

Données, codes, logiciels

Introduction aux enjeux

Claire Tignolet – 19/11/24

claire.tignolet@chartes.psl.eu



I. CONTEXTE

II. DÉFINITIONS

**III. ENJEUX ET OUTILS
SPÉCIFIQUES**

I. CONTEXTE

Open...

Science

Government Partnership

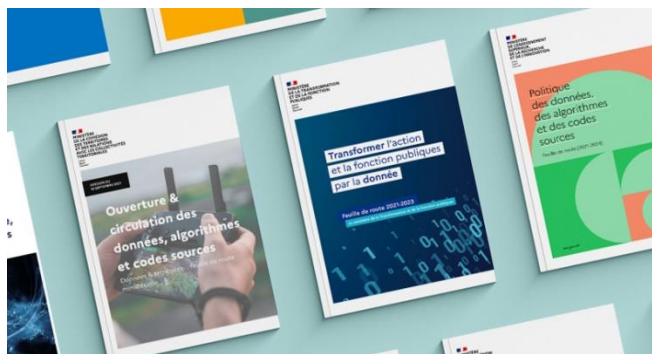


Les composantes de la science ouverte [selon l'Unesco \(2020\)](#) + [Recommandation sur une science ouverte \(2021\)](#)

<https://www.opengovpartnership.org/>

Politique des données, des codes sources et des algorithmes (rapport Bothorel 2020)

- Depuis 2021, déclinée en 15 feuilles de routes
- Gouvernance interministérielle à travers les AMDAC (Administrateurs ministériels des données, algorithmes et codes sources)



- Des actions couvrant tout le cycle de vie des données
- Recensement et valorisation des données de chaque ministère
- Renforcement de l'ouverture des algorithmes et des codes sources
- Promotion du recours aux logiciels libres
- Développement des compétences

Du côté de l'enseignement supérieur et de la recherche



5 objectifs

- Promouvoir, développer et mutualiser des services d'ouverture et d'usage des données
- Participer à la souveraineté et à l'indépendance technologique française et européenne
- Développer la culture et l'usage de la donnée
- Faire des données, des algorithmes et des codes un bien commun au service de tous
- Améliorer la visibilité internationale des données de l'enseignement supérieur, de la recherche et de l'innovation

Les plans nationaux pour la Science Ouverte



I- 2018-2021

1. Généraliser l'accès ouvert aux publications
2. Structurer et ouvrir les données de recherche
3. S'inscrire dans une dynamique durable, européenne et internationale

(bilan)



II- 2021-2024 : Généraliser la science ouverte en France

1. Généraliser l'accès ouvert aux publications
2. Structurer, ouvrir et partager les données de recherche
3. Ouvrir et promouvoir les codes sources produits par la recherche
4. Transformer les pratiques pour faire de la science ouverte le principe par défaut

Axe 2 : structurer, ouvrir et partager les données de recherche

4

Mettre en œuvre l'obligation de diffusion des données de recherche financées sur fonds publics

5

Créer Recherche Data Govv, la plateforme nationale fédérée des données de la recherche

6

Promouvoir l'adoption d'une politique de données sur l'ensemble du cycle des données de la recherche, pour les rendre faciles à trouver, accessibles, interopérables et réutilisables (FAIR)

Développer et structurer les offres d'accompagnement et les outils à disposition des chercheurs

- Créer Recherche Data Govv qui proposera une offre souveraine et certifiée (Core trust seal) de dépôt et de signalement des données de recherche ainsi qu'une offre d'accompagnement des chercheurs, à travers :
 - un réseau territorial d'« ateliers de la donnée » labellisés, mobilisant une large palette de compétences et de métiers pour un accompagnement de proximité ;
 - des centres de référence thématiques qui définissent les pratiques de gestion, de description et d'ouverture des données propres à une discipline ou un domaine de recherche.
- Généraliser la définition et la mise en œuvre effective des plans de gestion des données, qui garantissent la préservation économe, l'ouverture ou le partage de données documentées, en créant les conditions de leur réutilisation et de leur valorisation.

- Mettre en œuvre les recommandations du Comité international des rédacteurs de revues médicales (ICMJE) en développant les déclarations de partage des données (Data Sharing Statement), qui exposent publiquement les conditions et procédures d'accès aux données qui ne peuvent pas être ouvertes.
- Poursuivre le processus de certification (Core trust seal) des entrepôts de données français.
- Accompagner la transposition de la directive européenne sur le droit d'auteur en matière de fouille de texte et de données (TDM) afin de favoriser l'émergence de nouvelles connaissances.

Axe 3 : Ouvrir et promouvoir les codes sources produits par la recherche

7

Valoriser et soutenir la diffusion sous licence libre des codes sources issus de recherches financées sur fonds publics

8

Mettre en valeur la production des codes sources de l'enseignement supérieur, de la recherche et de l'innovation

9

Définir et promouvoir une politique en matière de logiciels libres

Définir et promouvoir une politique en matière de logiciels libres

- Établir une Charte nationale des logiciels libres issus de l'enseignement supérieur, de la recherche et de l'innovation.
- Développer le lien entre données et logiciels grâce au réseau des administrateurs des données, des algorithmes et des codes sources dans les établissements.
- Émettre des recommandations auprès des organismes financeurs pour accompagner au mieux le développement logiciel.
- Faire monter en compétence les structures de valorisation sur les modèles économiques associés à la production de logiciels libres.
- Soutenir Software Heritage et recommander son adoption pour l'archivage et le référencement des codes sources.

Reconnaître les codes sources comme une contribution à la recherche

- Créer un prix du logiciel libre pour la recherche qui récompense les équipes et projets exemplaires dans le domaine.
- Mieux valoriser les productions logicielles dans la carrière des chercheurs, des personnels d'accompagnement à la recherche et dans l'évaluation des structures de recherche.
- Suivre dans le temps la production de codes et logiciels de la recherche française pour en identifier les dynamiques, l'ouverture et les impacts grâce au baromètre de la science ouverte.
- Construire un catalogue des logiciels issus de la recherche en utilisant un schéma de métadonnées normalisé et partagé entre tous les acteurs de l'enseignement supérieur, de la recherche et de l'innovation.

Intégrité scientifique :

Bonnes pratiques en matière de données



ALLEA, [The European Code of Conduct for Research Integrity](#), v3 2023

- Les chercheurs, les institutions et les organismes de recherche **assurent une gestion, une curation et une conservation** adéquates de toutes les données, métadonnées, protocoles, codes, logiciels et autres matériaux de recherche pendant une période raisonnable et clairement définie.
- Les chercheurs, les institutions et les organismes de recherche s'assurent que l'accès aux données est **aussi ouvert que possible, mais aussi fermé que nécessaire** et, le cas échéant, **conforme aux principes FAIR** (autrement dit, facile à trouver, accessible, interopérable et réutilisable) en ce qui concerne la gestion des données.
- Les chercheurs informent les participants à la recherche de la façon dont leurs données seront utilisées, réutilisées, consultées, conservées et supprimées, en **conformité avec le RGPD**.
- Les chercheurs, les institutions et les organismes de recherche font preuve de **transparence** concernant la manière de consulter ou d'utiliser leurs données, métadonnées, protocoles, codes, logiciels et autres matériaux de recherche.
- Les chercheurs, les institutions et les organismes de recherche reconnaissent que les données, métadonnées, protocoles, codes, logiciels et autres matériaux de recherche sont des **produits légitimes de la recherche**, qui peuvent être cités.
- Les chercheurs, les institutions et les organismes de recherche veillent à ce que les **contrats ou les accords** relatifs aux résultats de la recherche comprennent des **dispositions équitables et justes** en ce qui concerne la gestion de leur utilisation, de leur propriété, et/ou de leur protection en vertu des droits de propriété intellectuelle.

➤ **Article D211-2 du Code de la recherche**

(qui reprend en partie le [Décret n° 2021-1572 du 3 décembre 2021](#) relatif au respect des exigences de l'intégrité scientifique dans les établissements de recherche)

Les établissements publics et fondations reconnues d'utilité publique mentionnés au troisième alinéa de l'article [L. 211-2](#) :

- 1° Veillent à ce que les travaux de recherche qu'ils conduisent ou auxquels ils participent respectent les exigences de l'intégrité scientifique ;
- 2° Assurent la formation des personnels et des étudiants au respect de ces exigences ;
- 3° Promeuvent la diffusion des publications en accès ouvert et la **mise à disposition des méthodes, protocoles, données et codes sources associés aux résultats de la recherche** ;
- 4° Définissent les **conditions de conservation, de communication et de réutilisation des résultats bruts** des travaux scientifiques menés en leur sein ;
- 5° Veillent à ce que tout signalement relatif à un éventuel manquement aux exigences de l'intégrité scientifique soit traité selon une procédure établie au regard des recommandations du Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur définies en application des dispositions de l'article [L. 114-3-1](#).

Pour mémoire, le texte du [Décret n° 2021-1572 du 3 décembre 2021](#)

« Article 2

Les établissements publics et fondations reconnues d'utilité publique mentionnés au [troisième alinéa de l'article L. 211-2 du code de la recherche](#) :

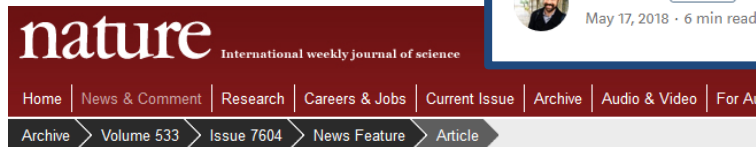
(...) 3° **Promeuvent** la diffusion des publications en accès ouvert et **la mise à disposition des méthodes et protocoles, des données et des codes sources associés aux résultats de la recherche** afin d'en garantir la traçabilité et la reproductibilité. Ils incitent à la publication des résultats de recherche dits négatifs ; (...)

Article 6

[Ils] définissent une **politique de conservation, de communication et de réutilisation des résultats bruts** des travaux scientifiques menés en son sein. A cet effet, ils veillent à la mise en œuvre par leur personnel de **plans de gestion de données** et **contribuent aux infrastructures** qui permettent la conservation, la communication et la réutilisation des données et des codes sources. »

- Les évolutions des textes, relevées par [l'Office français de l'intégrité scientifique](#)

Reproductibilité, Réplicabilité, Robustesse...



NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

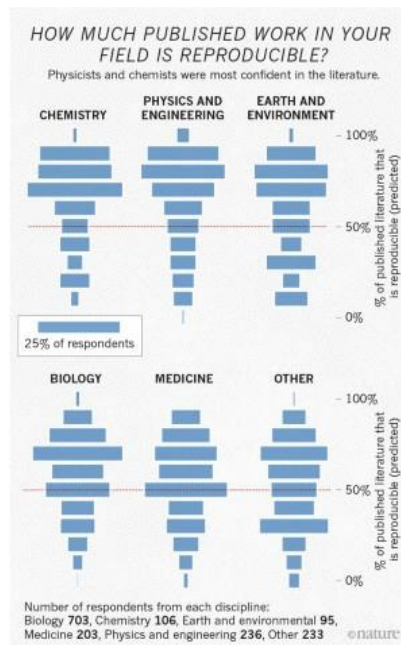
Survey sheds light on the 'crisis' rocking research.

Monya Baker

25 May 2016 | Corrected: 28 July 2016

[Enquête de
mai 2016](#)

Plus de 70% des chercheurs ont échoué à reproduire l'expérience d'un collègue ; et plus de 50% n'ont pas réussi à reproduire leur propre expérience...



A. Hoquet, « Reproduire un résultat scientifique, plus facile à dire qu'à faire », [The Conversation](#), mars 2020 - « L'embellissement en science est banal », [TheMetaNews](#), octobre 2023

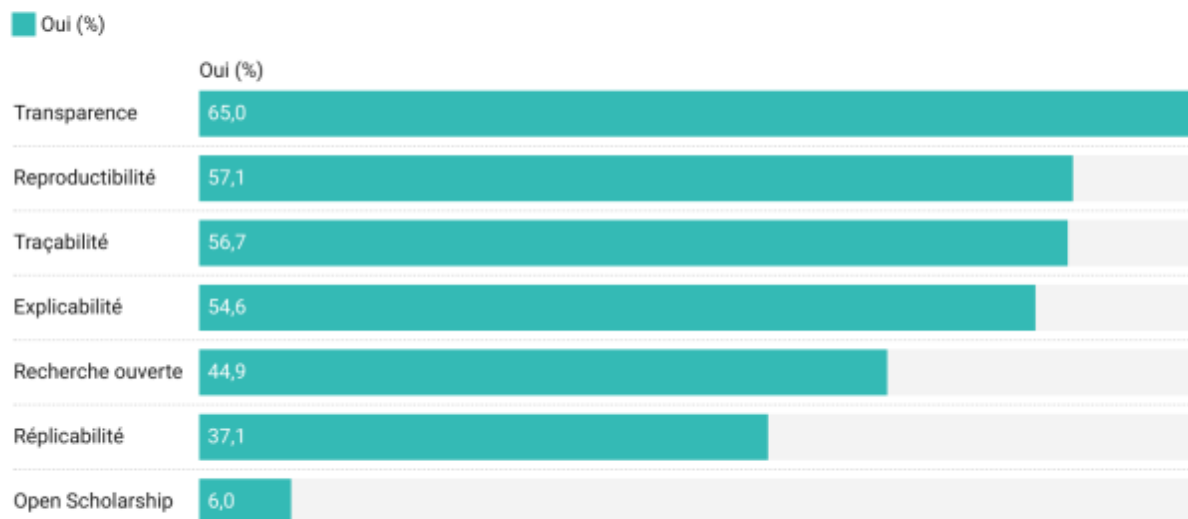
Vers la notion de qualité

Élargir **les enjeux de reproductibilité à ceux de qualité en recherche**. Employer d'autres termes tels que ceux de transparence, de traçabilité, d'explicabilité notamment avec des communautés SHS.



Quels mots représentent un ou des critères de qualité en recherche pour vous?

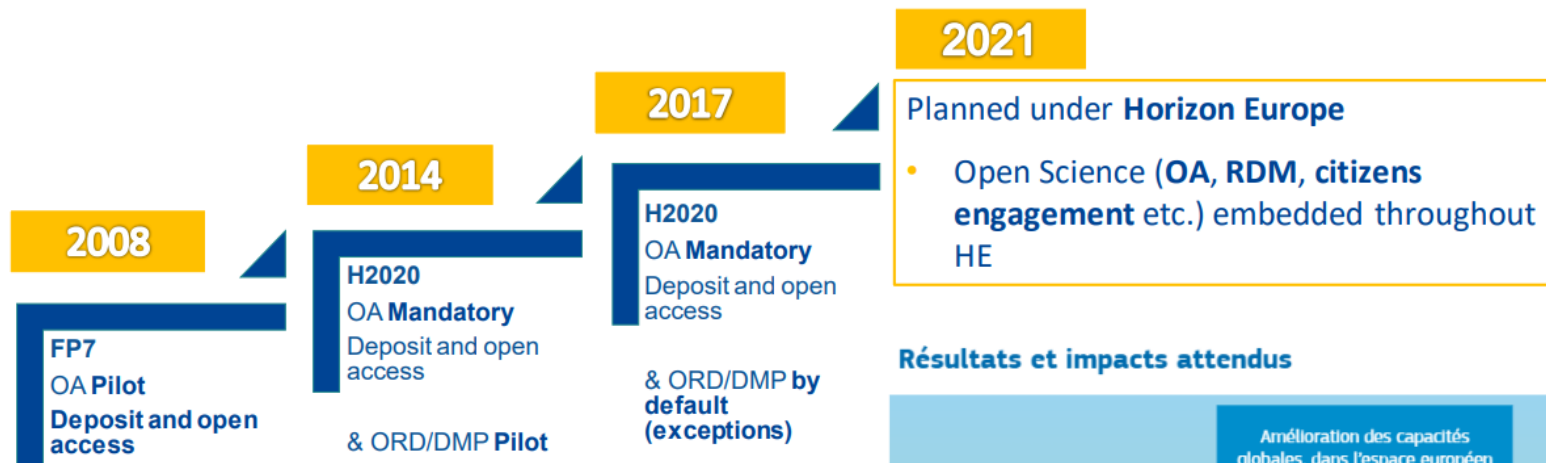
Réponse à choix multiples.



Graphique: Groupe de travail Réussir l'appropriation de la science ouverte • Source: Questionnaire "données et science ouverte" • Créé avec Datawrapper

Les attentes des programmes cadres européens

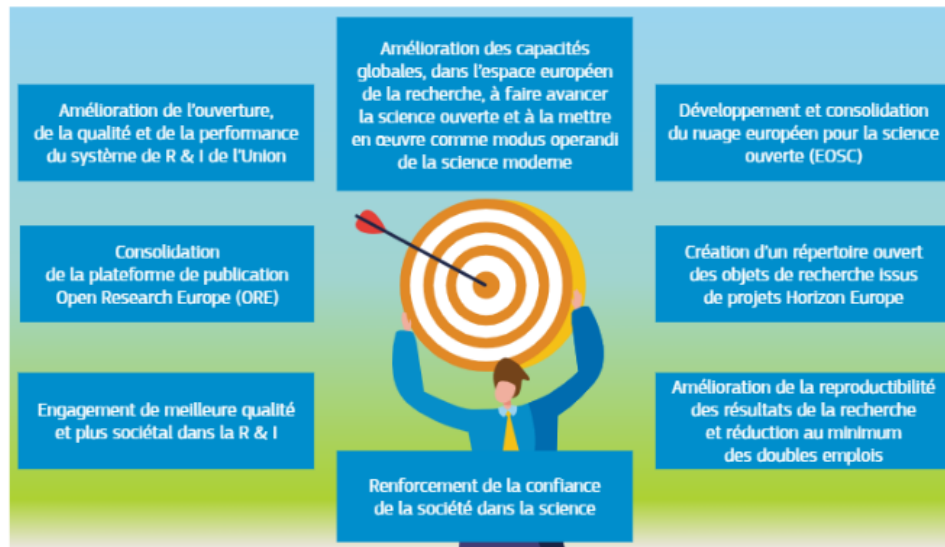
Evolution of open science policy across FPs



Exploitation and Open Science in Horizon Europe

Fiche : La science ouverte dans Horizon Europe

Résultats et impacts attendus





[Horizon Europe - Programme Guide v1.5 - 2022](#)

Voir RDM in Horizon Europe :
[Guide OpenAire](#)

the proposal stage. Proposers should be aware of both mandatory and recommended practices and integrate them into their proposals.

Open science practices include early and open sharing of research (for example through preregistration, registered reports, pre-prints, or crowd-sourcing); research output¹⁸ management; measures to ensure reproducibility of research outputs; providing open access to research outputs (such as publications, data, software, models, algorithms, and workflows); participation in open peer-review; and involving all relevant knowledge actors including citizens, civil society and end users in the co-creation of R&I agendas and contents (such as citizen science).

These practices are explained and relevant resources provided in a separate section further below (open science practices and resources).

Mandatory open science practices

- Some open science practices are **mandatory for all beneficiaries per the grant agreement**. They concern:
 - open access to scientific publications under the conditions required by the grant agreement;
 - responsible management of research data in line with the FAIR principles of 'Findability', 'Accessibility', 'Interoperability' and 'Reusability', notably through the generalised use of data management plans, and open access to research data under the principle 'as open as possible, as closed as necessary', under the conditions required by the grant agreement;
 - information about the research outputs/tools/instruments needed to validate the conclusions of scientific publications or to validate/re-use research data;
 - digital or physical access to the results needed to validate the conclusions of scientific publications, unless exceptions apply;
 - in cases of public emergency, if requested by the granting authority, immediate open access to all research outputs under open licenses or, if exceptions apply, access under fair and reasonable conditions to legal entities that need the research outputs to address the public emergency¹⁹.

These obligations are described in the Model Grant Agreement (Article 17) and detailed guidelines on complying with them are provided in the Annotated Grant Agreement (Article 17).

- Some open science practices are **mandatory per specific work programmes or call conditions**, which may provide for additional obligations to adhere to open science practices.

Recommended open science practices

La politique des agences de financements en France



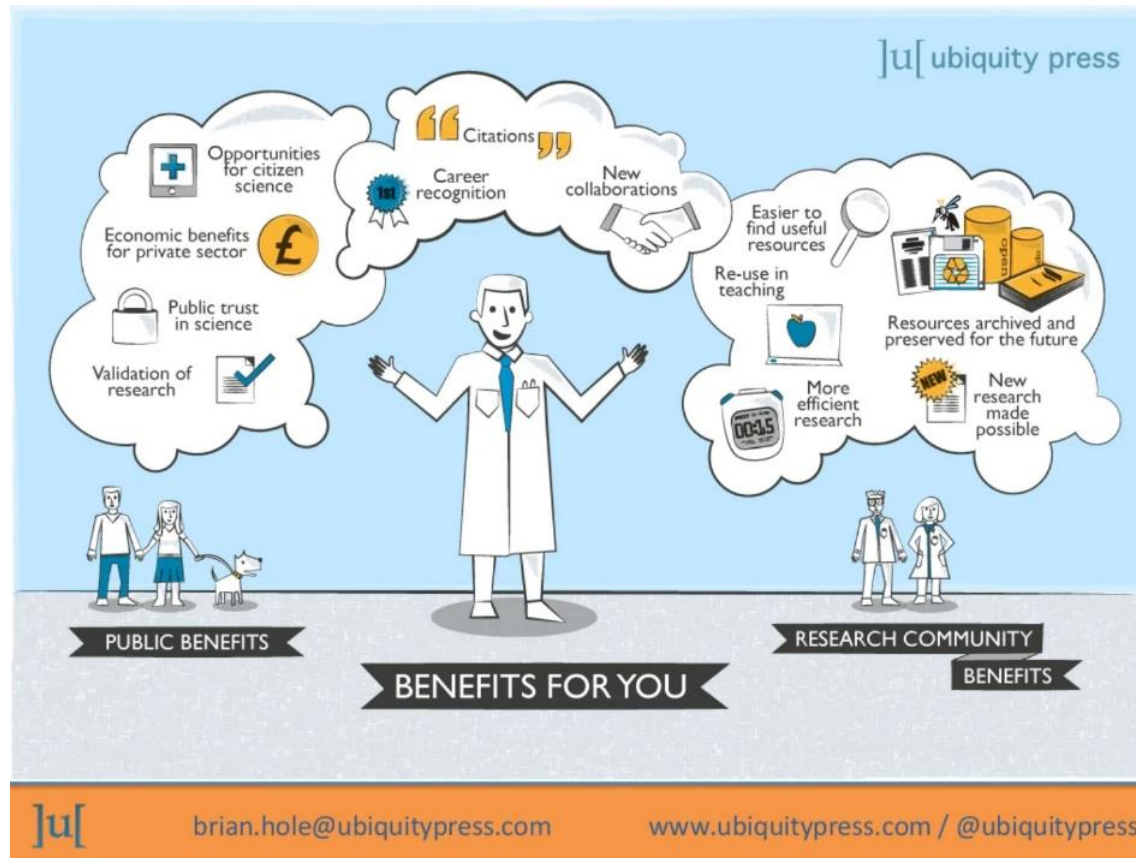
Une politique commune depuis mars 2022

- Favoriser le libre accès pour les publications scientifiques issues des projets financés
- Promouvoir le partage et l'ouverture des données de la recherche
- Partager et harmoniser nos pratiques et modalités d'évaluation des projets
- Informer et sensibiliser aux pratiques de science ouverte
- Ouvrir les données relatives aux projets financés

Dans la Feuille de route 2023

- Déployer un modèle commun de plan de gestion des données structuré
- Prendre part aux discussions sur la réforme de l'évaluation de la recherche

De multiples bénéfices



<https://fr.slideshare.net/brianhole/from-open-access-to-open-data>

II. DÉFINITIONS

1. Données
2. Codes, algorithmes, logiciels
3. Cadre juridique

II. DÉFINITIONS

1. Données

Données de la recherche

« Les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. »



OCDE, *Recommandation du Conseil concernant l'accès aux données de la recherche financée sur fonds publics*, 2021

(1^e occurrence : Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics, 2007).



« Cette expression **ne couvre pas** les carnets de laboratoire, analyses préliminaires ou projets de documents scientifiques, programmes de travaux de recherche futurs, examens par les pairs, communications personnelles avec des collègues, ou objets matériels (échantillons de laboratoire, souches bactériennes ou animaux de laboratoire, par exemple). »



« **Autres objets numériques pertinents** au regard de la recherche financés sur fonds publics : métadonnées, algorithmes, flux de travail, modèles et logiciels (y compris le code connexe) issus de la recherche financée en tout ou partie sur fonds publics et utilisés dans le cadre de travaux de recherche et de développement. »

Typologie des données de recherche

Données d'observation

- capturées en temps réel
- habituellement uniques, impossible à reproduire

Ex.: mesures sismiques, images d'une étoile en fin de vie, enquêtes sociologiques, fouilles archéologiques...

Données d'expérimentation

- obtenues à partir d'équipements de laboratoire
- souvent reproductibles, parfois coûteuses

Ex.: séquences peptidiques, résultats de réactions chimiques, observations sur des individus en situation de test...

Données de simulation

- générées par des modèles informatiques
- souvent reproductibles si le modèle est correctement documenté

Ex.: modélisation du changement climatique, « reproduction » du Big Bang, modèles économiques...

C.L. Borgman, *Big data, little data, no data*, 2015 / *Qu'est-ce que le travail scientifique des données ?* [2020](#)

+ Données compilées ou dérivées

- issues du traitement de données brutes
- souvent reproductibles mais coûteuses

Ex.: bases de données compilées, corpus textuel préparé pour le TDM...

+ Données « de référence »

- validées par la communauté
- réutilisables

Ex.: décodage du génome humain, certaines données astronomiques...

Records – Traces (C. Borgman)

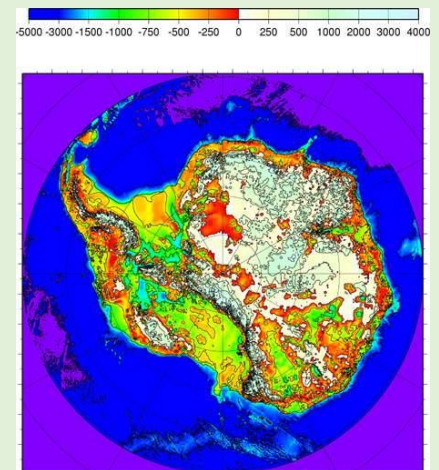
- documents témoignant d'un phénomène ou d'une activité humaine
- uniques ou non

Ex.: fonds de photographies, documents d'archives, textes de loi, ouvrages littéraires...

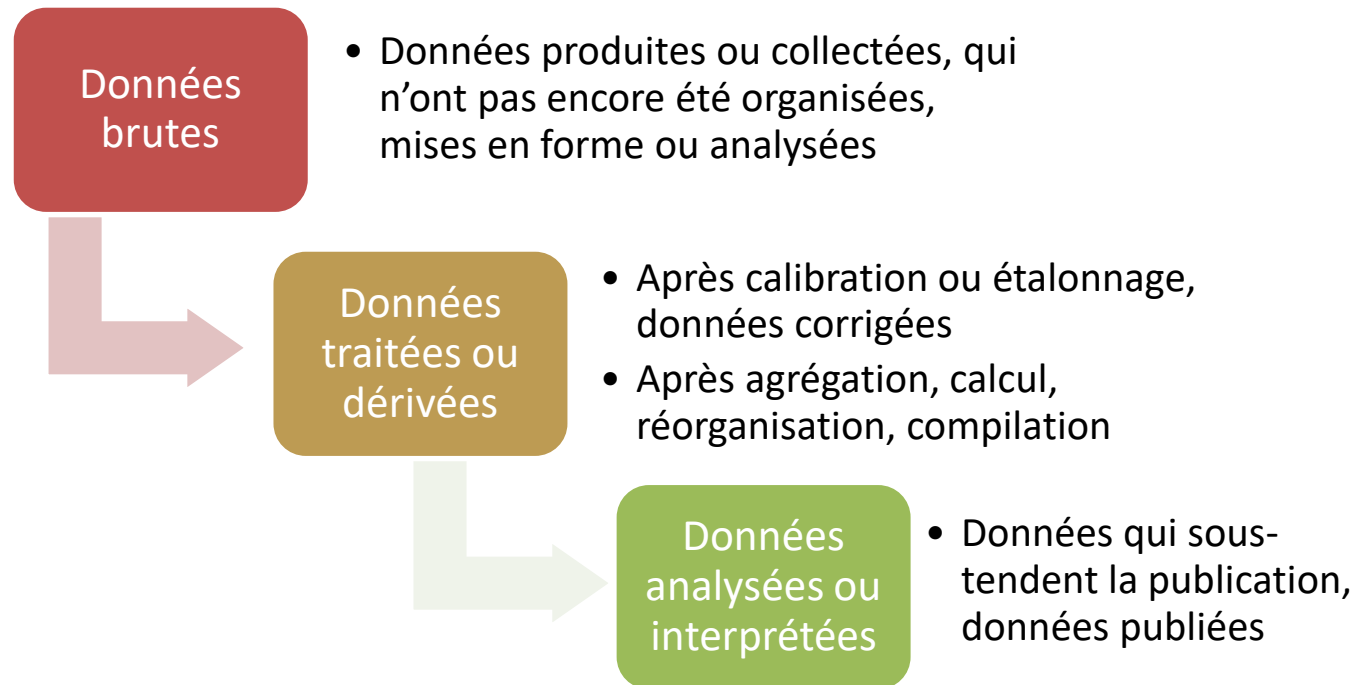
Exemples en sciences de l'environnement

- Données issues d'un poisson-robot dans la lagune de Venise
- Indice de sensibilité environnementale du littoral français.
- Données sur la population de l'Alouette des champs, dans la Zone Atelier Plaine et Val de Sèvre, au sud de Niort, en Nouvelle-Aquitaine.
- Inventaire National du Patrimoine Naturel
- Carte de l'altitude du socle antarctique

[Source : Parcours sur la gestion des données de recherche en sciences de l'environnement - Doranum 2021](#)



Le niveau de traitement



Variations : Typologie des données du [CERN](#) pour le LHC ([data policy 2020](#))

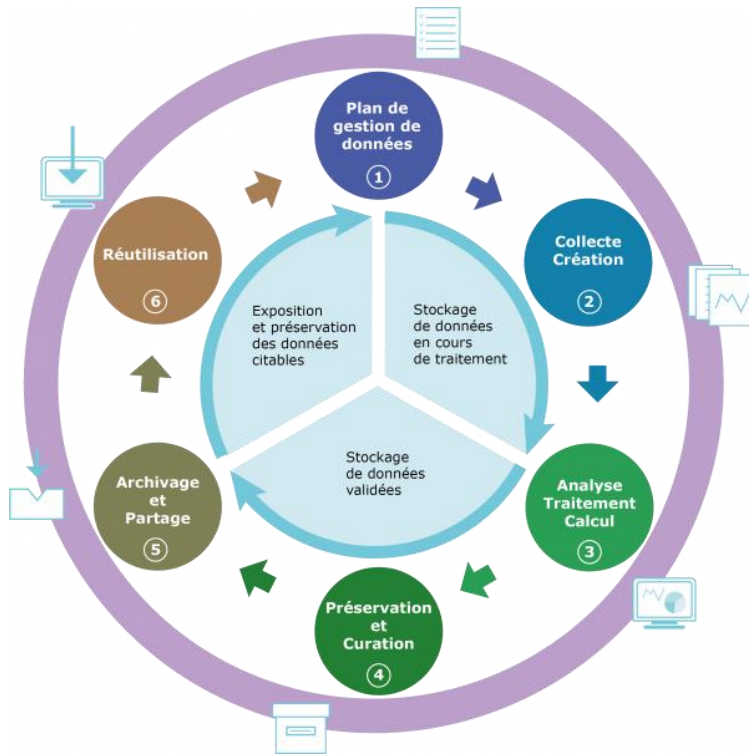
1. Données à l'appui des résultats publiés (*published results*)
2. Ensembles conçus à des fins pédagogiques ou de vulgarisation (*outreach and education data*)
3. Données requises pour produire les résultats scientifiques (*reconstructed data*)
4. Données brutes (*raw data*)

Des données... données ?

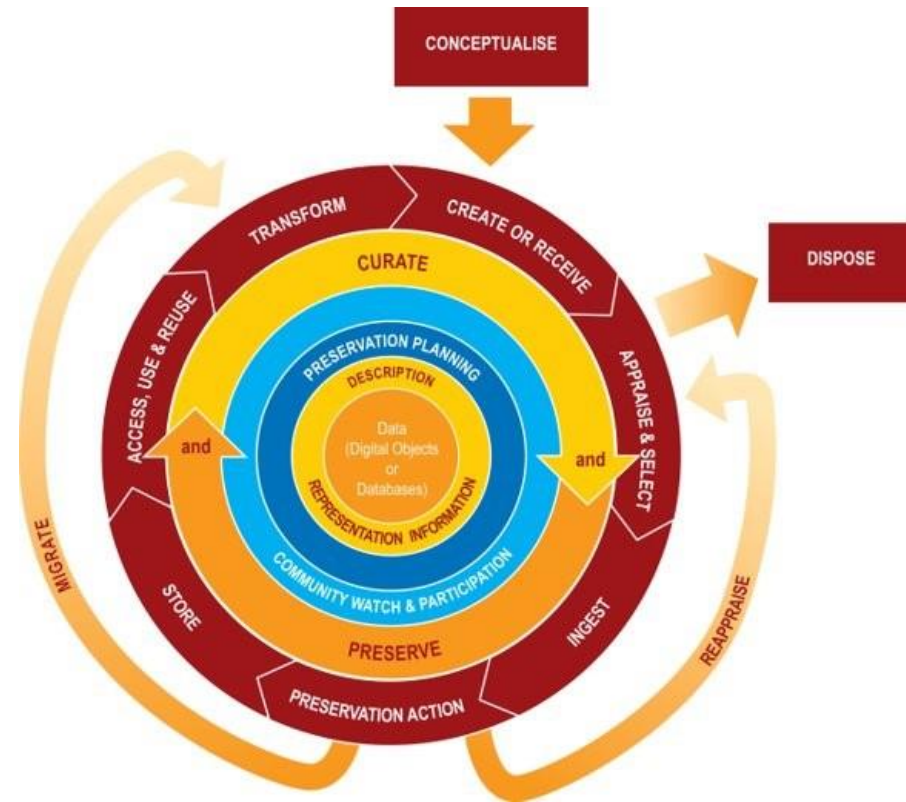


- *Capta* ([H. Becker 1952](#))
- *Obtenues* ([B. Latour 1993](#))
- *Cooked* (G. Bowker dans [L. Gitelman 2013](#))
- *Data Travel* ([S. Leonelli 2016](#))
- *Datafication* ([F. Jaton et D. Vinck 2016](#) à propos de [G. Bastin et J.-M. Francony 2016](#)), notamment en SHS ([O. Hamzaoui et G. Marrel, 2023](#))
- *Mise en données* ([C. Gruson-Daniel et C. de Quatrebarbes 2018](#))

Le cycle de vie des données



[Cycle de vie des données](#),
Science ouverte Grenoble Alpes,
2022, Inspiré de [Research Data
Lifecycle](#), UK Data Archive



[DCC Lifecycle](#)

Les principes FAIR

scientific **data**

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [scientific data](#) > [comment](#) > article

[Open Access](#) | [Published: 15 March 2016](#)

The FAIR Guiding Principles for scientific data management and stewardship

[Mark D. Wilkinson](#), [Michel Dumontier](#), [IJsbrand Jan Aalbersberg](#), [Gabrielle Appleton](#), [Myles Axton](#), [Arie Baak](#), [Niklas Blomberg](#), [Jan-Willem Boiten](#), [Luiz Bonino da Silva Santos](#), [Philip E. Bourne](#), [Jildau Bouwman](#), [Anthony J. Brookes](#), [Tim Clark](#), [Mercè Crosas](#), [Ingrid Dillo](#), [Olivier Dumon](#), [Scott Edmunds](#), [Chris T. Evelo](#), [Richard Finkers](#), [Alejandra Gonzalez-Beltran](#), [Alasdair J.G. Gray](#), [Paul Groth](#), [Carole Goble](#), [Jeffrey S. Grethe](#), ... [Barend Mons](#)  [+ Show authors](#)

[Scientific Data](#) **3**, Article number: 160018 (2016) | [Cite this article](#)

502k Accesses | **4861** Citations | **2035** Altmetric | [Metrics](#)

 An [Addendum](#) to this article was published on 19 March 2019

Abstract

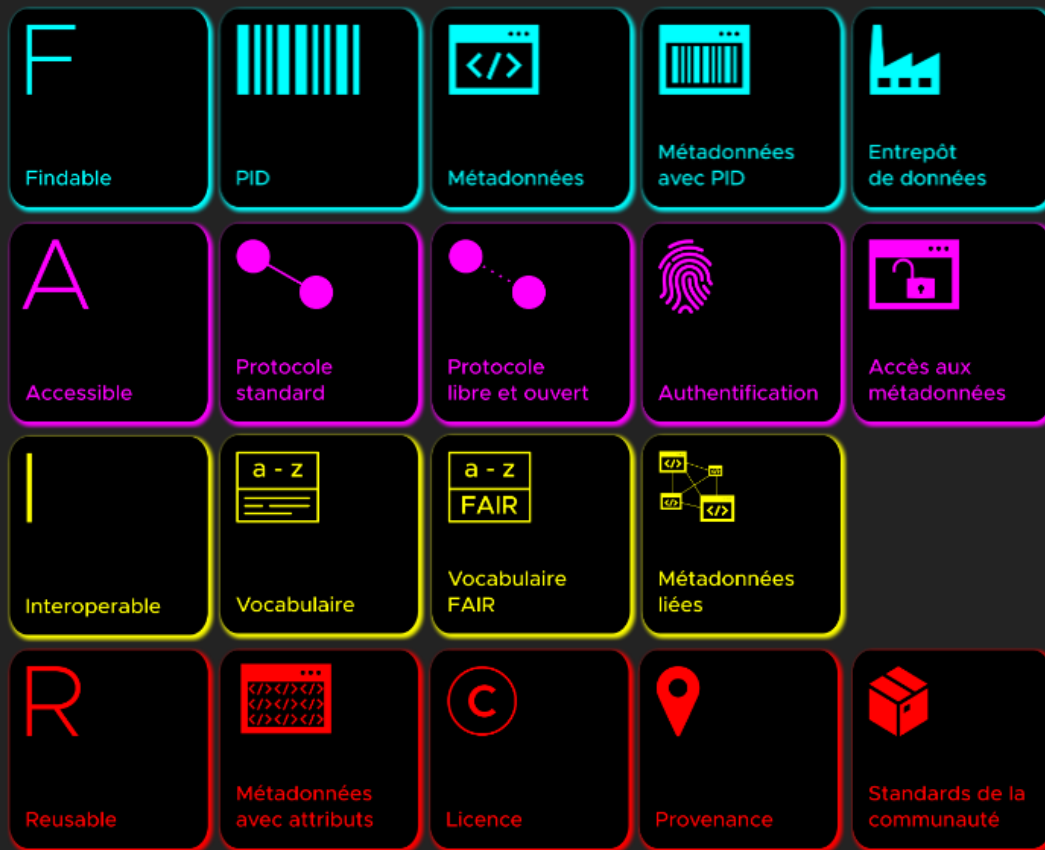
There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A

« [The FAIR Guiding Principles for scientific data management and stewardship](#) », Scientific data, 2016



Les principes FAIR

Les chercheurs s'appuient sur les connaissances scientifiques antérieures, notamment sur les résultats publiés dans les articles scientifiques. La reproductibilité des résultats, ainsi que leur croisement, ne sont cependant envisageables qu'avec des données originelles et leurs conditions d'obtention. C'est pourquoi la science ouverte vise à faciliter l'accès aux publications scientifiques et aux données de la recherche. Cette facilitation s'accompagne d'un certain nombre de mesures pour rendre les données scientifiques facilement découvrables, accessibles, interopérables et réutilisables. Ce sont les principes FAIR : Findable, Accessible, Interoperable, Reusable.



Guide pratique

En français : <https://dorum.fr/appli-fair-aware-pleine-page-vf/>

En anglais : <https://www.fairsfair.eu/fair-aware>



FACILE A TROUVER

1. Savez-vous qu'un jeu de données déposé dans un entrepôt de données doit se voir attribuer un identifiant international, unique, pérenne et qui dirige vers une page de présentation des données ? ⓘ
2. Savez-vous que lorsque vous déposez un jeu de données dans un entrepôt, vous devez fournir certains détails (appelés métadonnées de référencement) afin de rendre vos données faciles à trouver, compréhensibles et réutilisables par d'autres personnes ? ⓘ
3. Savez-vous qu'un entrepôt qui donne accès à vos données doit fournir les métadonnées qui décrivent celles-ci dans un format apte à être lu par des machines aussi bien que par des humains ? ⓘ

ACCESSIBLE

4. Savez-vous que l'accès à votre jeu de données peut nécessiter d'être restreint (soumis à autorisation) et que vos métadonnées devraient comprendre des informations sur les licences applicables en cas de ré-utilisation de données ? ⓘ
5. Savez-vous que vos métadonnées doivent rester disponibles sur le long terme, même dans le cas où les données ne sont plus accessibles ? ⓘ



INTEROPERABLE

6. Savez-vous que les métadonnées qui décrivent vos jeux de données doivent se baser sur des vocabulaires contrôlés (standardisés, structurés sous la forme d'ontologies), autant que possible ? ⓘ

REUTILISABLE

7. Savez-vous que vos métadonnées doivent inclure des informations relatives à la provenance des données collectées et/ou générées ? ⓘ
8. Savez-vous que les métadonnées décrivant vos données doivent suivre des standards reconnus par la communauté scientifique dans votre domaine de recherche ? ⓘ
9. Savez-vous que vos données doivent être déposées de préférence dans un format de fichier ouvert (pour faciliter leur réutilisation) qui soit pris en charge par votre entrepôt, pour une conservation à long terme. ⓘ
10. Savez-vous que la conservation de vos jeux de données à long terme exige l'intervention de professionnels de la gestion et de la conservation des données ? ⓘ



Pour tester des données :

[FAIR data assessment tool \(ARDC\)](#)
[Automated FAIR data assessment tool \(FUJI\)](#)

Des données faciles à trouver ?



FACILE A TROUVER

1. Savez-vous qu'un jeu de données déposé dans un entrepôt de données doit se voir attribuer un identifiant international, unique, pérenne et qui dirige vers une page de présentation des données ? ⓘ

2. Savez-vous que lorsque vous déposez un jeu de données dans un entrepôt, vous devez fournir certains détails (appelés métadonnées de référencement) afin de rendre vos données faciles à trouver, compréhensibles et réutilisables par d'autres personnes ? ⓘ

3. Savez-vous qu'un entrepôt qui donne accès à vos données doit fournir les métadonnées qui décrivent celles-ci dans un format apte à être lu par des machines aussi bien que par des humains ? ⓘ

- **L'idée** : permettre à toute personne intéressée d'avoir connaissance de l'existence de données en interrogeant n'importe quel moteur de recherche, sans devoir explorer précisément la plateforme sur laquelle elles sont conservées
- **La solution** : décrire les caractéristiques de mes données de manière précise et compréhensible, à la fois pour un ordinateur et pour un humain, de façon à ce qu'elles soient renvoyées comme résultat d'une requête

La mise en pratique :

- Décrire ses données avec un maximum d'informations, notamment de mots-clés, par des **métadonnées**
- (Faire) attribuer un identifiant unique et pérenne à son jeu de données
 - **DOI** ou assimilé
- Déposer ses données dans un entrepôt indexé par les moteurs de recherche
 - Chercher un **entrepôt compatible avec le protocole OAI-PMH** :

Des données accessibles ?



ACCESSIBLE

4. Savez-vous que l'accès à votre jeu de données peut nécessiter d'être restreint (soumis à autorisation) et que vos métadonnées devraient comprendre des informations sur les licences applicables en cas de ré-utilisation de données ? ⓘ

5. Savez-vous que vos métadonnées doivent rester disponibles sur le long terme, même dans le cas où les données ne sont plus accessibles ? ⓘ

- **L'idée** : une fois identifiées par un individu ou un moteur de recherche, les données doivent pouvoir être obtenues (éventuellement sous certaines conditions)
- **La solution** : bien préciser les conditions selon lesquelles on souhaite rendre des données accessibles ou non, et les automatiser autant que possible

La mise en pratique :

- Déposer les données dans un **entrepôt interrogeable** par internet
- Choisir un entrepôt qui intègre un **contrôle d'accès** (si on le souhaite)
- Détailler les **conditions d'accès** à chacun des jeux de données
- Si des **programmes ou logiciels** sont indispensables à l'utilisation des données, les **fournir** ou expliquer comment récupérer ces logiciels
- Même si les données ne sont pas accessibles, faire en sorte que leur description précise et les **informations bibliographiques** les concernant le soient
 - Ces informations sont à préciser dans les **métadonnées** qui décrivent les données, à créer directement dans un **formulaire au moment du dépôt** dans l'entrepôt, ou en amont du dépôt **au sein d'un fichier à part**

Des données interoperables ?

INTEROPERABLE

6. Savez-vous que les métadonnées qui décrivent vos jeux de données doivent se baser sur des vocabulaires contrôlés (standardisés, structurés sous la forme d'ontologies), autant que possible ? ⓘ

- **L'idée** : faire en sorte que les données puissent être utilisées conjointement à d'autres jeux de données, ou par d'autres interfaces techniques que celles qui ont permis de les produire/collecter
- **La solution** : créer et décrire les données selon des pratiques très répandues ou très bien documentées

La mise en pratique :

- Utiliser des **formats de fichiers ouverts**, non-propriétaires, si possible normés
- Décrire les données en **termes non-ambigus** et employés par la communauté la plus vaste possible (ex.: un identifiant-chercheur plutôt qu'un nom-prénom)
- Associer à les données des mots-clés qui emploient un **vocabulaire de description courant** dans la discipline concernée
- Si on a besoin d'employer des termes très spécifiques pour décrire les données, proposer des **équivalents** dans des vocabulaires plus répandus ou mieux connus

Des données réutilisables ?



REUTILISABLE

7. Savez-vous que vos métadonnées doivent inclure des informations relatives à la provenance des données collectées et/ou générées ? ⓘ
8. Savez-vous que les métadonnées décrivant vos données doivent suivre des standards reconnus par la communauté scientifique dans votre domaine de recherche ? ⓘ
9. Savez-vous que vos données doivent être déposées de préférence dans un format de fichier ouvert (pour faciliter leur réutilisation) qui soit pris en charge par votre entrepôt, pour une conservation à long terme. ⓘ
10. Savez-vous que la conservation de vos jeux de données à long terme exige l'intervention de professionnels de la gestion et de la conservation des données ? ⓘ

- **L'idée** : indiquer très précisément comment et par qui les données pourront être réutilisées
- **La solution** : fournir ces informations de manière à ce qu'elles soient compréhensibles à la fois par un individu et par une machine

La mise en pratique :

- Choisir une **licence de diffusion** parmi celles recommandées par l'institution de rattachement, la communauté disciplinaire ou le gouvernement
- Intégrer ces informations de licence dans les **métadonnées** qui décrivent les données (à préciser au moment du dépôt dans l'entrepôt)
- Préciser la façon dont on souhaite être crédité(e) par les réutilisateurs des données, en fournissant par exemple un **modèle de citation**

Les principes CARE

#BeFAIRandCARE



FAIR Guiding Principles for scientific data management and stewardship can be accessed here

<https://www.gida-global.org/care>

Voir aussi « [The CARE data principles for Indigenous data governance](#) », Data Science Journal, 2020

II. DÉFINITIONS

2. Codes, algorithmes, logiciels

Codes, algorithmes, logiciels...



[Rapport Bothorel](#), 2020

- **Code** : ensemble d'instructions exécutables par un ordinateur
 - Code source : instructions rédigées et lisibles par l'humain (langage de programmation)
 - « Exécutable » : lisible par les machines (binaires)
- **Algorithme** : « description d'une suite finie et non ambiguë d'étapes (ou d'instructions) permettant d'obtenir un résultat à partir d'éléments fournis en entrée » (CNIL)
 - Algorithme public : procédure administrative dont tout ou partie est informatisée et qui intervient dans un processus de décision pour les citoyens.
- **Logiciel** : environnement comprenant le code source et l'exécutable, la documentation, les licences associées, des exemples d'utilisation...

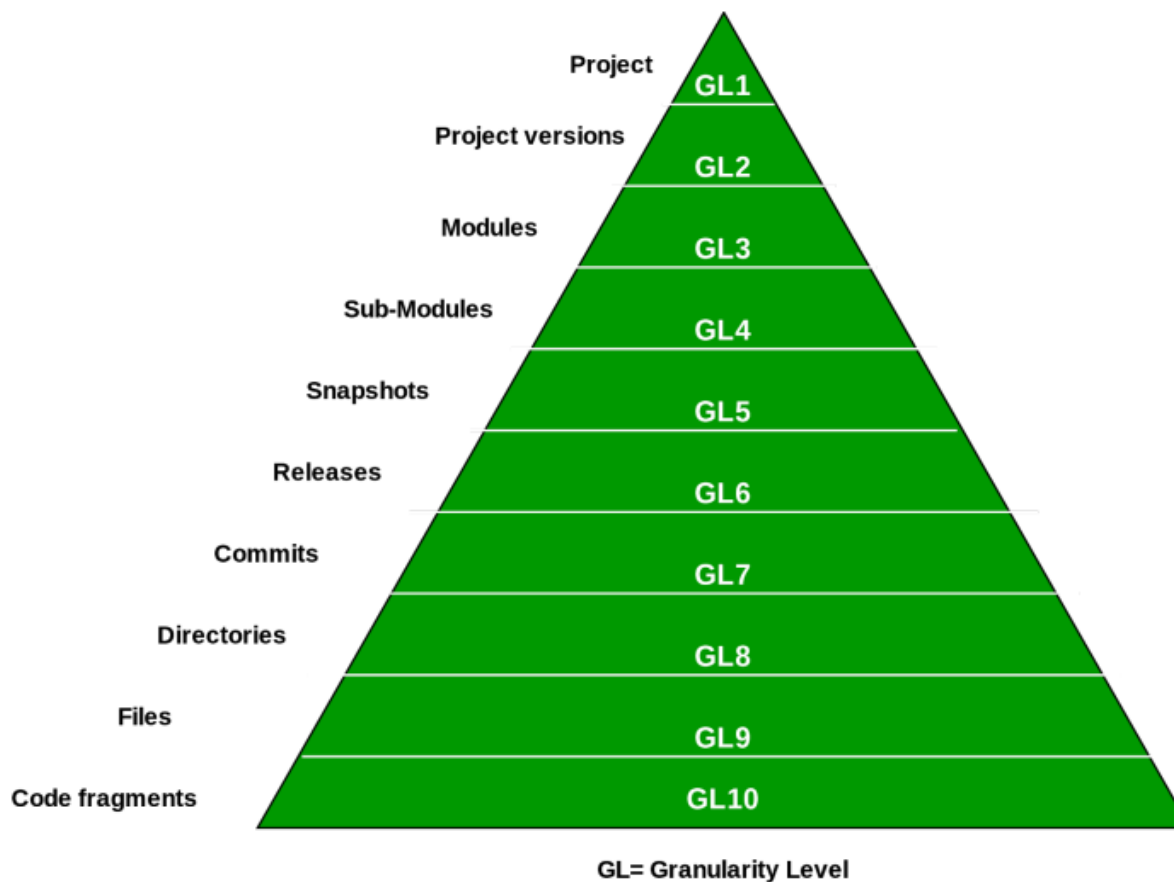
Les logiciels de recherche

- **Les logiciels de recherche** sont développés pour répondre à des **besoins spécifiques de la science**. Ils sont conçus, maintenus, et utilisés par des scientifiques (chercheurs et ingénieurs) et institutions de recherche, éventuellement dans une dimension internationale.
- Ils peuvent découler de **travaux de recherche** comme ils peuvent les favoriser, notamment par des publications avant/sur/autour/avec le logiciel.
- Ceux-ci peuvent se formaliser de **différentes façons** (une plateforme, un intergiciel, un workflow ou une bibliothèque, module ou greffon d'un autre logiciel) et être ainsi en interaction dans un **écosystème** ou au contraire plus autonomes.

Définition du COSO, citée par Violaine Loubet ([Formation ADBU-MediaNormandie](#), 2023)

Un outil pour la recherche, un produit de la recherche, un objet de recherche

Niveaux de granularité d'un projet logiciel



[RDA/Force11 Software Source Code Identification WG](#), 2020

Typologie de projets logiciels

6.2. Types of software that require different levels of management

This section gives examples of software types and groups them according to the software management levels as defined in Section 6.1.

6.2.1. Software that requires low level management

- A script that is used to create and format a single figure for a publication, for example, when using a plotting package such as *ggplot2* (R) or *Matplotlib* (Python).
- Software written as part of a project to automate an administrative or routine process, e.g. monitoring a process or generating document templates.
- Software written specifically for the analysis of a single experiment, data processing, and presentation of its results.³²

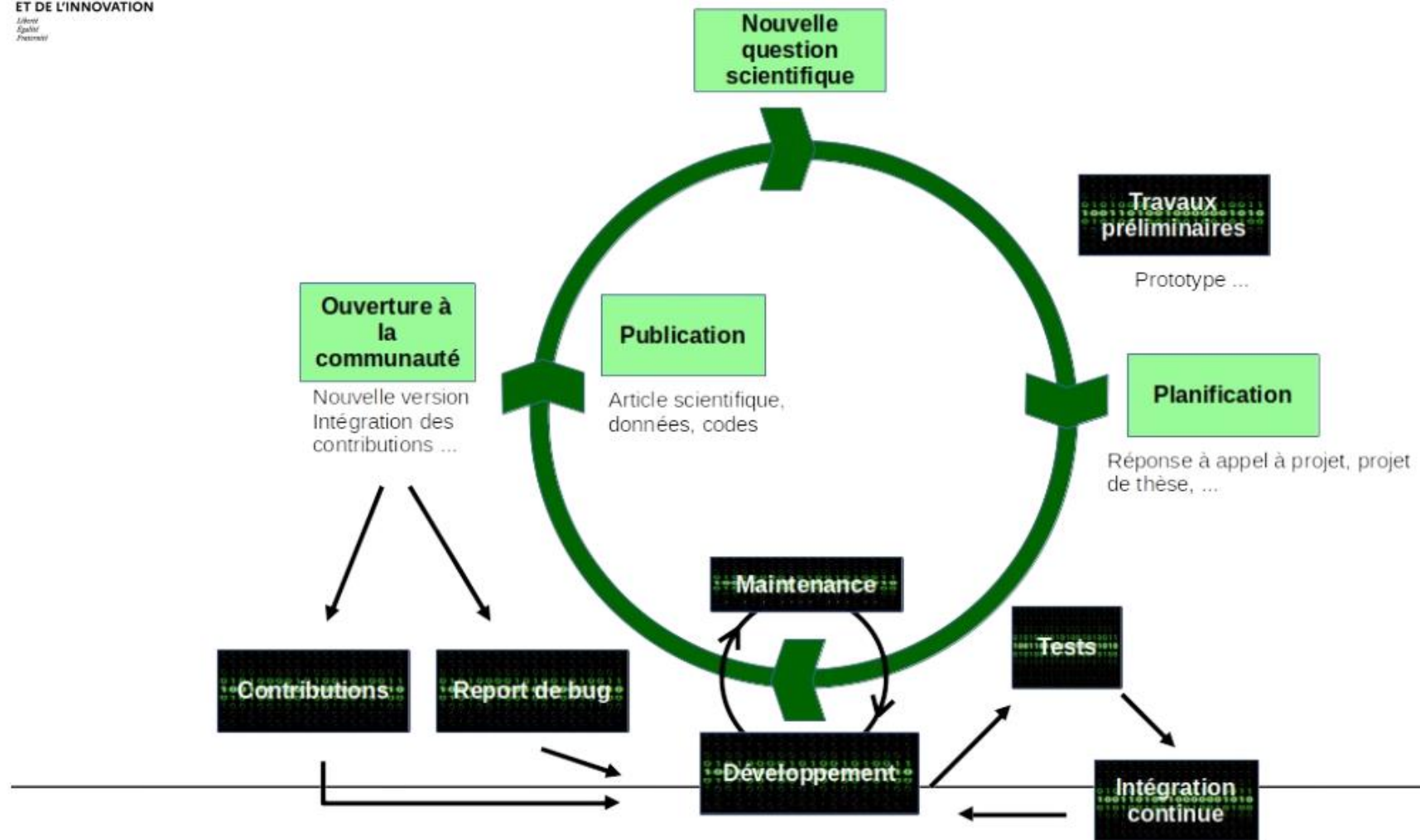
6.2.2. Software that requires medium level management

- Software that implements a new or higher performance algorithm and can be applied to different input data.
- Simulation software that implements one or more models and/or numerical methods, e.g. computational fluid dynamics, chemical interactions, planetary evolution, partial differential equation solvers, numerical integration, etc.

6.2.3. Software that requires high level management software

- Software used in production on which an institute, department, or instrument depends on for their operation, e.g. software that is used for pre-processing all data coming from a particular telescope.
- Software that cannot be rewritten during a project's lifetime. If one requires functionality from high-impact software, replacing it may threaten a project.

Cycle de vie d'un code de recherche



Codes de recherche ≠ données de recherche

- Les données de recherche sont plutôt passives, les codes sont intrinsèquement **vivants**
 - On ne change en général pas les données, collectées dans un contexte bien défini
 - On change éventuellement la façon dont on les traite et on les analyse (grâce à des codes)
 - Les codes sont associés à une (ou des) **action(s)** : création de connaissances, transformation d'informations, visualisation, ...
 - Le code peut être réutilisé tel que, en reproduisant son environnement et toutes ces dépendances mais on a surtout envie de le **modifier pour l'adapter** à nos besoins propres ou **l'enrichir de nouvelles fonctionnalités**
- Les codes s'appuient sur des **dépendances et tout un environnement logiciel et matériel** qui évolue sans cesse
 - Cela complexifie les questions de **reproductibilité**
- Les codes représentent un travail de création, et correspondent à un **cadre juridique différent de celui des données**

Violaine Loubet ([Formation ADBU-MediaNormandie](#), 2023)

FAIR4RS

➤ [Sur l'adaptation des principes](#)

F: Software, and its associated metadata, is easy for both humans and machines to find.
F1. Software is assigned a globally unique and persistent identifier. • F1.1. Components of the software representing levels of granularity are assigned distinct identifiers. • F1.2. Different versions of the software are assigned distinct identifiers. F2. Software is described with rich metadata. F3. Metadata clearly and explicitly include the identifier of the software they describe. F4. Metadata are FAIR, searchable and indexable.
A: Software, and its metadata, is retrievable via standardized protocols.
A1. Software is retrievable by its identifier using a standardized communications protocol. • A1.1. The protocol is open, free, and universally implementable. • A1.2. The protocol allows for an authentication and authorization procedure, where necessary. A2. Metadata are accessible, even when the software is no longer available.
I: Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.
I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards. I2. Software includes qualified references to other objects.
R: Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software).
R1. Software is described with a plurality of accurate and relevant attributes. • R1.1. Software is given a clear and accessible license. • R1.2. Software is associated with detailed provenance. R2. Software includes qualified references to other software. R3. Software meets domain-relevant community standards.

Table 1: The FAIR Principles for Research Software

Voir les listes de recommandations :

- [Fair Software](#)
- [Metrics for automated Software assessment](#) (article sur le [blog open Science Pasteur](#))
- [Software Sustainability institute](#)
- [Library Carpentry](#)

[FAIR Principles for Research Software](#)
[\(FAIR4RS Principles\) \(1.0\)](#)

II. DÉFINITIONS

3. Cadre juridique

Données de recherche : de multiples droits



Voir [Cours introductif sur les aspects juridiques et éthiques](#), DoRANum, 2022
Et Agnès Robin, *Droit des données de la recherche*, 2022

Des documents administratifs

- *Code régissant les relations entre le public et l'administration (CRPA)*
 - **Principe d'ouverture par défaut et de libre réutilisation**
- *Loi pour une République numérique (art. 30 II = Code de recherche L533-4)*
 - Précision sur le cadre d'application propre à la recherche

Périmètre : « activités de recherche **financées au moins pour moitié** par des dotations » (publiques, nationales, européennes) sous réserve de

 - Respecter le cas échéant les **droits spécifiques et réglementations particulières**
 - Avoir été **rendues publiques** par le chercheur, l'établissement ou l'organisme
 - Être **achevées**

La loi pour une République numérique (octobre 2016)

Article 30

II.- Dès lors que les données issues d'une activité de recherche financée au moins pour moitié par des dotations de l'Etat, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne ne sont **pas protégées par un droit spécifique ou une réglementation particulière** et qu'elles ont été **rendues publiques** par le chercheur, l'établissement ou l'organisme de recherche, **leur réutilisation est libre.** »

Les documents administratifs sont librement communicables à toute personne qui en fait la demande, sauf exceptions légales.

Si possible, mise à disposition exhaustive et immédiate, pour tous
Sinon, mise à disposition ciblée, après examen et sur demande

Communication obligatoire

- Données **géographiques informatisées** (directive INSPIRE): communication obligatoire d'office
- Données relatives à des émissions de substances dans l'**environnement** (convention d'Aarhus): communication obligatoire sur demande

Communication interdite

- Données présentant des **risques** pour la défense nationale
- Données à risque pour la **sécurité** de l'État, la sécurité publique, la sécurité de l'établissement
- Données protégées par des **secrets** (industriel et commercial, médical, financier...)

Communication sous conditions

- Données présentant des risques pour la protection du **potentiel scientifique** et technique de la nation
- Données protégées par le **droit d'auteur** ou autre droit de propriété intellectuelle
- Données **personnelles**
- Données **statistiques**
- Données liées à un contrat avec un tiers

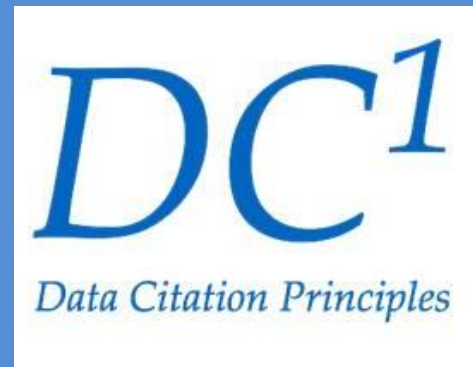
Données et Auctorialité

A. Robin, *Droit des données de la recherche*, 2022, §455-457

- Jeux de données comme nouvelle forme de publication
- Des règles de citation qui doivent être aussi exigeantes que pour les autres produits de la recherche

FORCE11, [Joint Declaration of Data Citation Principles](#), 2014

- *Importance*
- *Credit and Attribution*
- *Evidence*
- *Unique identification*
- *Access*
- *Peristence*
- *Specificity and Verifiability*
- *Interoperability and Flexibility*



Logiciels et créations informatiques

- Les « logiciels, y compris le matériel de conception préparatoire » (documentation, code, structure de données...) sont considérés comme des **œuvres de l'esprit** => protégeables au titre du **droit d'auteur** ([CPI, L 112-2](#))
- Interface graphique : protégeable par droit d'auteur
- Algorithmes : assimilés aux formules mathématiques, ne sont donc pas protégeables par le droit d'auteur

Voir « [Comment protéger un logiciel ou une création informatique ?](#) »,
Techniques de l'ingénieur, 2017

Des règles spécifiques pour le logiciel

- Dès lors que le logiciel est conçu dans l'exercice des fonctions du salarié ou sur instruction de l'employeur
 - Droits moraux : seule la **paternité** est conservée par **l'auteur du logiciel**
 - Droits patrimoniaux (**exploitation**) : dévolus à **l'employeur**
- Sans autorisation, les utilisateurs peuvent
 - Faire une copie de sauvegarde
 - Observer, tester le fonctionnement du logiciel
 - Reproduire le code si c'est indispensable pour son interopérabilité

Voir « [Comment protéger un logiciel ou une création informatique ?](#) »,
Techniques de l'ingénieur, 2017

Logiciels et brevets

- Les programmes d'ordinateur ne sont pas considérés comme des inventions ([CPI, L611-10](#))
- Le logiciel n'est pas brevetable en tant que tel
- Sont brevetables les **inventions mises en œuvre par ordinateur**
 - Sous réserve du caractère technique de l'invention = dès lors qu'il y a « effet technique supplémentaire »
 - Si l'invention répond aussi aux critères de nouveauté et d'activité inventive

III. ENJEUX ET OUTILS SPÉCIFIQUES ?

1. Gestion
2. Diffusion et Préservation
3. Valorisation

III. ENJEUX ET OUTILS SPÉCIFIQUES ?

1. Gestion

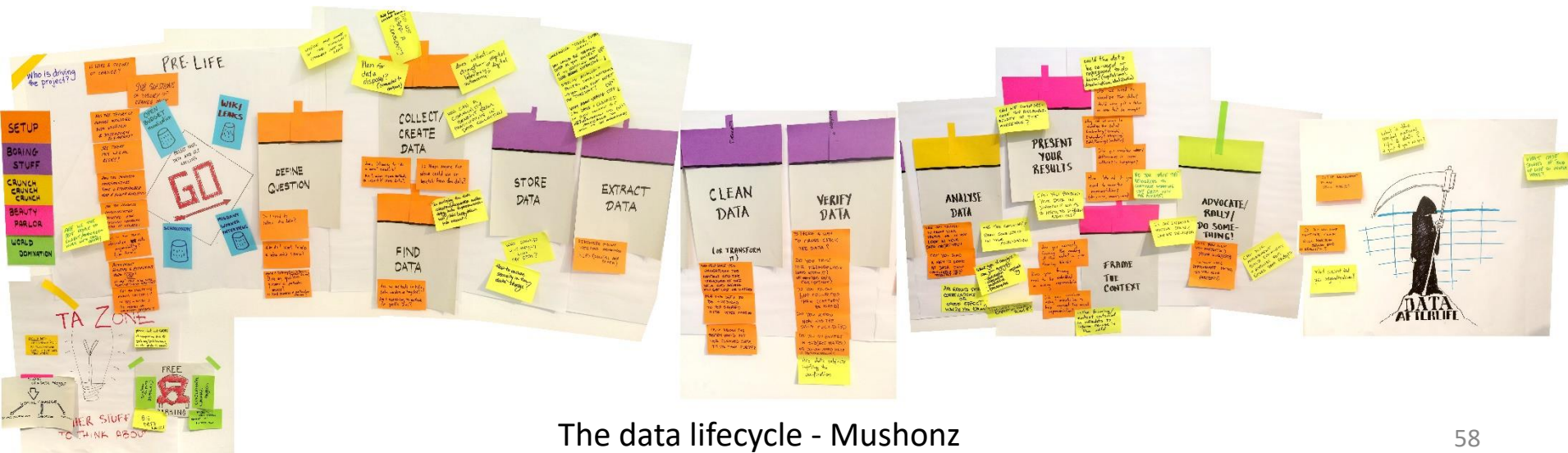
Le plan de gestion de données (PGD) Data management plan (DMP)

Un document formalisé

Il décrit la façon dont les données produites seront obtenues, documentées, utilisées, gérées et disséminées, pendant la phase de recherche et une fois le projet terminé.

Un document évolutif

- À compléter au fur et à mesure du projet
- À envoyer aux financeurs à différents stades du projet



The data lifecycle - Mushonz

Principales questions d'un PGD

- Comment la gestion et le partage des données sont-ils financés, en particulier à long terme ?

Ressources

- En quoi consiste le projet ?
- Qui sont les partenaires ?
- Quelle est la politique de gestion des données ?
- Qui est responsable de la gestion des données ?

Responsabilités dans le projet

- Quelles données seront produites/utilisées au cours du projet ? (type, format, volume et accroissement...).
- Comment seront-elles produites ou transformées ?

Collecte des données

- Comment les données seront-elles identifiées, décrites ?
- Quels standards de métadonnées utilisera-t-on ?
- Comment seront générées les métadonnées ?

Documentation des données

What's the plan?

D'après S. Cocard et D. L'Hostis,
« Pourquoi et comment rédiger un plan de gestion de données ? », 2019 ([en ligne](#)).

- Qui pourra accéder aux données ?
- Les données seront-elles publiées ?
- Comment ?
- Dans quel délai ?
- Sous quelle licence ?

Accès et partage des données

- Comment, où, par qui, seront stockées, sauvegardées et sécurisées les données ?

Sauvegarde des données

- Qui sera propriétaire des données produites ?
- Des données externes seront-elles utilisées ?

Propriété intellectuelle

- Des données sensibles seront-elles produites ou utilisées ?
- Comment sera assurée leur anonymisation ?

Éthique

- Quel est le plan d'archivage et de préservation à long terme ?

Archivage et préservation des données

Modèles de PGD

Sylvie Cocard (INRAE), [Rédiger un plan de gestion de données](#), 2022

Verburg, M., & Grootveld, M. (2022, février 22). **FAIR-Aware Additional guidance to the Science Europe DMP assessment rubric.**

<https://doi.org/10.5281/zenodo.6088215>



- Informations sur le PGD
- Informations sur le projet ou la structure
- Présentation générale des données
- Organisation et documentation des données
- Droits de propriété intellectuelle
- Confidentialité
- Stockage et sécurité des données
- Partage des données à l'issue du projet
- Archivage et conservation des données après la fin du projet

[Trame projet v2](#)
43 questions
[Trame structure](#)
39 questions



- Description des données et collecte ou réutilisation de données existantes
- Documentation et qualité des données
- Stockage et sauvegarde pendant le processus de recherche
- Exigences légales et éthiques, codes de conduite
- Partage des données et conservation à long terme
- Responsabilités et ressources en matière de gestion des données

[ANR Modèle de PGD](#)
15 questions
[Guide en français](#)



- Résumé descriptif des données
- Données FAIR
- Allocation de ressources
- Sécurité des données
- Aspects éthiques
- Autres



- Data Summary
- FAIR data
- Other research outputs
- Allocation of resources
- Data security
- Ethics
- Other issues

[H2020 FAIR DMP \(v.f.\)](#)
9 questions
[Horizon Europe Data](#)

this template is recommended
for Horizon Europe beneficiaries



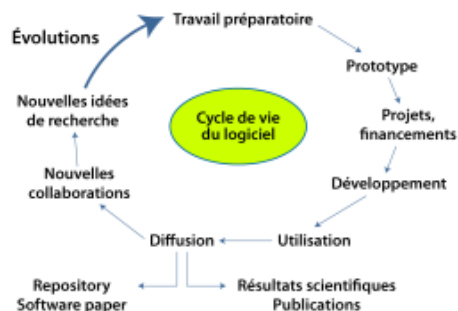
- Description des données et collecte ou réutilisation de données existantes
- Documentation et qualité des données
- Exigences légales et éthiques, codes de conduite
- Traitement et analyse des données
- Stockage et sauvegarde pendant le processus de recherche
- Partage des données et conservation à long terme

[Science Europe](#)
« [modèle structuré](#) »
(+ [version anglaise](#)) 21

Software management plan (SMP)

Presoft : Preservation for research software

Cycle de vie du logiciel de la recherche



Plan de gestion du logiciel ou SMP

- Nouvel outil pour la pérennisation du logiciel
- Feuille de route pour gérer le cycle de vie du logiciel
- Doit être adapté au logiciel
- Doit être adapté aux besoins de l'équipe et des chercheurs
- Outil de réflexion, de planification
- Permet d'identifier les étapes, les acteurs, les responsabilités
- Outil pour mieux gérer la production scientifique
- À ne pas confondre avec un plan de développement

Bénéfices du SMP

- Une vue d'ensemble de chaque logiciel
- Un suivi facile dans une équipe
- L'amélioration des méthodes (gain de temps)
- L'identification de ce qui marche bien
- La connaissance des productions scientifiques
- Une meilleure visibilité des productions importantes
- La mise en œuvre d'une politique d'équipe, de laboratoire, d'institution...

Procédure

- Les étapes suivent le cycle de vie du logiciel
- Mise à jour du SMP à chaque étape
- Renseigner les nouvelles informations
- Renseigner les nouvelles actions à entreprendre
- Niveau de détail à définir en fonction du besoin

Modèle PRESOFT de SMP



Références

- Article vs. Logiciel : questions juridiques et de politique scientifique dans la production de logiciels, T. Gomez-Diaz, 1024 – Bulletin de la société informatique de France, numéro 5, mars 2015, pp. 119–140
- Free software, Open source software, licenses. A short presentation including a procedure for research software and data dissemination, T. Gomez-Diaz, septembre 2014, présenté à la Conférence EGI, mai 2015
- Research Software Sustainability: Report on Knowledge Exchange workshop, février 2016
- The Software Sustainability Institute. Checklist for a Software Management Plan, 2016



Journées Développement
Marseille, 4-7 juillet 2017

Mots clés : logiciel de la recherche, pérennisation, préservation, accès libre, reproductibilité, science ouverte.



Figure 1. Software Management Plan requirements grouped by their focus.

Netherlands eScience center, NWO,
DRC, [Practical guide to SMP](#) (2022)

Software Management Plan

Quick definitions

- Software Management Plan: refers to activities during the project
- Software Sustainability Plan: refers to activities beyond the end of the project
- Software Stewardship Plan: refers to the entire lifecycle

Software stewardship is the management and oversight of an organization's software assets to help provide business users with high-quality software that is easily accessible in a consistent manner.

The SMP is NOT for editors.

One+ line of code is software.

Survey / List of Questions

Where can the source code be found? Or how can it be obtained?

1. Documentation			
1) What type of documentation is available for the software?	checklist		
	User documentation		

[Elixir SMP](#)

5.1. Forges externes commerciales

On considère ici les forges telles que github.com, gitlab.com, bitbucket.org, sr.ht, etc.

Forces	Faiblesses
• Intégration des fonctionnalités « tout en un » • Gratuité d'un ensemble de fonctionnalités de base • Disponibilité acceptable pour la plupart des projets • Pour GitHub : une ergonomie devenue familière • Pour sr.ht : un support humain très réactif • Visibilité internationale • Communauté internationale	• Dépendance à l'égard de la politique commerciale de la société • Pas de contrôle sur les ressources allouées aux fonctionnalités • Pas de possibilité d'être associé aux décisions stratégiques • Non souveraineté • Pas de garantie de pérennité
Opportunités	Menaces
• Nombreuses ressources disponibles pour la formation à ces outils	• Dépendance à la réglementation du pays où siègent ces sociétés

Voir aussi les SWOT

5.2 Forges externes communautaires

(Apache, Eclipse, OW2 et FSF)

5.4 Forge auto-hébergée à l'échelon national

5.3. Forges auto-hébergées à l'échelon local

Forces	Faiblesses
• Fonctionnalités sur mesure • Disponibilité sur mesure • Souveraineté • Résilience • Support (dépend des instances) • Pérennité (dépend des instances) • (Proximité) • Visibilité au niveau de l'institution • Communauté au niveau de l'institution	• Disponibilités (dépend des instances) • Support (dépend des instances) • Multiplicité des solutions disponibles • Difficulté d'accès en dehors de l'institution porteuse
Opportunités	Menaces
• De facto, catalogue institutionnel de logiciels • Mise en œuvre d'une politique institutionnelle de développement des logiciels de recherche • Expertise maintenue au niveau de l'institution • Diffusion de bonnes pratiques	• Evolution de la solution choisie (périmètre des fonctionnalités sous licence libre vs fonctionnalités devenant payantes) • Expertise difficile à trouver nécessaire pour maintenir la solution au niveau de l'institution

III. ENJEUX ET OUTILS SPÉCIFIQUES ?

2. Diffusion et Préservation

État des lieux en France

Explorez les données du [Baromètre science ouverte 2023](#) et répondez aux questions suivantes

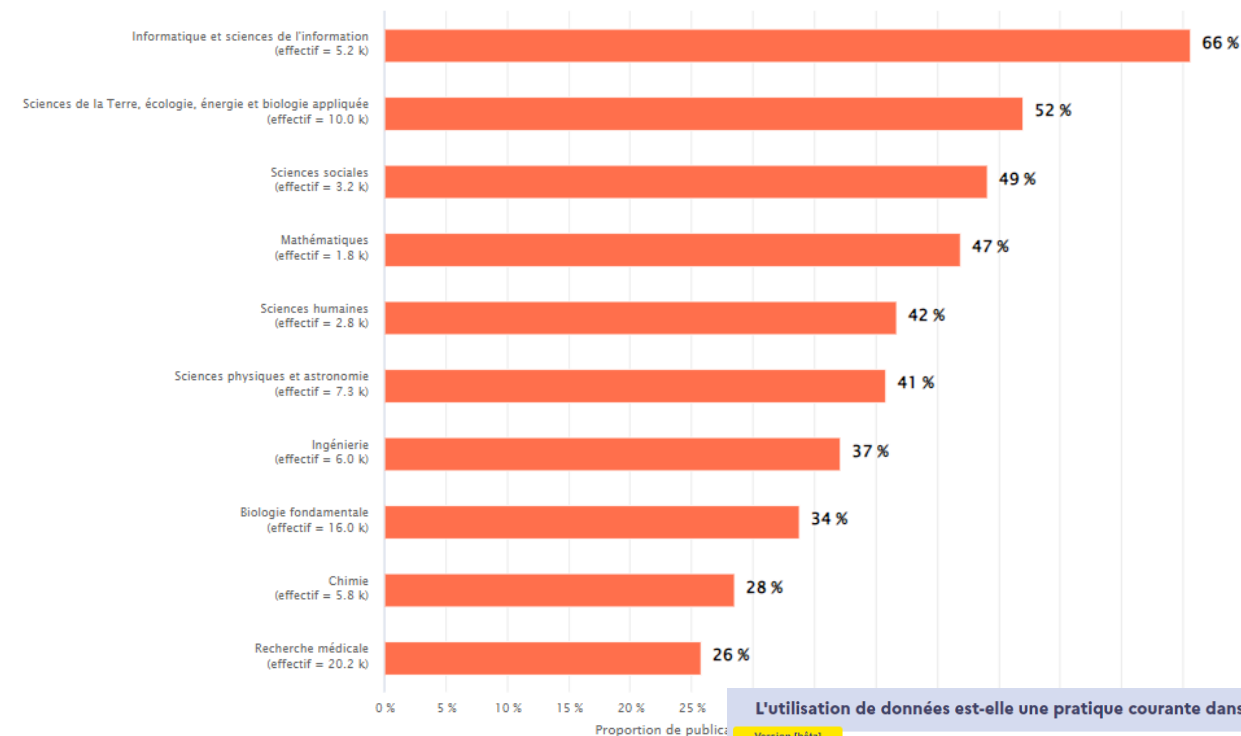
- Dans quelles disciplines produit-on le plus de données ?
- Quelle est la proportion des publications indiquant la disponibilité des données (*Data Availability Statement*) ? Dans quelles disciplines surtout ?
- En 2022, combien de publications mentionnent l'utilisation de codes ou logiciels ? Et la création ?



Proportion de publications françaises qui mentionnent avoir produit leurs données par discipline

Trier par :

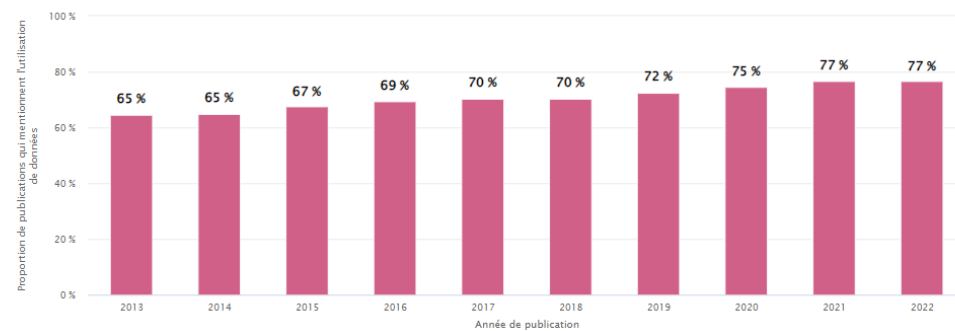
☐ Plus grand effectif ☒ Plus fort taux de création



L'utilisation de données est-elle une pratique courante dans les publications en France ?

Version [bêta]

Proportion de publications françaises qui mentionnent l'utilisation de données par année de publication

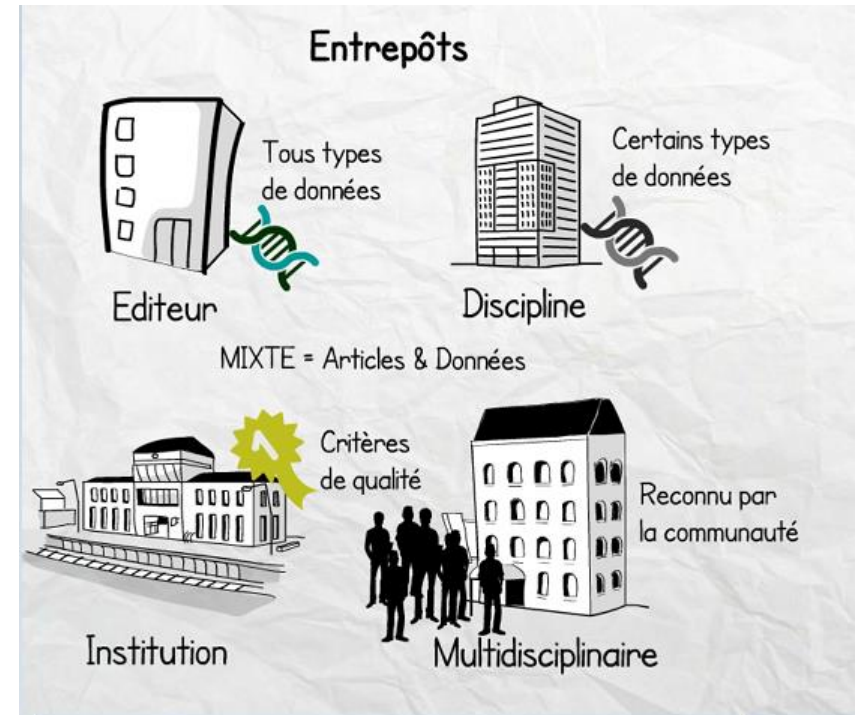


Commentaire

Ce graphique montre, par année de publication, la proportion de publications pour lesquelles une mention d'utilisation de données a été détectée parmi les publications analysées.

Les entrepôt de données

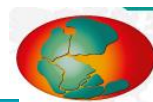
« Les entrepôts de données sont des **plateformes web** sur lesquelles les chercheurs peuvent **déposer et rechercher des données scientifiques**. ils proposent des **services spécifiques à l'activité de recherche** (description des jeux de données, choix des conditions d'accès, attribution de licence...) et offrent une garantie de **sécurité**. La plupart présente un système de recherche détaillé avec de nombreux filtres. »



Les entrepôts de données de recherche, [Cours Doranum sur Callisto](#), 2021

multidisciplinaires

disciplinaires



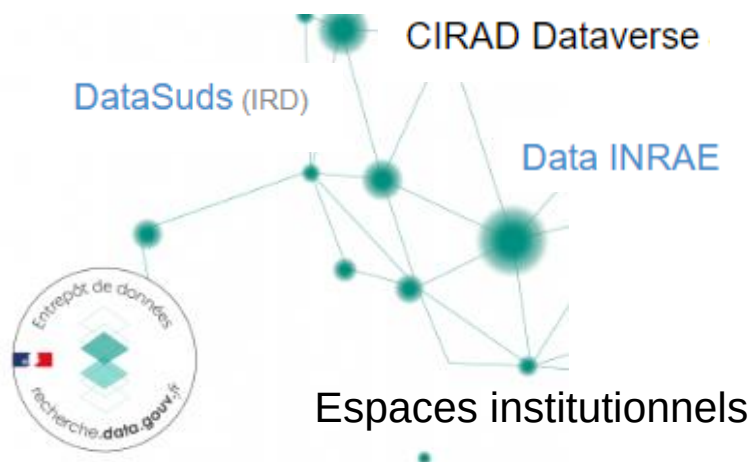
PANGAEA.

Data Publisher for Earth & Environmental Science



International

France



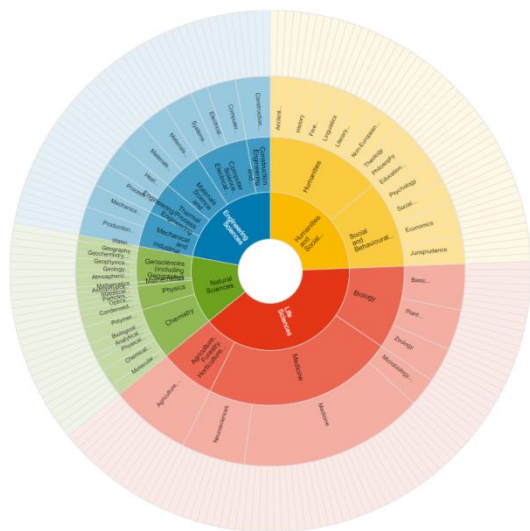
data.sciencespo



Annuaires et listes d'entrepôts



re3data



OAD Data repositories

Contents [hide]	
1	Archaeology
2	Astronomy
3	Biology
4	Chemistry
5	Computer Science
6	Energy
7	Engineering
8	Environmental sciences
9	Geology
10	Geosciences and geospatial data
11	Linguistics
12	Marine sciences
13	Medicine



Core certified repositories

Search markers	
	CLARIN-LV repository https://repository.clarin.lv/ CoreTrustSeal certification 2020-2022 (valid till 16 January 2023) Raina bulvāris 29, Central District, Riga, Latvia
	Ocean Biology Distributed Active Archive Center https://oceancolor.gsfc.nasa.gov/ CoreTrustSeal certification 2020-2022 (valid till 16 December 2022) NASA Goddard Space Flight Center, Greenbelt Road, Greenbelt, MD, USA
	The Digital Archaeological Record (tDAR) https://www.tdar.org/

Les critères de choix d'un entrepôt

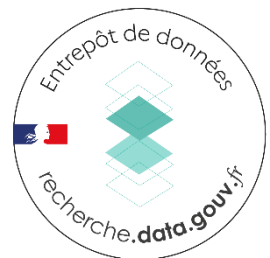
- Dans mon périmètre disciplinaire ?
Recommandé par mes collègues ? Les financeurs ? Mon employeur ?
- Accepte mes fichiers (format, taille...) ?
- Accepte mes métadonnées ? Fournit un identifiant pérenne ?
- Permet de modifier mes données et accéder aux anciennes versions ?
- Propose licences de diffusion ? Permet de restreindre l'accès ?
- Procédure de validation du dépôt ?
Coûts éventuels associés ?
- Certifié / reconnu ? [Core Trust Seal](#) ?
[Principes TRUST](#) ?
- Voir S.-A. Sansone et al., *Data Repository Selection: Criteria That Matter*, Zenodo ([en ligne](#)).
- [Sélectionner un entrepôt thématique de confiance](#), note méthodologique du COSO, 2023

Autant que possible, privilégier un entrepôt (par ordre de préférence) :

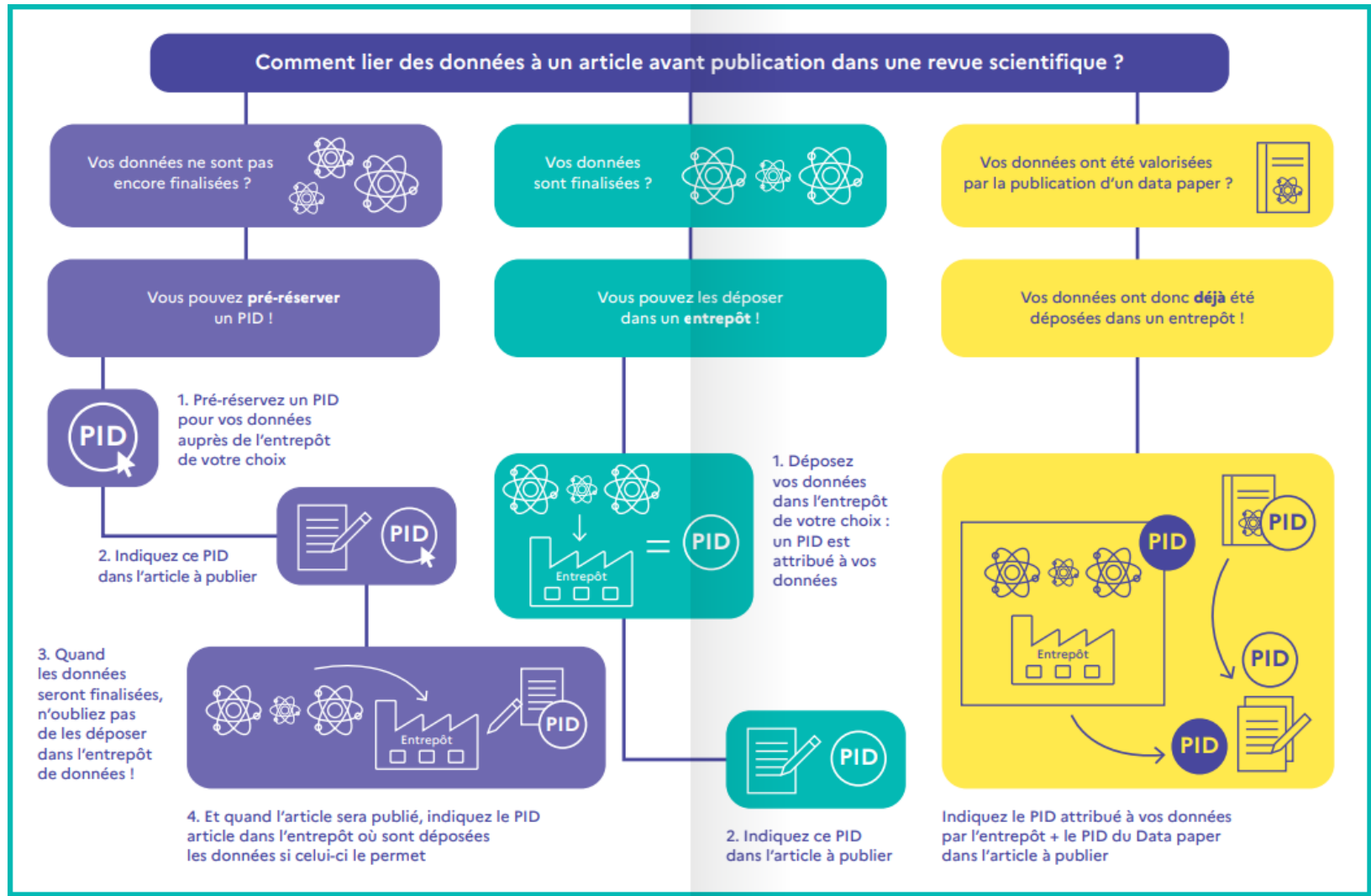
1. Spécialisé dans la gestion de son type spécifique de données
2. Spécialisé dans sa discipline
3. Institutionnel
4. Généraliste

➤ [Sélectionner un entrepôt thématique de confiance, COSO, 2024](#)

➤ [Logigramme Recherche Data gouv](#)



Lier article et données



[Partager les données liées aux publications scientifiques,](#)
Guide pour les chercheurs, Ouvrir la science, 2022

Lier article et code



Exemple : [ORE](#)

Voir aussi « [Journal Production Guidance for Software and Data Citations](#) », Scientific Data, 2023

Research and Innovation

Open Research Europe

Search

SUBMIT YOUR RESEARCH

Browse Gateways & Collections How to Publish About Resource Hub Blog

Sign in

Home > [For Authors](#) > [Open Data, Software and Code Guidelines](#)

Submit your Research

My Submissions

Article Guidelines

Article Guidelines (New Versions)

Open Data, Software and Code Guidelines

Open Data, Software and Code

- [What is required when submitting an article](#)
- [Further guidance](#)
- [Limited exceptions to these guidelines](#)
- [The FAIR Data Principles](#)
- [Open Research Europe-approved repositories](#)

[Programme Guide](#)

What is required when submitting an article

1. [Your dataset\(s\) must be deposited in an appropriate data repository.](#)
2. [Your dataset\(s\) must have a license applied which allows reuse by others \(CC0 or CC-BY\).](#)
3. [Your dataset\(s\) must have a persistent identifier \(e.g. a DOI\), allocated by a data repository.](#)
4. [You must provide a data availability statement as a section at the end of your article, including elements 1-3.](#)
5. [You must include a data citation and add a reference to data to your reference list.](#)
6. [Your dataset\(s\) should not contain any sensitive information, for example in relation to human research participants.](#)
7. [You should share any related software and code.](#)
8. [Your dataset\(s\) must be useful and reusable by others, adhere to any relevant data sharing standards in your discipline and align with the FAIR Data Principles.](#)
9. [Your dataset\(s\) should link back to your article, if possible.](#)

If you fail to adhere to these guidelines when submitting, the publication of your article may be delayed, and your article may ultimately be rejected.

7. You should share any related software and code.

All articles should include details of any software and code that are required to view the datasets described or to replicate the analysis.

For software

For all software used, please state the version, details of where the software can be accessed, and any variable parameters that could impact the outcome of the results. If you have coded software in-house, the source code should be written in (or be compatible with) an Open Source programming language, and should be archived under an open license and shared. For code stored in GitHub, you should [create a 'public registration'](#) for your project to obtain a DOI.

Information about software should be included in a software availability statement, which you can add to the end of your article, before the references list.

When drafting the statement, please include:

- Software available from: URL for the website where software can be downloaded from, if applicable.
- Source code available from: URL for versioning control system (for example GitHub).
- Archived source code at time of publication: DOI and citation for project in Zenodo (please select the appropriate DOI for the version which underlies your article).
- License: Must be an open license and preferably an [OSI-approved license](#).

Where third-party proprietary software has been used, a non-proprietary, Open Source alternative software should be suggested by the author to allow for the replication of the analysis or research by all readers. We recognize that there may be cases where this may not be feasible. Please see the [limited exceptions to these guidelines](#) for more information.

Additionally, if there are ethical or privacy considerations as to why the source code may not be made available, please contact the [editorial team](#).

For analysis code

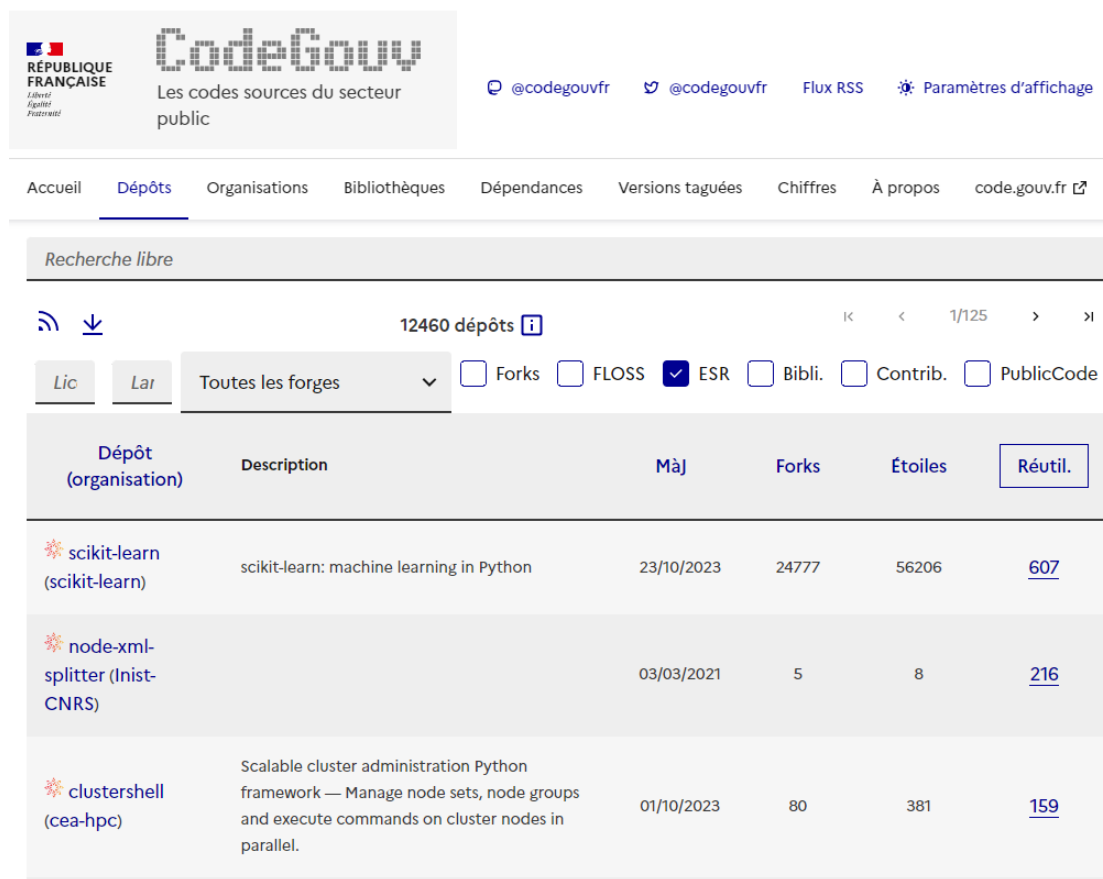
If you have created custom analysis code, this should be archived under an open license and shared. For analysis code stored in GitHub, you should [create a 'public registration'](#) for your project to obtain a DOI. We recommend using an [OSI-approved license](#), but CC-BY 4.0 is also acceptable.

Préserver et signaler les codes

- Pour assurer leur description
- Pour faciliter leur recherche
- Pour permettre leur citation
- Pour les valoriser

Dépôts ESR dans

<https://code.gouv.fr/public/#/repos>



The screenshot shows the CodeGouv website interface. At the top, there's a header with the French Republic logo, the CodeGouv logo, and the tagline "Les codes sources du secteur public". Navigation links include "Accueil", "Dépôts", "Organisations", "Bibliothèques", "Dépendances", "Versions taguées", "Chiffres", "À propos", and "code.gouv.fr". A search bar labeled "Recherche libre" is present. Below the search bar, there are filters for "Lic", "Lar", "Toutes les forges", and checkboxes for "Forks", "FLOSS", "ESR" (checked), "Bibli.", "Contrib.", and "PublicCode". A table lists repositories with columns: "Dépôt (organisation)", "Description", "Màj", "Forks", "Étoiles", and "Réutil.". The table contains three entries: "scikit-learn (scikit-learn)", "node-xml-splitter (Inist-CNRS)", and "clustershell (cea-hpc)".

Dépôt (organisation)	Description	Màj	Forks	Étoiles	Réutil.
scikit-learn (scikit-learn)	scikit-learn: machine learning in Python	23/10/2023	24777	56206	607
node-xml-splitter (Inist-CNRS)		03/03/2021	5	8	216
clustershell (cea-hpc)	Scalable cluster administration Python framework — Manage node sets, node groups and execute commands on cluster nodes in parallel.	01/10/2023	80	381	159

Déposer le code source d'un logiciel

Une collaboration étroite entre Hal-Inria (l'archive ouverte d'Inria), le CCSD et Software Heritage a abouti à la mise en place d'un processus permettant de déposer des codes sources de logiciels via HAL, et leur archivage pérenne dans Software Heritage. Le dépôt dans HAL permet la citabilité, l'archivage étant pris en charge par Software Heritage.

Pour être transféré à Software Heritage, le fichier déposé doit être sous licence libre et ne peut pas être sous embargo.

Quel logiciel déposer dans HAL/Software Heritage ?

La fonctionnalité de dépôt actuelle permet de déposer seulement une archive aux formats zip ou tar.gz. On ne peut donc pas archiver tout un historique de développement, mais seulement un répertoire contenant des fichiers sources à un moment donné. L'usage de cette fonctionnalité est pour l'instant déconseillé aux chercheurs qui utilisent des systèmes de contrôle de version et qui souhaitent archiver aussi l'historique de développement.

Préparation du code source

Avant de commencer le dépôt, préparer le code source du logiciel en ajoutant les fichiers suivants:

- **README** : décrit le logiciel que vous déposez, voir aussi Make a README
- **AUTHORS** : contient la liste des auteurs et des contributeurs éventuels
- **LICENSE** : décrit les droits d'utilisations du code source déposé (à choisir en collaboration avec les services de valorisation des organismes de rattachement des auteurs). voir aussi la section Choisir une licence pour un code source de logiciel dans la documentation.

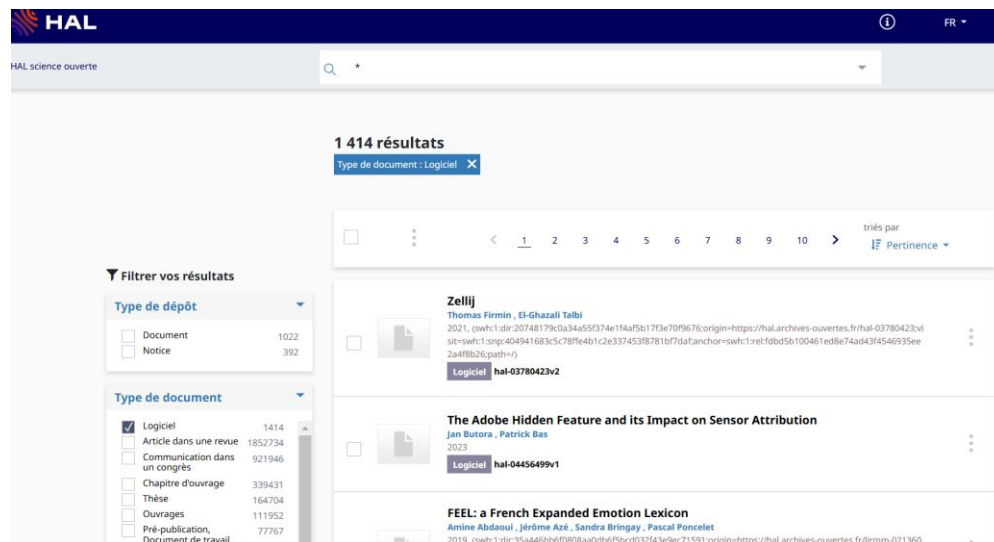
Puis créer une archive .zip et nommer l'archive avec le nom du logiciel et sa version.

Le formulaire de dépôt sur HAL

Cliquer sur l'onglet Dépôt.

<https://doc.archives-ouvertes.fr/deposer/deposer-le-code-source/>

[Fiche pratique](#)



The screenshot shows the HAL search interface. At the top, the HAL logo and navigation links are visible. The search bar contains the text 'HAL science ouverte'. Below the search bar, the results are displayed as a list of items. The first item is 'Zellij' by Thomas Firmin, El-Ghazali Talbi, with a HAL ID of hal-03780423v1. The second item is 'The Adobe Hidden Feature and its Impact on Sensor Attribution' by Jan Butora, Patrick Bas, with a HAL ID of hal-04456499v1. The third item is 'FEEL: a French Expanded Emotion Lexicon' by Amine Abdou, Jérôme Aze, Sandra Brinay, Pascal Poncelet, with a HAL ID of hal-04456499v1. The results are filtered by 'Type de document : Logiciel' and sorted by 'Pertinence'.

Software Heritage

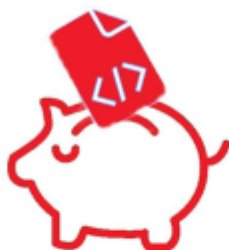


Modalités d'alimentation de la bibliothèque



Des opérations de sauvetage

[En savoir + sur le sauvetage des contenus Gitorious](#)



Des partenaires exportent leurs données



Du moissonnage automatisé



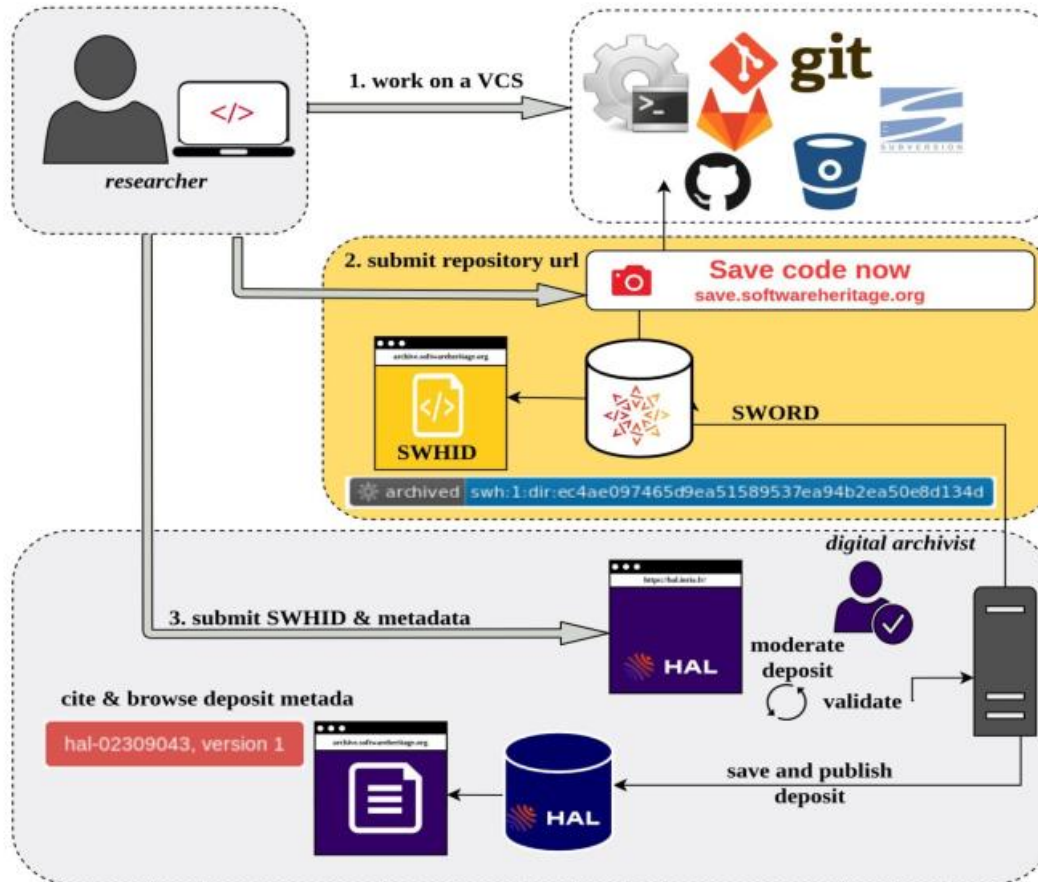
Voir si la forge de votre institution est archivée



De l'archivage sur mesure :

- Un projet
 - Une collection de projets (forge)
- Sans création de compte
Service accessible à tous,
pas besoin d'être l'auteur du code*

Connexion HAL / Software Heritage



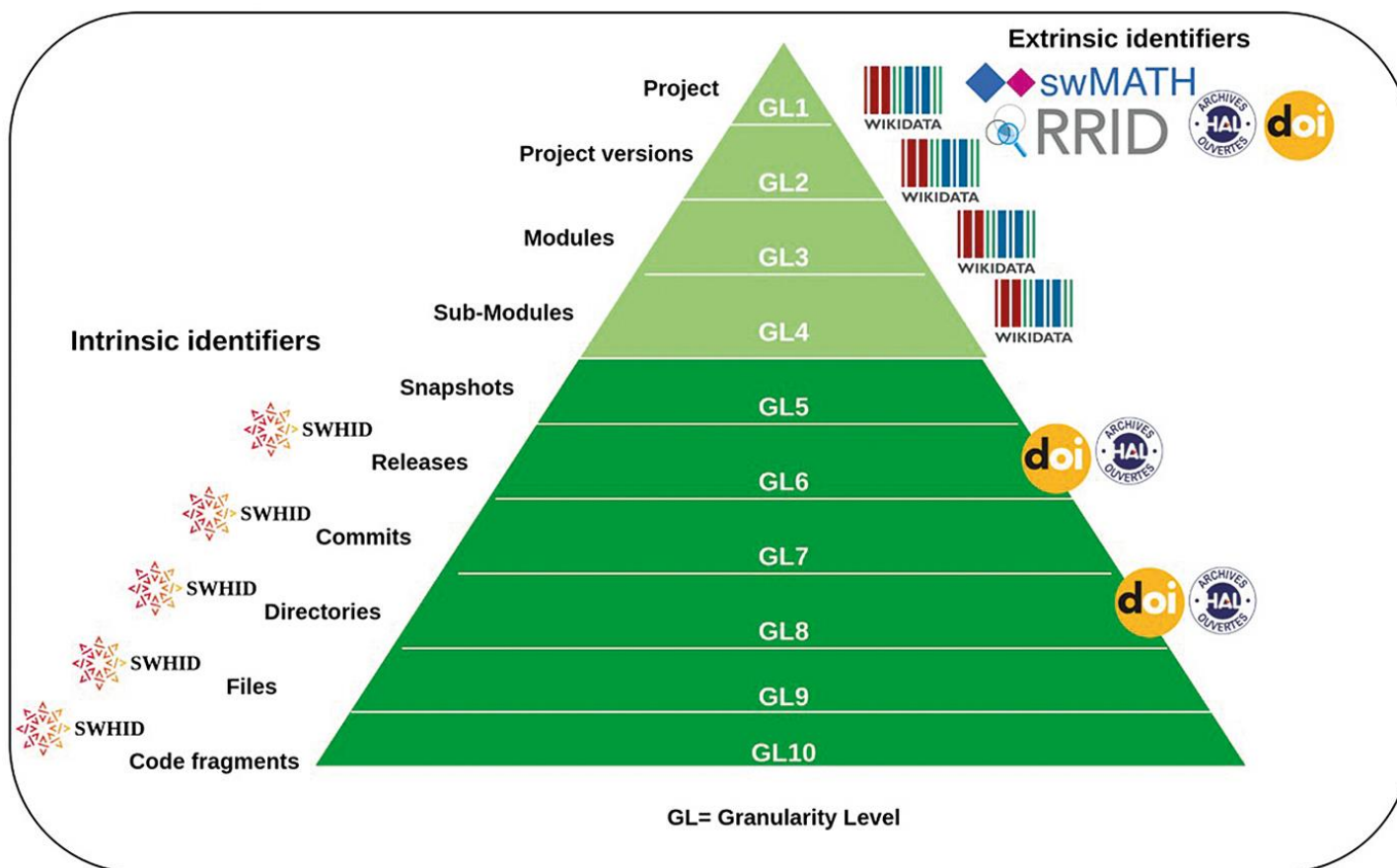
[Référencer les logiciels de la recherche](#),
webinaire Cellule Data
UGA, 2023

[Guide du déposant](#)

[Guide du modérateur
HAL](#)

[Extension de
navigateur SWH](#)

Un écosystème de référencement

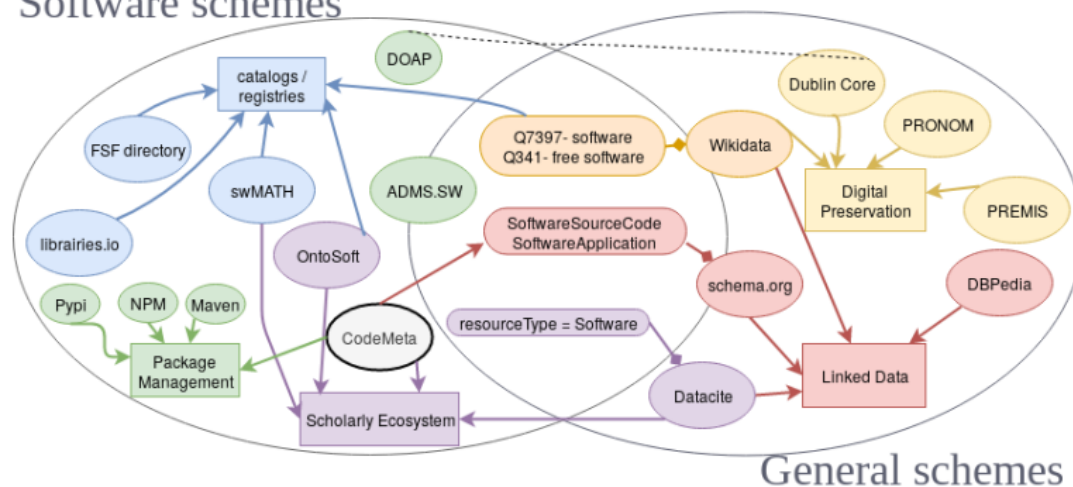


Niveaux de granularité et identifiants
Interview [« Garantir la cohérence des données
constitue le cœur de notre activité », BBF 2023](#)

Les métadonnées des codes et logiciels

Le millefeuille des vocabulaires pour décrire le logiciel

Software schemes



“As yet, there is no single ontology or metadata vocabulary for software description, or indeed agreed-upon taxonomies for these and other possible sets of criteria.”

Morrissey, S. M. (2020). *Preserving Software: Motivations, Challenges and Approaches*. Digital Preservation Coalition.
<https://doi.org/10.7207/twgn20-02>

[CodeMeta](#), un pivot entre les vocabulaires

MédiaNormandie | 2023-05-12 | S. Granger | CC-BY-SA

34

Software Heritage, Sabrina Granger ([Formation ADBU-MediaNormandie](#), 2023)

Voir aussi <https://fair-impact.eu/events/fairimpact-events/developing-guidelines-metadata-collection-and-curation-research-software>

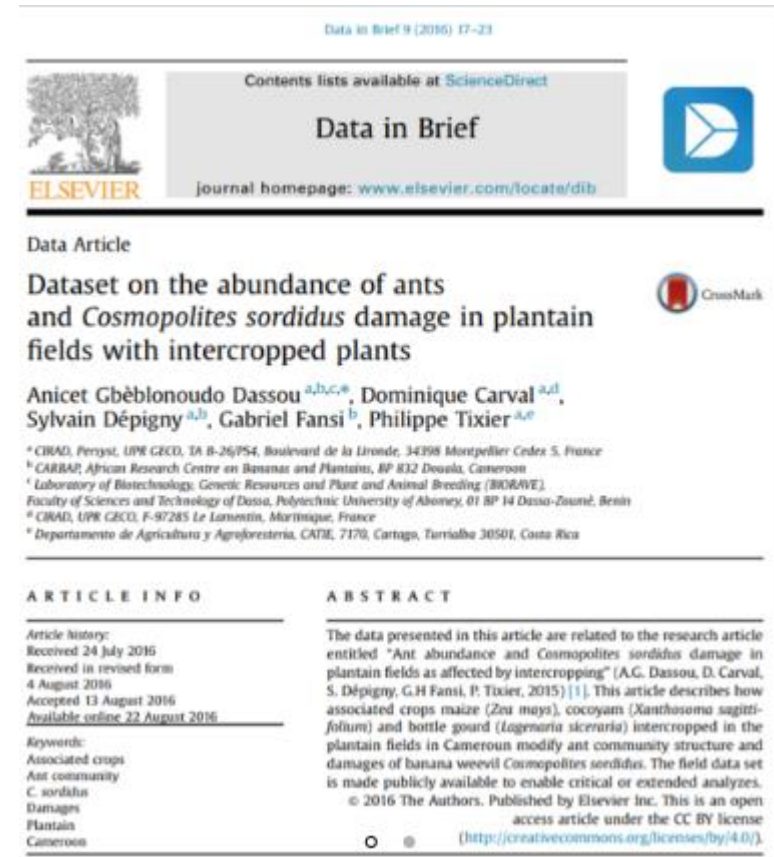
III. ENJEUX ET OUTILS SPÉCIFIQUES ?

3. Valorisation

Data papers et software papers

De nouveaux types de publications scientifiques, décrivant les données ou les logiciels plutôt que les résultats de la recherche et destinés à en faciliter la réutilisation ([Fiche Data papers](#) SOCLE ; [Présentation Software papers](#), Open Science Pasteur)

- Liste de revues publiant des Software papers [Software Sustainability Institute](#)
- Listes de revues publiant des Data papers : [DATAACC](#) ; [CoopIST-Cirad](#)

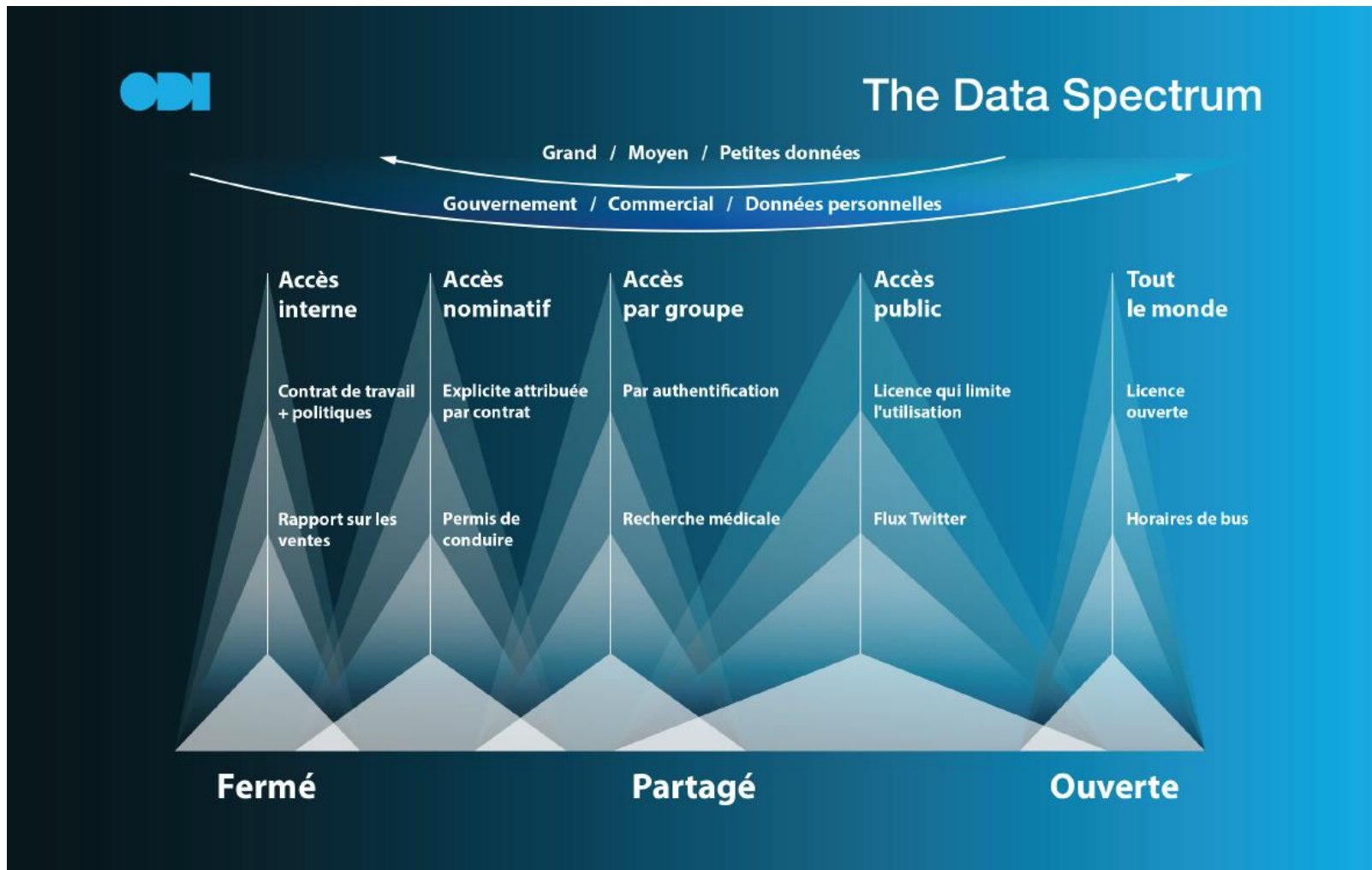


De nouvelles données à publier



Schéma adapté de Report on integration of data and publications. Opportunities for Data Exchange (Reilly S. et al., 2011) - [Tutoriel Form@doct, 2016](#)

Aussi ouvert que possible...



... aussi fermé que nécessaire ?

Clarifier : quels degrés d'ouverture pour les codes sources ?

Nous proposons de distinguer les quatre degrés d'ouverture suivants :

-  **Niveau A - contributif** : Le code source est publié, les contributions extérieures sont activement recherchées et traitées.
-  **Niveau B - ouvert** : Le code source est publié, les contributions extérieures sont traitées mais non activement recherchées.
-  **Niveau C - publié** : Le code source est publié mais les contributions extérieures ne sont pas traitées.
-  **Niveau D - non-communicable** : Le code source n'est pas communicable au public.

<https://guides.etalab.gouv.fr/logiciels/#a-quoi-sert-il>

Licences de diffusion et de réutilisation

Données



- **Etalab** : paternité
([Équivalente à CC-BY 2.0](#))



- **ODC-ODbL** : Open Database License : partage, réutilisation et adaptation sous condition d'attribution et de licence à l'identique

+ licences homologuées

Voir [Licences de réutilisation autorisées – Etalab](#)

Choix de licences : [Licentia by Inria](#)

Logiciels

- Copyleft fort, imposant les mêmes termes (GNU GPL , CeCILL-A...)
- Copyleft faible, pour combiner différentes licences (GNU LGPL, CDDL, Mozilla MPL...)
- Non copyleft, sans obligation de diffusion dans les mêmes termes (MIT, Apache, BSD, CeCILL-B)

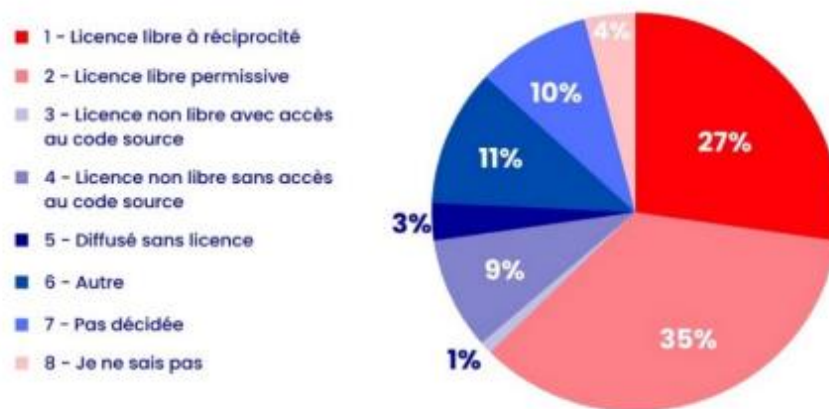
Voir [Livret Bleu du Logiciel Libre – Fondamentaux Juridiques](#), 2016

Logiciel libre

Logiciel dont le code source est accompagné d'une licence permettant 4 libertés :

- Exécuter le programme
- Redistribuer des copies
- Étudier le programme et l'adapter à ses besoins
- Apporter des améliorations au programme et les publier

Diffusion des logiciels de recherche LICENCES CHOISIES



62% des logiciels sont sous licence libre
10 % des logiciels recensés sont diffusés avec une licence privative, ce qui confirme le lien fort entre le logiciel libre et la recherche

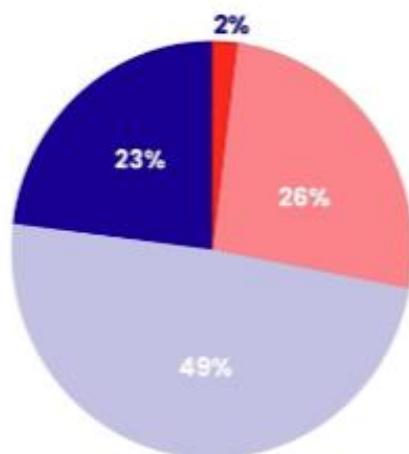
L'accès au code source favorise la reproductibilité des résultats de la recherche par d'autres équipes

Etat des lieux de la production et de la valorisation des logiciels issus de la recherche publique (novembre 2023)

Modalités de valorisation

Publication scientifique (72 % des cas)
Services autour du logiciel (23% des cas)

**PART DES LOGICIELS AYANT FAIT L'OBJET
DE PUBLICATION SCIENTIFIQUE**



■ Je ne sais pas ■ Non ■ Oui, plusieurs publications ■ Oui, une seule publication

Source : Enquête nationale, Juin 2019 (ARIST) - Citeris SARL

Différentes modalités de valorisation cohabitent.

La **valorisation académique** représente la première forme de valorisation, 72% des logiciels issus des travaux de recherche font l'objet d'une publication scientifique.

Concernant la **valorisation économique** :

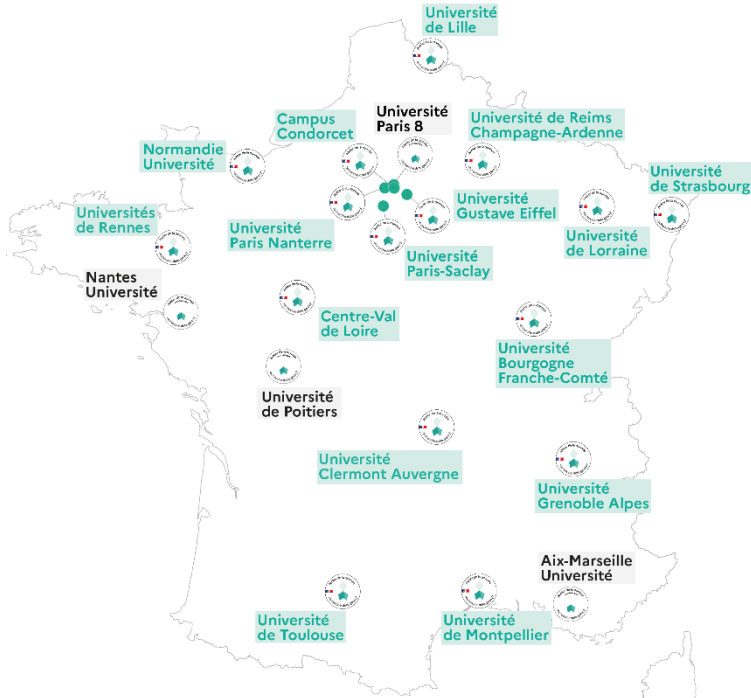
- La diffusion sous licence privative, observée dans 10% des cas constitue la voie la plus directe de valorisation ;
- Le développement de services autour du logiciel, dont le cœur scientifique reste ouvert et accessible sous la forme d'une licence libre est observé dans 23% des cas. Il s'agit de services de formation, de documentation, d'intégration, de développement à façon de modules ou de versions spécifiques destinées à la commercialisation et protégées par des licences propriétaires en réponse aux besoins spécifiques d'un partenaire industriel, voire co-développés avec des entreprises.

Les logiciels de recherche conduisent également à la création de start-up dans 4% des cas.

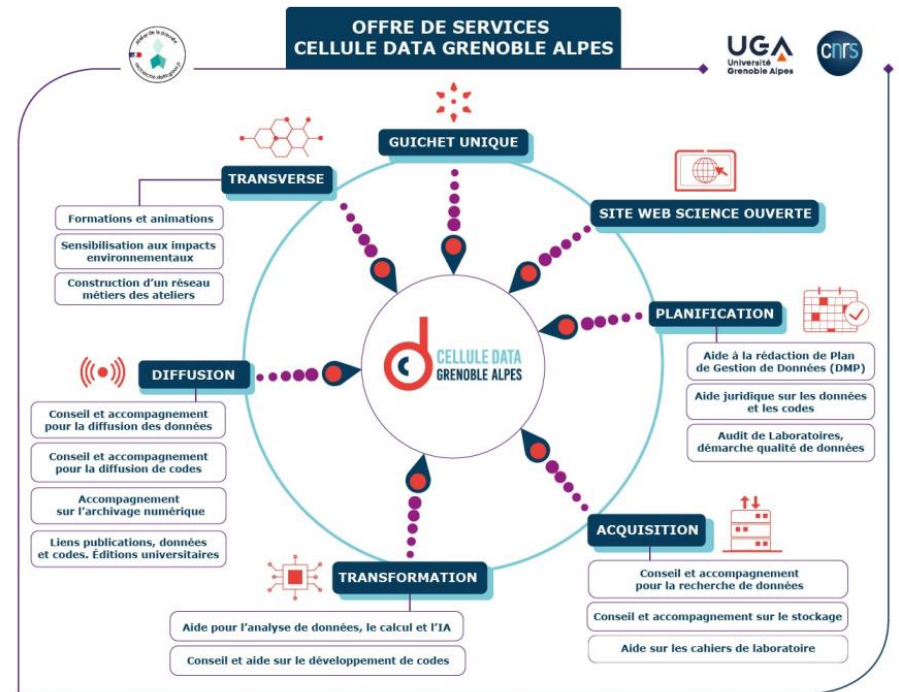
Ces différentes modalités de valorisation économique à l'œuvre illustrent qu'il n'y a **pas de contradiction entre licence libre, valorisation économique et pratiques de partage des chercheurs.**

Données	Tout type de données produites dans le cadre de la recherche, définies comme enregistrements factuels qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche.	Données tabulaires, images, sons, vidéos, données 3D, séquençage, données d'observation, données d'instruments, bases de données relationnelles, textes, annotations, etc.	Principe « aussi ouvert que possible, aussi fermé que nécessaire », principes FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable)	https://zenodo.org/ https://www.humanum.fr/consortiums
Comité pour la science ouverte, <i>Types de documents, productions et activités valorisées par la science ouverte et éligibles à une évaluation</i>, décembre 2019 (en ligne).				
Plans de gestion de données	Plans élaborés dans le cadre de projets de recherche, incluant les formats, la description des données et les règles de partage.		Renseignement et dépôt dans une plateforme spécialisée	https://dmp.opidor.fr/public_plans
Logiciels	Tout type de code et d'interface	Distinguer (1) les services accessibles en ligne (ex. application web), (2) les logiciels téléchargeables-réutilisables (mais dont le code n'est pas ouvert), (3) les logiciels libres et <i>open-source</i> . En qualifier la réutilisabilité : documentation, tutoriel communauté d'utilisateur, communauté de développement.	Dépôt dans une archive ouverte adéquate gérant les versions et moissonnée par Software Heritage	www.softwareheritage.org
Protocoles de recherches	Description de méthodes, de démarche de recherche	Design d'essais cliniques	Dépôt dans une archive ouverte adéquate	https://www.protocols.io/ https://cos.io/rr/ https://clinicaltrials.gov/

Conclusion



Réseau des Ateliers de la donnée



Retour d'expérience –
Cellule Data Grenoble Alpes

([Formation ADBU-MediaNormandie](#), 2023)

[Ambassadeurs SWH](#)

[Réseau Devlog](#)

Ressources généralistes / Données

- [Dorandum](#)
 - [Parcours interactif sur les données de la recherche](#) (juin 2020, mis à jour 2023)
- [Guide de bonnes pratiques sur la gestion des données de la recherche](#), Atelier Données CNRS (2023, v2)
- Espace Callisto du [Centre de ressources Compétences – Recherche Data Gouv](#) (2022-)
- [Passeport Données de la recherche](#) (2024)
- C. Borgman, [*Qu'est-ce que le travail scientifique des données ?*](#), (2015, traduction française 2020)

Ressources disciplinaires / Données

- [Centres de référence thématiques](#)
Recherche Data Gouv

Astronomie et astrophysique



Système Terre et Environnement



Sciences Humaines et Sociales



→ catalogue/entrepôt : <https://data.progedo.fr/>

HN Huma-Num^{IR*} → catalogue/entrepôt : <https://nakala.fr/>

Biologie-Santé



- Espace Callisto Doranum avec des cours [thématiques et disciplinaires](#) : bioinformatique, informatique théorique, sciences de l'environnement, économie, droit...
- [Dataacc](#) : gestion des données en physique et chimie
- Guide du consortium [MASA](#)
- Guide [Données et recherches participatives](#)

Ressources généralistes / Codes et logiciels

- [The Turing Way](#) Guides for Reproducible Research, Project design, Collaboration, Communication...(2020-)
- [MOOC recherche reproductible](#) INRIA (2020-)
- [Passeport pour la Science ouverte – Codes et logiciels](#) (2022)
- [Ouvrir les codes sources](#), Guide Etalab
- [Accompagner la préservation et la diffusion des logiciels](#) dans les établissements de l'ESR, Formation ADBU-MédiaNormandie (2023)