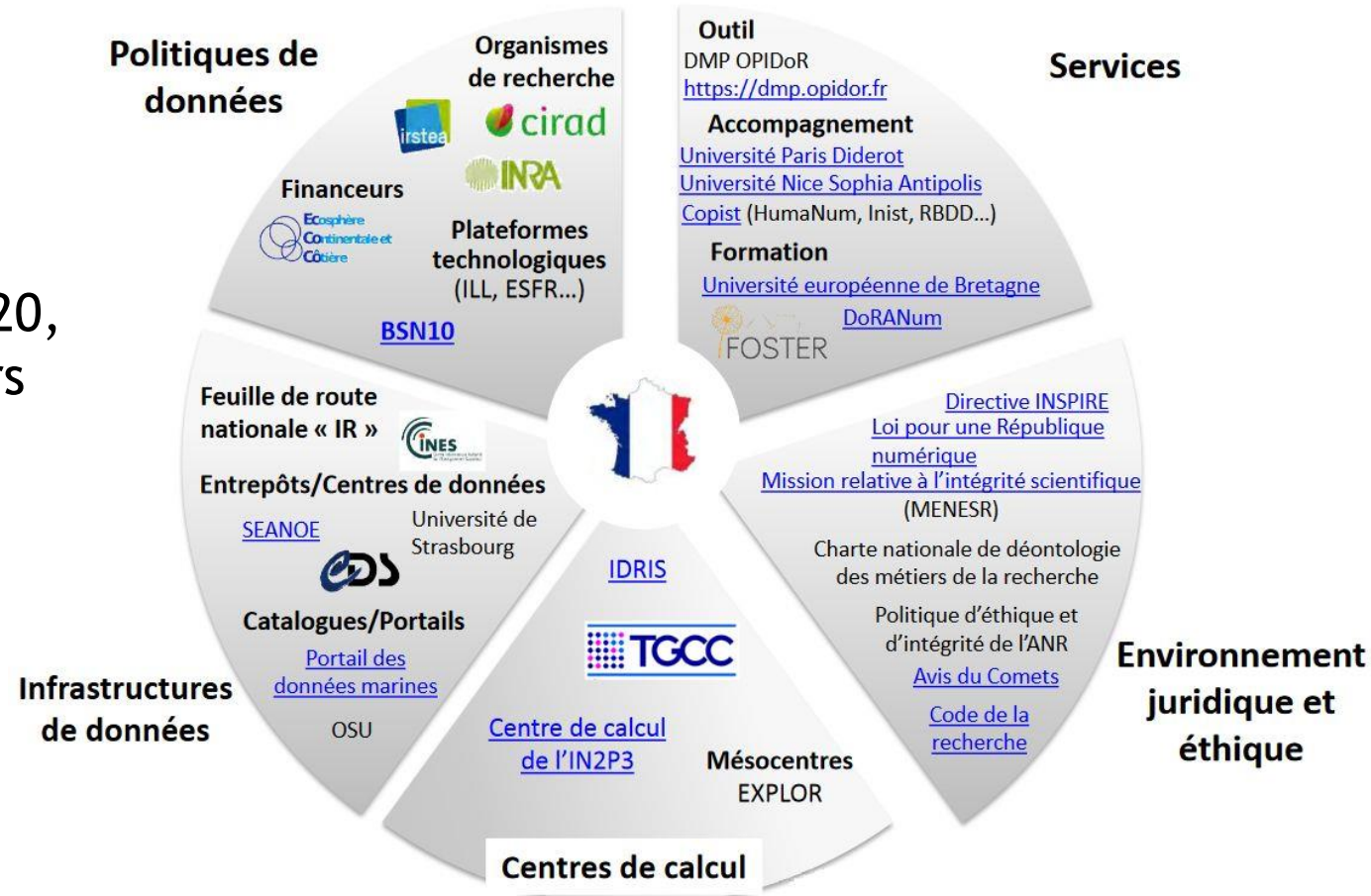


14 services pour exploiter des publications scientifiques via l'IA (2023)

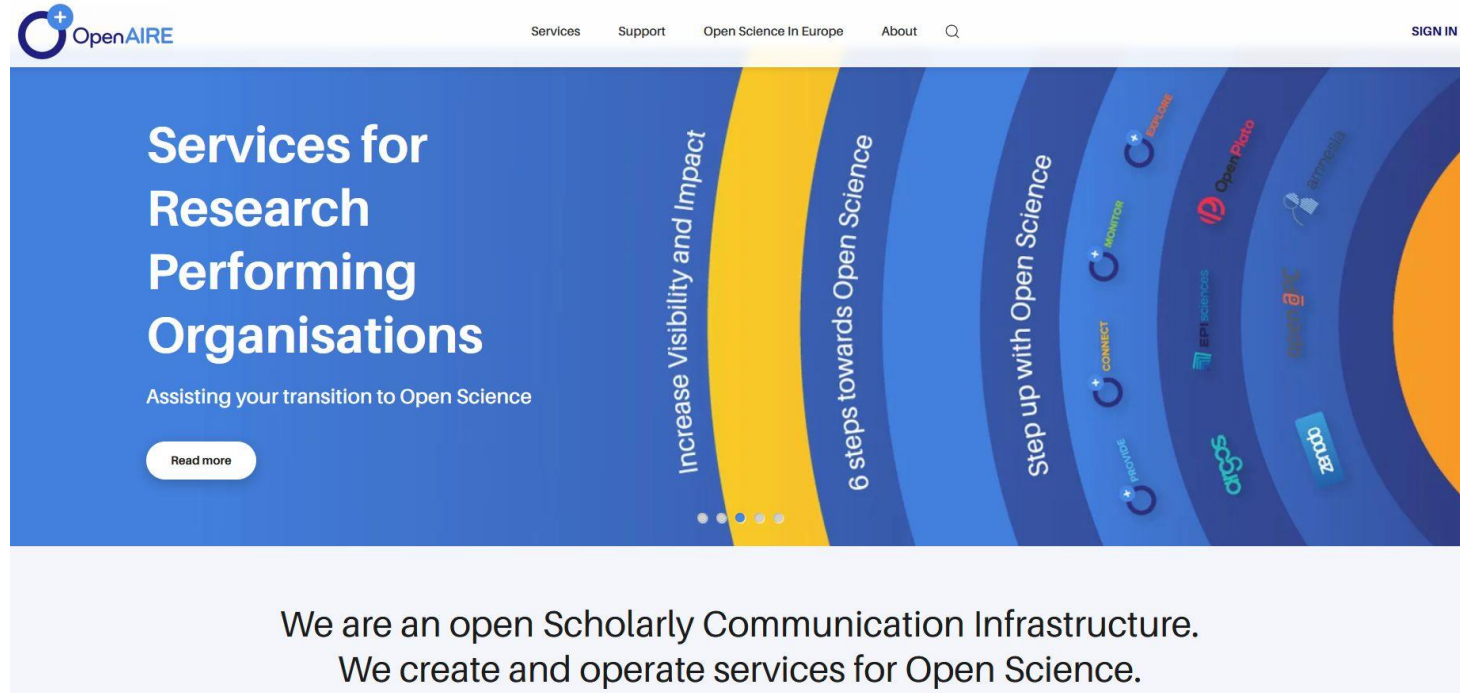
Comparatif de 50 services de traitement de corpus documentaires via l'IA (2023)

Données de la recherche : paysage national

Graphique de 2020, illustrant toujours notre paysage national dans ses grandes lignes.

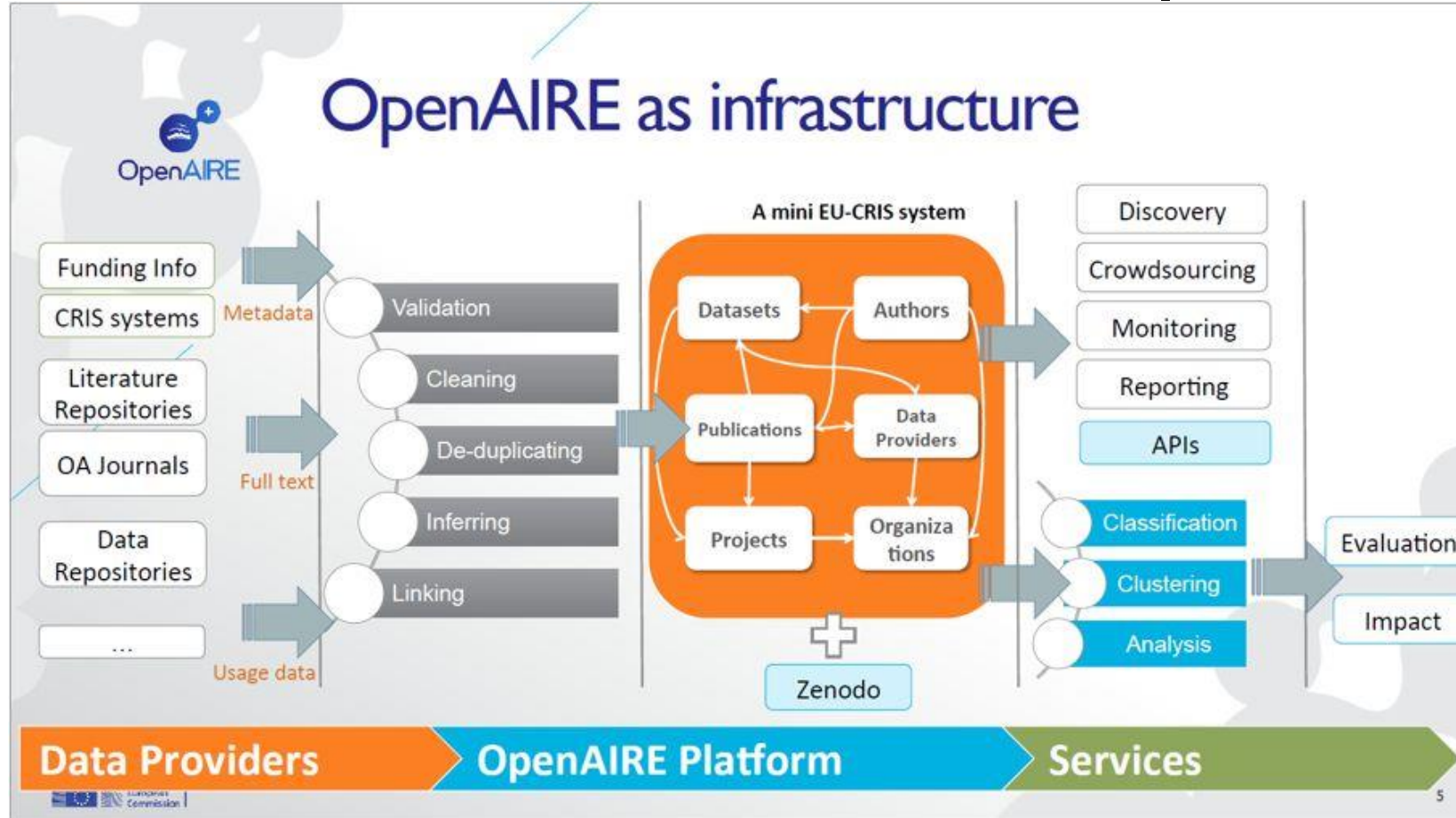


Données de la recherche : OpenAIRE



Open Access Infrastructure for Research in Europe : infrastructure d'accès ouvert, robuste, durable et participative, responsable de la gestion, de l'analyse, de la manipulation, de la fourniture et (surtout) de la mise en réseau d'un très large éventail de publications scientifiques et de données de la recherche, le tout au niveau européen.

Données de la recherche : OpenAIRE



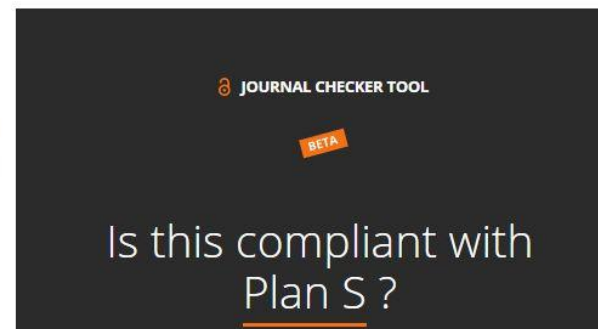
Science Ouverte et Plan S



About Plan S

Plan S is an initiative for Open Access publishing that was launched in September 2018. The plan is supported by cOAlition S, an international consortium of research funding and performing organisations. Plan S requires that, from 2021, scientific publications that result from research funded by public grants must be published in compliant Open Access journals or platforms.

[Read more](#)



- Initiative lancée en 2018.
- L'[ANR](#) (Agence nationale de la recherche) pour la France.
- Pour accélérer la transition vers un accès complet et immédiat aux publications scientifiques.

Plan national pour la Science Ouverte



Second plan national 2021-2024.

- Accès ouvert aux publications
- Ouverture des données de la recherche, y compris les codes logiciels
- Dynamique durable et internationale

<https://www.ouvrirlascience.fr/plan-national-pour-la-science-ouverte/>

EOSC : European Open Science Cloud



- EOSC lancé en novembre 2018 par la Commission européenne.
- Science Ouverte et données FAIR.
- Pour un accès unifié aux infrastructures européennes déjà existantes.