



RÉPUBLIQUE
FRANÇAISE

Liberté
Égalité
Fraternité

recherche.data.gouv.fr



Construction du catalogue des données de Recherche Data Gouv

L'apport des identifiants pérennes

Richard Moreno – INRAE
Responsable Plateforme des données de Recherche Data Gouv

25/11/2025





RÉPUBLIQUE
FRANÇAISE

*Liberté
Égalité
Fraternité*

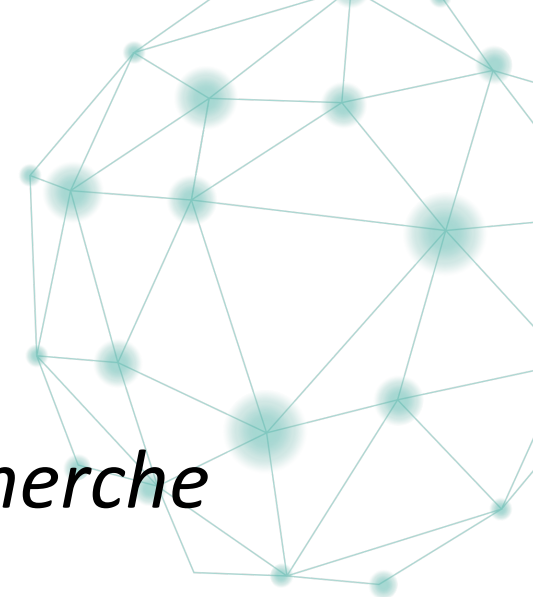
recherche.data.gouv.fr

Recherche Data Gouv



Recherche Data Gouv - un écosystème

*construit par et pour les établissements de l'ESR
pour accompagner
la gestion, le partage et l'ouverture des données de recherche*



- ## Objectifs
- Ne pas laisser d'équipes de recherche sans solution pour ouvrir ou partager leurs données
 - Faire de l'accompagnement un élément central du dispositif
 - Reconnaître, fédérer et amplifier les initiatives existantes

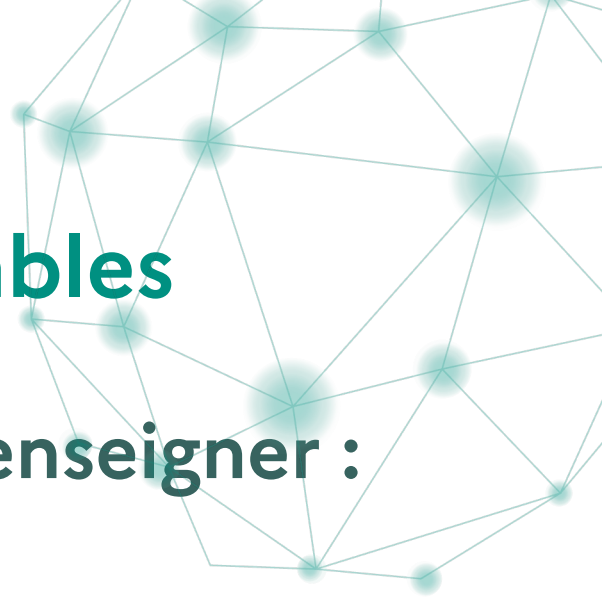


Concevoir nativement des données réutilisables

Travail scientifique de FAIRisation des données, renseigner :

- Les origines et les transformations des données
- Les guides de codification
- Les modèles et la description détaillée des méthodes de collecte, de nettoyage, d'analyse
- Le code spécifique à un instrument
- Les logiciels conçus pour des applications propres au domaine

Plus la réutilisation est loin du point d'origine, en termes de temps, de problématique de recherche, de discipline... plus il devient difficile de retrouver et d'interpréter des données





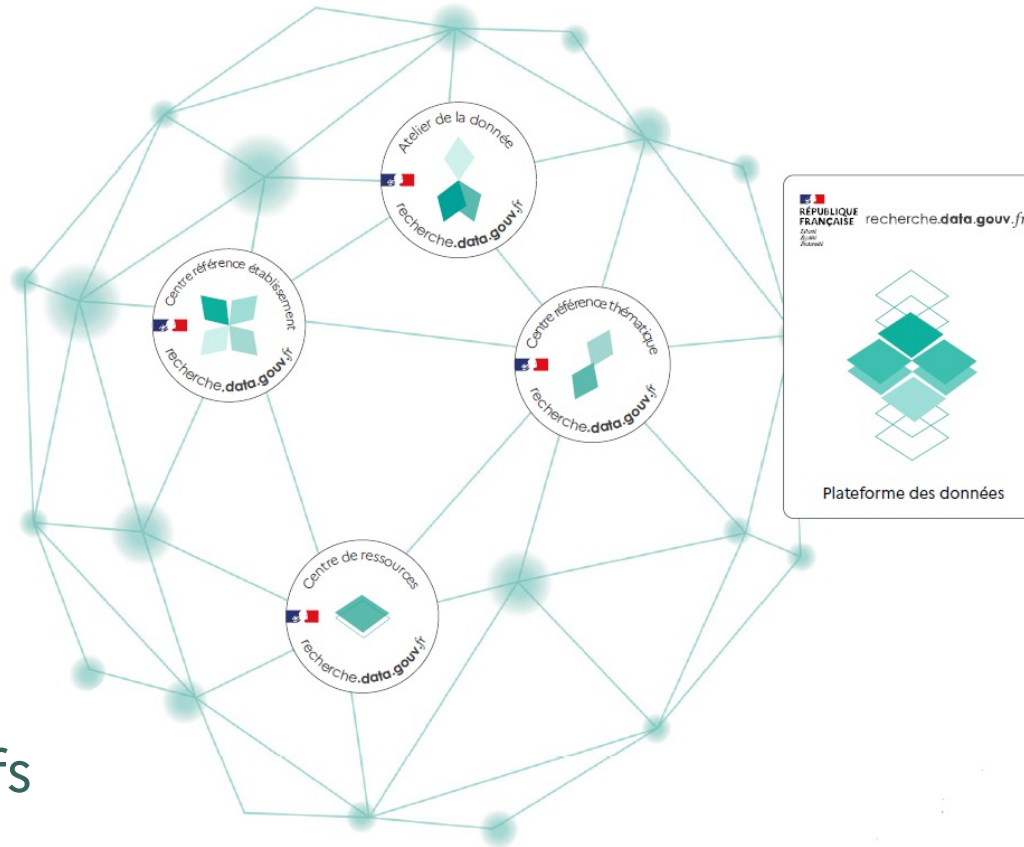
RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.fr

Concrètement, Recherche Data Gouv, c'est :

Une fédération nationale de centres de compétences

Maillage de dispositifs d'accompagnement sur la gestion et la diffusion des données de recherche, au plus près des équipes de recherche



Une plateforme nationale des données de recherche

Solution souveraine pour déposer, publier, signaler, découvrir, rendre accessible et réutilisable les données de recherche grâce à :

- un entrepôt national de données de recherche ;
- un catalogue des données publiées sur l'entrepôt lui-même ou sur des entrepôts externes.

Plateforme nationale des données de la recherche

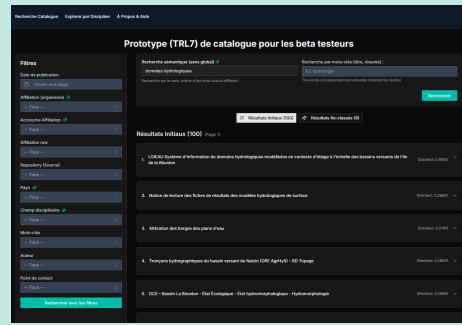
Unité mixte de services « Plateforme » depuis le 01/01/2025



Portail



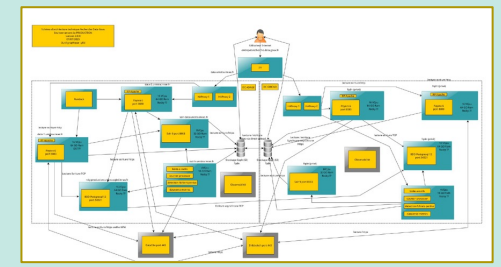
Entrepôt



Catalogue



Centre de ressources
Plateforme CreP



Infrastructure matérielle



40 personnes - 22 ETP
 5 tutelles + CNRS/INIST + U. Paris Nanterre, Rennes
 Soutien régulier DoRANum et GIS URFIST



RÉPUBLIQUE
FRANÇAISE

*Liberté
Égalité
Fraternité*

recherche.data.gouv.fr

Introduction au catalogue des données de Recherche Data Gouv



Le catalogue « augmenté »

La situation de départ : les limites du mot clé exact et des silos de données

- Recherche en **silos** : L'utilisateur doit interroger chaque entrepôt séparément
- **Dépendance aux mots-clés exacts**

Objectifs du catalogue

- **Accès unifié** : un seul point d'entrée pour tout explorer
- **Compréhension du sens** : le moteur comprend l'intention > orientation découverte des données
- **Découverte contextuelle** : le système suggère des publications, codes et projets liés
- **Polyglotte** : le système est insensible à la langue utilisée
- **Sérendipité et efficacité** : découvrir des ressources inattendues

Recherche par filtres

Une méthode classique pour trouver exactement ce qui est recherché grâce à des critères comme l'auteur, la date ou l'institution.

→ Idéale si on sait déjà ce que l'on veut

Recherche intelligente par le sens

Une méthode guidée par l'IA qui comprend le sens de la question, quelle que soit la langue utilisée

Exemple : en tapant "données sur la sécheresse en milieu agricole", on peut tomber sur "étude du stress hydrique des cultures"

→ Explorer un sujet sans chercher des mots précis

Combinaison des deux modalités de recherche = recherche hybride

- Dépasser les limites : la recherche comprend différents langages et domaines de recherche
- Obtenir une réponse : même si on ne connaît pas les bons mots
- Découvrir plus : le système propose des résultats liés (publications, codes, données).
- Explorer à la place de rechercher : la recherche devient intuitive et interactive



RÉPUBLIQUE
FRANÇAISE

*Liberté
Égalité
Fraternité*

recherche.data.gouv.fr

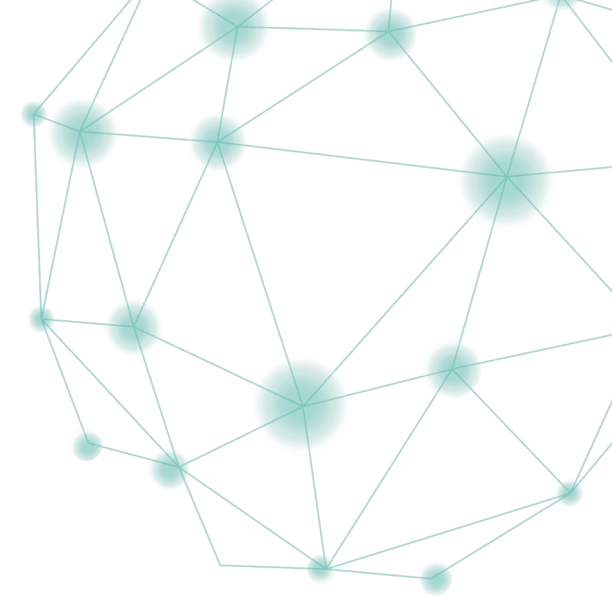
Introduction au catalogue des données de Recherche Data Gouv





Catalogue des données de la recherche française

- Données de la recherche
 - Produite par une activité de recherche
- Données de la recherche française
 - Produites ou coproduites par au moins un auteur affilié à un établissement de recherche ou d'enseignement français ;
 - ou dans des entrepôts de données gérés par des établissements français de la recherche, sans faire de distinction sur l'affiliation des auteurs des données (français ou étrangers)
- Données hébergées dans
 - Potentiellement tout type d'entrepôt dans la mesure où elles sont signalées manuellement
 - —> Uniquement les entrepôts priorisés par les acteurs de Recherche Data Gouv concernant le référencement automatisé





Pourquoi et pour qui un catalogue?

Dans un contexte de multiplicité des entrepôts de données (thématiques, généralistes, institutionnels)

- **Enjeux pour les scientifiques**

- Découvrir des entrepôts
- Trouver des données d'intérêt
- S'adapter à l'accroissement d'information (méta-analyses, deep search assistée par IA, ...)

- **Enjeux pour les établissements de recherche et d'enseignement**

- Avoir une vision globale de la production des données produites
- Mesurer l'impact des données partagées

- **Enjeux au niveau national**

- Avoir une vision globale des données produites par la recherche française
- Mesurer l'impact des données partagées

- **Enjeux au niveau européen et international**

- Proposer un point d'accès unique des données de la recherche française



Grands principes et hypothèses

L'objectif n'est pas de réaliser un méga catalogue de toutes les données ESR

- Uniquement de permettre la découverte de jeux de données
 - **Pour permettre une recherche de type « 2 step search »**
 - 1/ Recherche au niveau des jeux de données, d'une manière assez générique
 - 2/ une fois le jeu de données identifié
 - *L'utilisateur est redirigé via URL ou API sur l'entrepôt ou l'infrastructure de recherche concerné*
 - *Il y trouve tous les outils pour chercher et traiter les données du jeu de données*
 - **Cette « découverte » a principalement pour but de réduire le nombre de jeux**
 - Passer de quelques centaines de milliers à quelques uns – quelques dizaines

Méthode employée par les agences spatiales (CEOS) et par Data Terra



L'objectif n'est pas de réaliser un méga catalogue de toutes les données ESR

- Avantages de **rester au niveau jeu de données**
 - La partie commune à la recherche transversale peut rester simple
 - La complexité, spécificité (modèle de métadonnées, artefacts sémantiques, ...) reste au niveau de la recherche de la donnée elle-même
 - S'il y a lieu : certains jeux de données peuvent être constitués d'un seul fichier ou ne pas proposer de recherche avancée 'via API par exemple)
- On ne **refait pas le travail** des communautés scientifiques, entrepôts, ... précis et avancé sur les données
 - Action de permettre la découverte , le signalement de jeux de données et de mettre en avant les entrepôts, les organismes et projets associés

Aider à la découverte des données



Profiter des nouvelles technologies pour faciliter la découverte des données

- Le catalogue exposera 100000+ jeux de données
 - Utiliser des critères classiques et peu structurés est voué à l'échec
- Besoin de trouver des méthodes pour faciliter la découverte des jeux de données
 - S'appuyer sur des vocabulaires contrôlés avec des nombres d'items raisonnables
 - => **besoin d'aligner sur des vocabulaires contrôlés**
 - Capacité de proposer des recherches sémantiques vectorielles
 - En fait hybrides : mélangeant critères classiques et recherche sémantique vectorielle



Aider à la découverte de ressources

- **Capacité à rebondir d'un jeu de données vers d'autres ressources**

- Découvrir les autres productions d'un créateur (avec le catalogue, avec d'autres catalogues de ressources, via sa page ORCID)
- Découvrir des productions qui sont proches de cette ressource (autres JDD, publications, PGD, code, ...)

- **Capacité à réutiliser un jeu de données**

- Constat : les métadonnées sont nécessaire - mais peuvent ne pas être suffisantes
- Par exemple : besoin de connaître comment elles ont été produites (données utilisées, étalonnage, algorithme de traitement, validation, ...)
- Une façon d'y arriver est de lier les JDD à d'autres ressources (via le DOI par exemple)
- En cours : lien entre HAL et l'entrepôt RDG



Identifiants
pérennes
Recherche
sémantique
Lien avec d'autres
dispositifs

Auteur Affiliation

INRAE (1 865)

INRA - Institut National de la Recherche

Agronomique (399)

INRAE, UR RiverLy, Villeurbanne, France

(212)

UGSF, UMR 8576; CNRS, Université de Lille;

France (58)

Centre national de la recherche scientifique

(CNRS) - CERCEC Centre d'études russes,

caucasiennes, est-européennes et

centrasiatiques (UMR 8083, EHESS, CNRS);

France (54)

CNRS (51)

INRAE, UMR IGE, Grenoble, France (50)

CIRAD (41)

AgroParisTech (38)

INRAE - Institut national de recherche pour

l'agriculture, l'alimentation et l'environnement

(36)

Friedrich Schiller University Jena, Institute of

Geography, Jena, Germany (34)

Ecole des Hautes Etudes en Sciences

Sociales (EHESS) - CERCEC Centre d'études

russes, caucasiennes, est-européennes et

centrasiatiques (UMR 8083, EHESS, CNRS),

Université de Strasbourg - GEO Groupe

d'études orientales, slaves et néo-helléniques

(UR 1340); France (32)

Université Montpellier (32)

Institut national d'études démographiques

(INED) - UR12 "Mobilité, parcours et

territoires", Ecole des Hautes Etudes en

Sciences Sociales (EHESS) - CERCEC

Centre d'études russes, caucasiennes, est-

européennes et centrasiatiques (UMR 8083,

EHESS, CNRS); France (31)

Observatoire de Paris - PSL (30)

iv.fr

Exemples d'API de recherche

Ce dataset est issu de la base de données de l'Institut National de la Recherche Agronomique (INRAE). En cliquant sur le lien, vous accédez directement au dépôt source des données.

Dataveg project (INRAE)



21 nov. 2025 BaGaTeL Dataverse



PLAN DE GESTION DE DONNEES ANR-22-CE21-0011 UNDERPHE



21 nov. 2025 - UMR SPO



Berger, Patrick, 2023, "PLAN DE GESTION DE DONNEES ANR-22-CE21-0011 UNDERPHE", <https://doi.org/10.57745/N7YG7X>, Recherche Data Gouv, V2

Plan de gestion de données du projet UnderPhe ANR-22-CE21-0011

«

< Précédent

1

2

3

4

5

Suivant >

»



Prototype (TRL7) de catalogue pour les beta testeurs

Filtres

Date de publication

Choisir une plage

Affiliation (organisme)

-- Tous --

Affiliation brute

-- Tous --

Repository (Source)

-- Tous --

Pays

-- Tous --

Champ disciplinaire

-- Tous --

Mots-clés

-- Tous --

Auteur

-- Tous --

Point de contact

-- Tous --

Recherche sémantique (sens global)

consommation électrique

Recherche par le sens, même si les mots exacts diffèrent.

Recherche par mots-clés (titre, résumé) :

Ex: hydrologie

Trouve les correspondances textuelles (tolérant les fautes).

Rechercher

Résultats de la recherche (200)

Re-classement avancé

Retour aux résultats paginés

Exporter tout

Résultats Re-classés (10) (Top 10 après analyse sémantique)

1. Données issues de l'étude des comportements de consommation d'électricité (EXPESIGNO)

ENTREPÔT RECHERCHE DATA GOUV

(Score CE: 0.6505)

Le projet EXPESIGNO a pour principal objectif l'étude des comportements de consommation d'électricité de foyers lorsqu'ils font face à des signaux d'effacement et/ou de report. Ces foyers ont participé à une étude expéri...

2. The EVE Pilot: Usage Data from an Electric Car in France

ENTREPÔT RECHERCHE DATA GOUV

(Score CE: 0.6029)

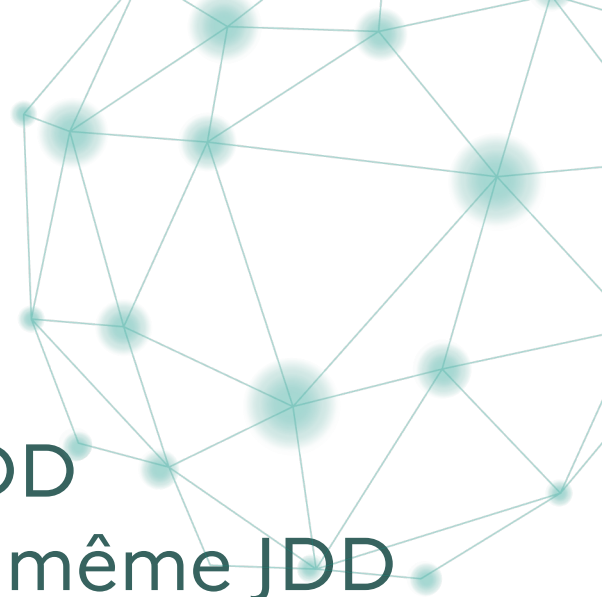
This dataset contains the usage data of a single electric car collected in as part of the EVE study (Enquête des Vehicules Electrique) run by the Observatoire du Transition Énergétique Grenoble (OTE-UGA). This dataset i...

3. Efficiency tests on A123 4.4Ah cells

ENTREPÔT RECHERCHE DATA GOUV

(Score CE: 0.5535)

This dataset contains characterization measurement that were performed at the University Gustave Eiffel and supervised by Eduardo Redondo-Iglesias (eduardo.redondo@univ-eiffel.fr). Four 4.4 Ah A123 cells were tested in a...



• DOI

- Essentiel pour identifier de façon précise un JDD
- Permet notamment de gérer les doublons : un même JDD peut être stocké ou référencés dans plusieurs entrepôts / catalogues
- Permet de relier des ressources à un JDD (ex publication)
- Permet de facilement remonter à l'entrepôt source (via la landing page)
- Tous les JDD n'ont pas de DOI – notamment les données anciennes
- => on doit aussi prendre en compte l'identifiant interne de l'entrepôt



Identifiants pérennes



- **Affiliation (organismes)**

- Utilisation du ROR (acronyme et étendu)
 - L'utilisation du ROR n'est pas généralisée, mais en progression
 - Le ROR c'est ~150000 entrée, donc gérable d'un point de vue informatique pour réaliser des alignements
-
- ROR ne couvre que les établissement de recherche.
 - Or nombre de co-auteurs sont hors de ce cadre
 - Doit être contextualisé : CNRS, AMU si hors contexte « France »



Identifiants pérennes / Vocab contrôlé

- **Disciplines / domaines scientifiques**

- Pour un catalogue multi-thématiques, il est important de pouvoir filtrer sur le domaine scientifique
- Or, c'est une information qu'on ne trouve pas en général dans les métadonnées
 - Sur Dataverse Recherche Data Gouv : mot-clé « thématique » qui est noyé dans les autres mots clés
- Choix actuel : thesaurus HCERES
 - Permet une classification sur un nombre raisonnable de valeurs (~100)
 - Métadonnées créées de façon synthétique via IA





Identifiants pérennes / Vocab contrôlé



- **Personnes (créateurs et points de contact)**
 - Utilisation d'ORCID
 - Landing Page d'ORCID contient potentiellement beaucoup d'information
- ORCID : des dizaines de millions d'entrées
 - **Complicqué** de récupérer l'ensemble, de le garder à jour dans une base locale
 - Complicqué de trouver directement une personne
- Attention humains : utilisation de l'IA réglementée



Identifiants pérennes / Vocab contrôlé

- **Autres Identifiants pérennes / Vocab contrôlé**
 - **Mots-clé**
 - Très grande diversité (plus que nécessaire pour un catalogue généraliste)
 - Certains sur des thesaurus (mais pas pour tous les JDD)
 - Au bilan, plus de mots-clés que de JDD !!
 - Piste : EuroVoc
 - *Multilingue*
 - *21 domaines / 127 micro thesaurus / 7300 concepts*



Identifiants pérennes / Vocab contrôlé

- **Autres Identifiants pérennes / Vocab contrôlé**
 - **Licences**
 - Pour aider à la découverte (limiter le nombre de choix)
 - **Financeurs**
 - Idem



Identifiants pérennes : Synthèse

- Les alignements que nous faisons sont utiles et nécessaires pour les anciennes données
- Mais comme tout traitement, il peut y avoir des erreurs
- Et certains alignements (ORCID) sont compliqués
- La meilleure option (pas d'erreur, moins couteuse) : renseigner les identifiants pérennes dans les métadonnées



Qualité des métadonnées => qualité de découverte et de la réutilisation

Le « chaos » actuel des métadonnées

- **Hétérogénéité** : les descriptions varient selon les chercheurs et les labos
- **Ambiguïté** : un même établissement peut être nommé de plusieurs façons (ex. "UT3", "Univ. Toulouse"...)
- **Manque de structure** : les mots-clés libres rendent la recherche difficile et confuse

Utilisation d'agents IA pour aligner les métadonnées sur des vocab contrôlés.

Voire en ajouter (champs disciplinaires)

Une richesse invisible faute de bonnes descriptions

- **Recherche fastidieuse** : Deviner tous les mots possibles pour trouver ce que l'on cherche...
- **Données cachées** : Des ressources précieuses restent introuvables car mal décrites
- **Perte de potentiel** : même le meilleur moteur ne peut pas deviner ce qui n'est

Utilisation d'agents IA pour analyser la qualité des métadonnées (sur le fond)
Avec un focus future sur identifiant pérennes

<=> analyse curation a posteriori
Puis au moment du dépôt

Faciliter le dépôt de jeux de données

Perspectives ~1 an

Dépôt assisté

- Décrire un jeu de données prend aujourd'hui plus de 30 minutes et est vécu comme une tâche pénible
- Possible de lier la saisie d'un DMP dans OPIDOR => **certaines champs ne sont saisis qu'une seule fois** (ROR, ORCID, ...)
- Au moment du dépôt, plusieurs documents sont disponibles (publications, documents techniques, Readme, PGD, ...) qui peuvent être exploité via IA pour aider à préremplir le formulaire de dépôt
- Le déposant valide et/ou corrige
- Résultat : **Moins de 5 minutes pour finaliser le dépôt, avec des données bien décrites**