

Module 4 : Partage et valorisation des données

Science Ouverte
Questions juridiques
Modalités de partage





Fanny Sébire - <https://orcid.org/0000-0002-6301-7147>

Anne-Caroline Delétoille - <https://orcid.org/0000-0002-8637-4040>

Cyril Pommier - <https://orcid.org/0000-0002-9040-8733>

Special thanks to:

Fredéric de Lamotte - <https://orcid.org/0000-0003-4234-1172>

Paulette Lieby - <https://orcid.org/0000-0002-9289-9652>

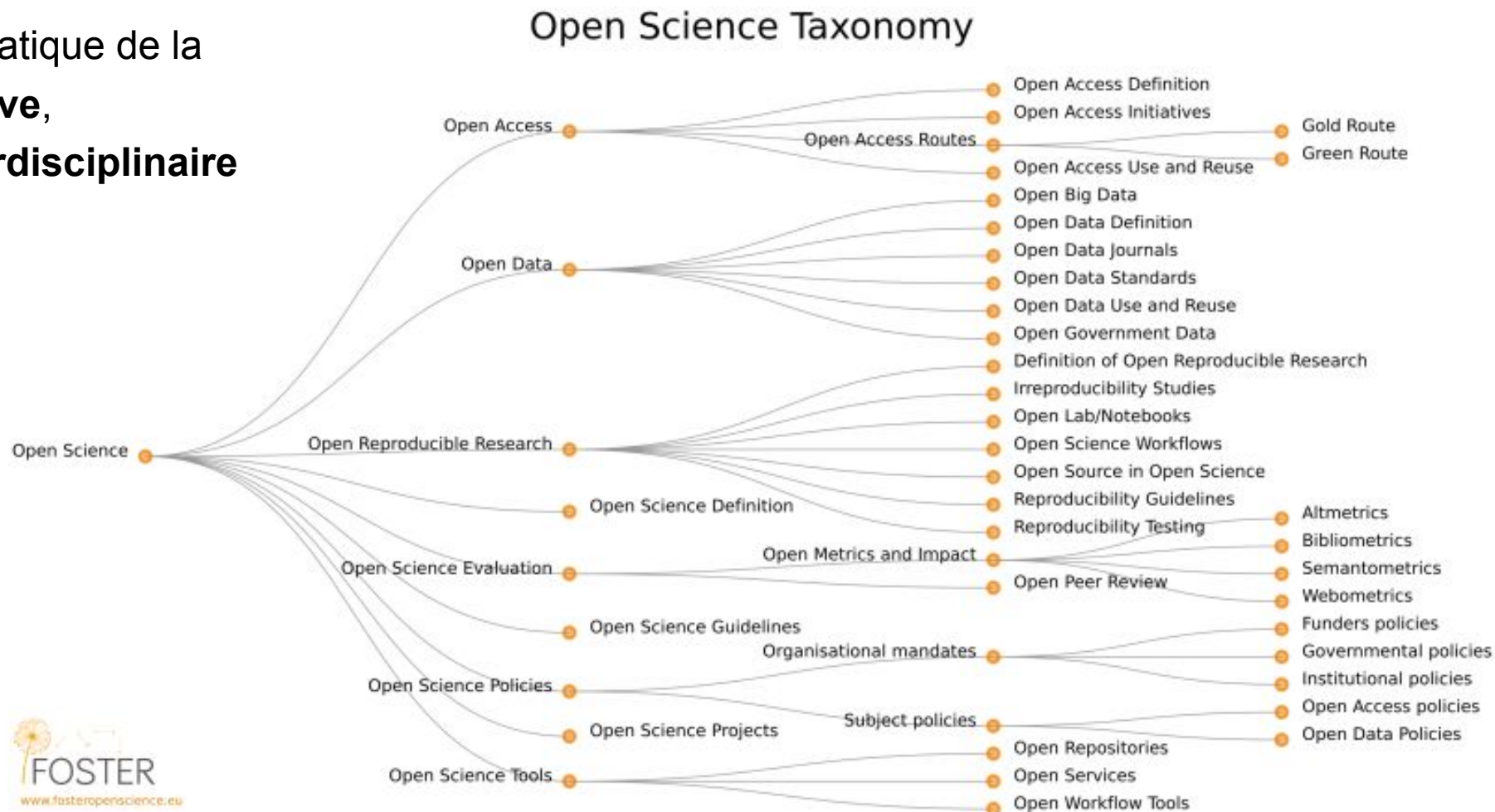
Lionel Maurel - <https://orcid.org/0000-0002-8667-5900>

Science ouverte : contexte national et international



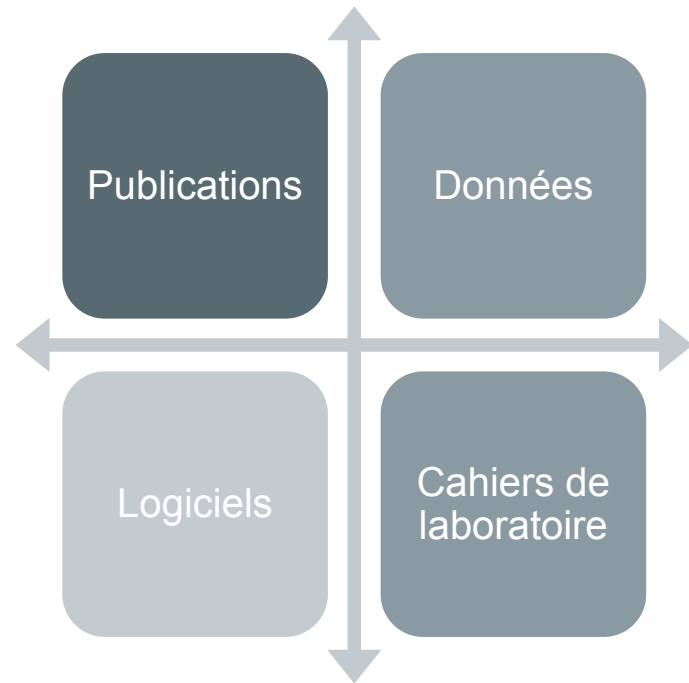


Nouvelle forme de pratique de la science : **collaborative**, **participative** et **interdisciplinaire**



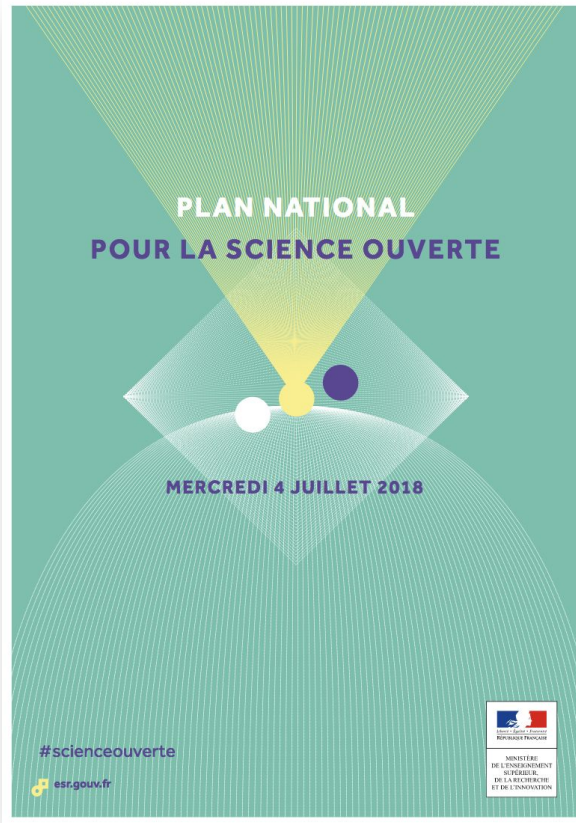
Source : Knoth, Petr; Pontika, Nancy (2015): Open Science Taxonomy. figshare. Figure. (<https://doi.org/10.6084/m9.figshare.1508606.v3>)

- Un des objectifs : rendre les **résultats de la recherche scientifique** accessibles à tous
- En permettant une diffusion et une réutilisation sans entrave des résultats de la recherche



Les ambitions :

- **démocratiser l'accès aux savoirs**
- **rendre la science plus cumulative**, plus fortement étayée par les données, plus transparente
- **augmenter l'efficacité de la recherche** en évitant de dupliquer les efforts, en réutilisant des données ou du matériel scientifique
- favoriser les **avancées scientifiques et l'innovation**
- favoriser la **confiance des citoyens dans la science**



2018



2021

Mesures

4

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

5

Créer Recherche Data Gov, 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)



Décret 2021-1572 du 03/12/2021 applicable aux établissements publics et fondations d'utilité publique

- L'intégrité scientifique = ensemble des règles et valeurs qui régissent les activités de recherche pour en garantir le caractère honnête et scientifiquement rigoureux.
- Nouvelles obligations pour les établissements :
 - Promouvoir la **mise à disposition des données et codes sources** associés aux résultats de la recherche ainsi que la publication des résultats négatifs
 - **Mettre en œuvre les plans de gestion de données**
 - Définir une politique de conservation, de communication et de réutilisation des résultats bruts et contribuer aux infrastructures qui le permettent

Montage du projet

Paragraphe sur la gestion des données

Budgéter la science ouverte (APC, data manager, stockage...)



Début du projet

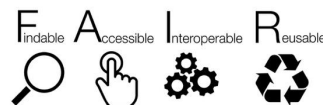
Plan de gestion des données (PGD)



Pendant le projet

Gestion des données suivant les principes FAIR

Mise à jour du PGD



Fin du projet

Partage des données « aussi ouvert que possible, aussi fermé que nécessaire »

Entrepôts de données

Open Access des publications





- **Horizon Europe – Grant Agreement [art. 17](#) :**

The beneficiaries must manage the digital research data generated in the action ('data') responsibly, in line with the FAIR principles and by taking all of the following actions:

- establish a data management plan ('DMP') (and regularly update it)
- as soon as possible and within the deadlines set out in the DMP, deposit the data in a trusted repository; if required in the call conditions, this repository must be federated in the EOSC in compliance with EOSC requirements
- as soon as possible and within the deadlines set out in the DMP, ensure open access — via the repository — to the deposited data, under the latest available version of the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a licence with equivalent rights, following the principle 'as open as possible as closed as necessary', unless providing open access would in particular:



- **ANR – [La science ouverte : un engagement de l'ANR](#)**

L'ANR participe à l'alignement européen et international en faveur de la structuration et de l'ouverture des données de la recherche. Le principe « aussi ouvert que possible, aussi fermé que nécessaire » est au cœur de sa démarche. L'Agence attire l'attention des coordinateurs sur l'importance de considérer la question de la gestion et du partage des données dès le montage du projet.

- **Bill and Melinda Gates Foundation - [Gates Foundation Open Access Policy](#)**

The Open Access Policy requires that underlying source data results are accessible and open immediately.



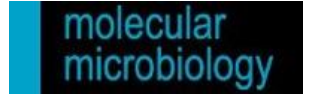
- **Wellcome Trust** – [Policy on data, software and materials management and sharing](#)

1. We expect our researchers to **maximise the availability of research data**, software and materials with as few restrictions as possible. As a minimum, the data underpinning research papers should be made available to other researchers at the time of publication, as well as any original software that is required to view datasets or to replicate analyses. Where research data relates to public health emergencies, researchers must share quality-assured interim and final data as rapidly and widely as possible, and in advance of journal publication.



Exigences des éditeurs

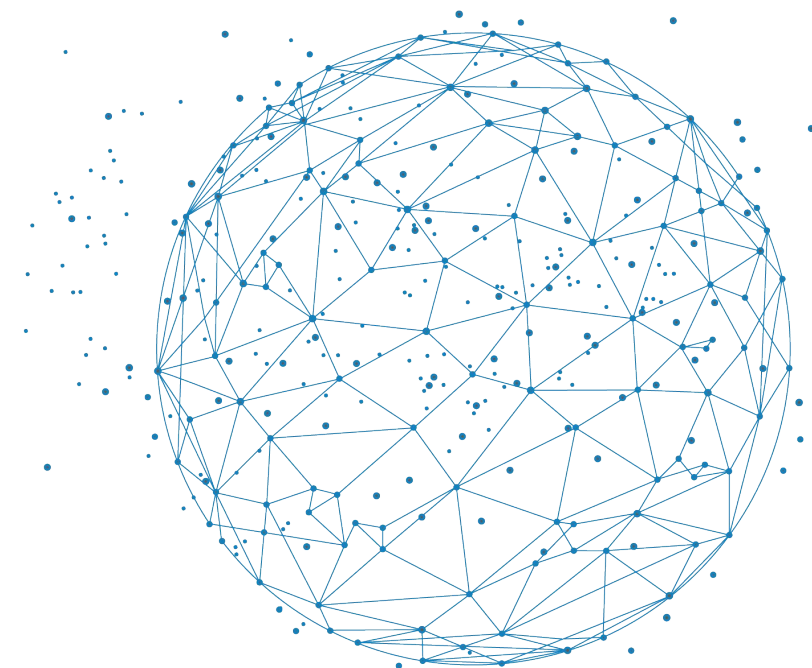
- Déposer les données associées à une publication dans un entrepôt de données



Politiques des établissements

- [Plan Données de la recherche du CNRS](#)
- [Charte de l'INRAE pour le libre accès aux publications et aux données](#)
- [Plan stratégique de l'INSERM](#)
- [Politique de gestion et partage des données de la recherche et codes logiciels de l'Institut Pasteur](#)

Questions juridiques





A votre avis...

- Quelles données doivent être ouvertes, quelles données doivent rester en accès restreint ?

<http://scrumblr.ca/open-closed-data>



*A priori**, les données issues d'une activité de la recherche sont soumises à une obligation de partage

Les données sont soumises à un principe d'ouverture par défaut et de libre utilisation (Loi République numérique 2016 LPRN)

*«*a priori*» : l'objet d'une partie de ce module est de définir les contours de cette obligation de partage



Retour sur la définition

- *Les « documents administratifs » visés par la LPRN sont tous les documents quels que soient leur date, leur lieu de conservation, leur forme et leur support qui sont produits ou reçus, dans le cadre de leur **mission de service public, par l'Etat, les collectivités territoriales ainsi que par les autres personnes de droit public ou les personnes de droit privé chargées d'une telle mission** (cf. EPICs)*
- Qualificatif important : on parle ici des documents **achevés**



Les données de la recherche sont des informations publiques :

- Principe d'**ouverture par défaut** et de **libre utilisation** (*Loi Lemaire - LPRN 2016*)
- Principe de **gratuité** (*Loi Valter 2015*) :
 - Seule une liste fermée d'administrations peuvent fixer des redevances de réutilisation (IGN, Météo France)
 - Articulation possible avec le dépôt de brevets et d'autres formes de valorisation
- **Protégées contre les risques d'accaparement**
 - Quels que soient les droits cédés à un éditeur, la partie concernant les données est nulle si elle va contre l'ouverture des données



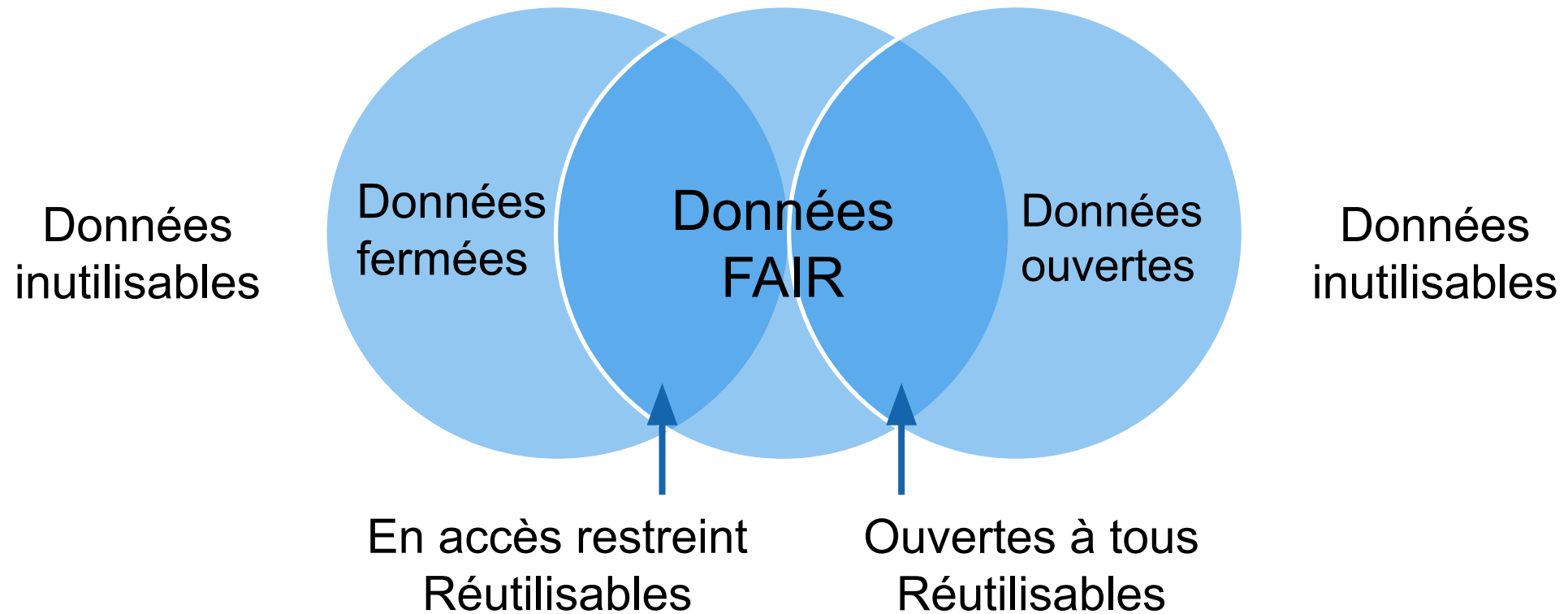
- La LPRN supprime le principe du régime dérogatoire où chaque établissement devait déterminer les modalités de la réutilisation des données
- Il y a des protections possibles pour certains types de données, il s'agit du **principe de l'exception**, fixé dans la loi et selon le principe de la commission européenne :

**« aussi ouvert que possible,
aussi fermé que nécessaire »**





Accessible ≠ Ouvert





*A priori**, les données issues d'une activité de la recherche sont soumises à une obligation de partage... mais :

- Propriété intellectuelle
- Données sensibles et secrets
- Brevets
- Bases de données
- Données à caractère personnel
- ...

sortent de l'ouverture par défaut



- **les projets partenariaux et les droits des tiers**
 - dans le cas de **recherche partenariale** : acteurs publics et privés
 - l'accès, la transmission, etc. aux données sont fixés par des règles, par notamment les **accords de consortium et les contrats**
- néanmoins, à priori, les données doivent être diffusées en open access si :
 - elles sont issues d'une activité de recherche financée **au moins pour moitié** par des fonds publics



Le **droit d'auteur**, pour les œuvres de l'esprit :

- les **textes (publications scientifiques...)**, **plans, dessins, graphiques** :
les chercheurs sont considérés comme des tiers par rapport à l'administration. Ils décident de la manière de diffuser, de partager leur œuvre.
- l'**image (photos, vidéos)**, qui sous certaines conditions d'originalité, peut aussi être considérée comme une œuvre :
 - la ligne de partage est fixée par l'**originalité**
- les **logiciels**
- etc...



Exception à l'exception en recherche pour la fouille de texte



Elles sont susceptibles de recevoir une double protection :

- Un droit particulier, appelé droit "*sui generis*" protège le contenu de la base de données, c'est-à-dire l'ensemble des données qu'elle contient.

Le droit *sui generis* peut protéger les bases de données, même à défaut d'originalité de celles-ci. L'objectif de la protection par le droit *sui generis* est de protéger les investissements réalisés dans le secteur des bases de données et d'empêcher la reprise des bases de données par des concurrents.

- Le droit d'auteur protège la structure de la base de données, si elle est originale.



- Données sensibles qui ne sont pas des données personnelles
 - données de biodiversité, PPST (protection du potentiel scientifique et technique de la nation), etc...
- Secret médical, secret des affaires, secret des procédés, secret militaire
- Secret des statistiques (loi de 1951) :
 - S'applique aux personnes collectant des données à visée statistique
 - Interdit la réutilisation des données pour réidentifier les personnes



Le RGPD en 5 points :

- Déclarer TOUS les traitements de données à caractère personnel au DPO (registre des traitements) et s'assurer de leur conformité
- Information des personnes concernées +/- consentement
- Sécurisation des supports et des transferts de données
- Droits des personnes (accès, modification, suppression et portabilité)
- PIA (Privacy Impact assessment) pour les données sensibles

DPO = Délégué à la protection des données

-> s'assure de la conformité des traitements de DcP (information des personnes, droits des personnes, registre de traitements de DcP)

-> Il N'est PAS responsable

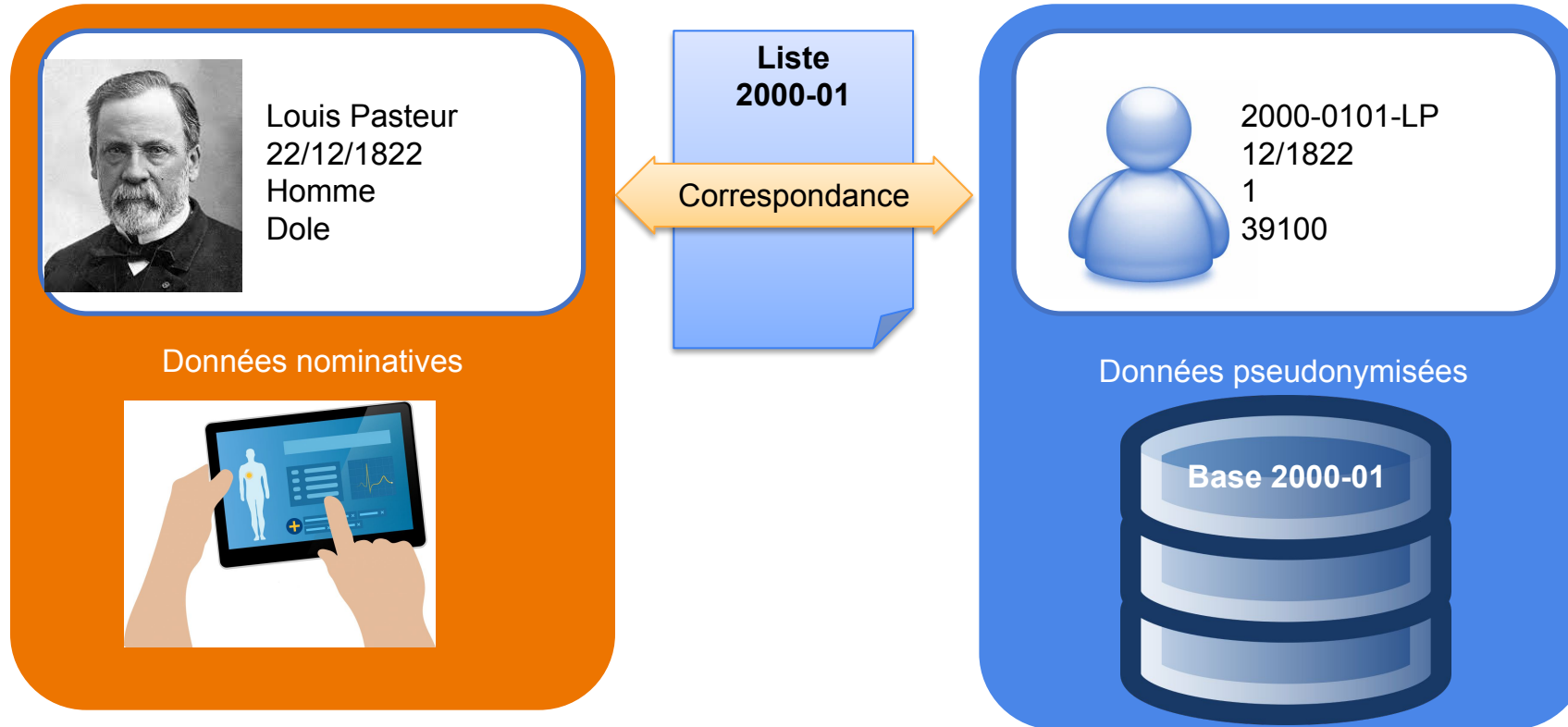


- *Loi Informatique et Liberté* + [RGPD](#) (UE-2018) : les données à caractère personnel **sont par principe fermées**. Il est possible d'envisager leur ouverture :
 - **Obligation** de recueillir le **consentement** +/- de **pseudonymiser** les données
 - ou **anonymiser** les données.
- Il existe néanmoins des **dérogations pour la recherche**, à étudier projet par projet
 - Consultez [l'arbre de décision](#) de UE
 - Rapprochez-vous de votre DPO !
- le RGPD n'autorise pas le transfert de données personnelles hors UE -> nécessite une analyse juridique du DPO

Choisissons une personne connue, combien faut-il de questions à [Akinator](#) pour la retrouver ?



<https://digital.essca.fr/donnees-personnelles-qui-est-ce>





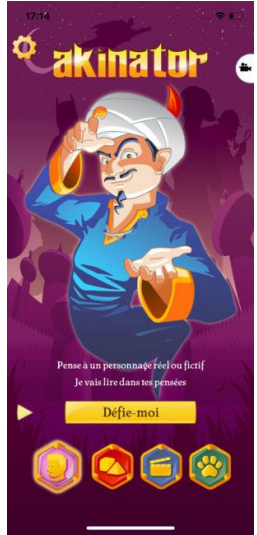
S'il est possible de ré-identifier une personne en croisant des données, ces données ne sont pas considérés comme anonymes !



La CNIL a publié un [guide sur l'anonymisation des données personnelles](#) (mai 2020)

Les autorités de protection des données européennes définissent **trois critères** qui permettent de s'assurer qu'un jeu de données est véritablement anonyme :

- **l'individualisation** : il ne doit pas être possible d'isoler un individu dans le jeu de données ;
- **la corrélation** : il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu ;
- **l'inférence** : il ne doit pas être possible de déduire, de façon quasi certaine, de nouvelles informations sur un individu.





La CNIL indique **2 procédés d'anonymisation** :

- **La généralisation**

- Modifie l'échelle des attributs des jeux de données, ou leur ordre de grandeur, afin de s'assurer qu'ils soient communs à un ensemble de personnes (ex : tranche d'âge).
- > Eviter l'individualisation d'un jeu de données et limite les possibles corrélations entre jeux de données

- **La randomisation :**

- Modification des attributs dans un jeu de données (perte de précision), tout en conservant la répartition globale.
- > Protège le jeu de données du risque d'inférence



- Origine raciale ou ethnique
- Opinions politiques, convictions religieuses ou philosophiques, appartenance syndicale,
- **Données concernant la santé,**
- **Données génétiques,**
- Données biométriques permettant d'identifier une personne physique de manière unique,
- Données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique.



- Quelles spécificités pour la recherche sur la personne humaine ?
 - **RGPD/CNIL** – données sensibles 2018
 - **Code de la Santé Publique** (CPS) + Loi Jardé 2016
- A garder en tête pour la gestion des données :
 - Principe de minimisation : ne collecter que les données nécessaire à la recherche
 - Consentement des participants à la recherche (CSP + RGPD)
 - Sécurité informatique des supports et transferts de données



Entrepôt de données de santé = **entrepôt de données dont le but est de permettre la réutilisation des données qu'il contient**

- CNIL, Référentiel nov 21
- La réutilisation de ces données nécessite des formalités spécifiques :
 - Conformité à une méthodologie de référence (type MR004)
 - Demande d'autorisation de recherche à la CNIL
- Inconvénient : Partage de données à l'étranger (et particulièrement hors UE) très compliqué
- Information des personnes



Open data et données de la recherche sur la personne humaine :

- Résumés de protocoles de recherche déposés sur des registres publics (clinicaltrial.gov; Eu clinical trial register)
- Résultats de la recherche
 - Engagement des promoteurs de rendre public les résultats des recherches (engagement OMS : [Joint statement on public disclosure of results from clinical trials](#))
 - Dispositions légales sur l'information sur les résultats globaux des recherches (CSP)



« Déclaration universelle sur le génome humain et les droits de l'homme », UNESCO, 11 novembre 1997.

- art. 1 : "Le génome humain sous-tend l'unité fondamentale de tous les membres de la famille humaine, ainsi que la reconnaissance de leur dignité intrinsèque et de leur diversité. Dans un sens symbolique, il est le patrimoine de l'humanité."
- art.4 : "Le génome humain en son état naturel ne peut donner lieu à des gains pécuniaires."

Principes des Bermudes, 1996 : Les séquences génomiques primaires devraient

- faire partie du **domaine public**, être accessibles librement et gratuitement
- être **mises à disposition rapidement** (« on a daily basis »)

Les séquences finalisées (annotées) devraient être versées **dans des bases de données publiques**



Attention à la ré-identification à partir de données génétiques !

- triangulation possible des données génétiques
- quantité de données disponibles dans le domaine public
- évolutions techniques

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,^{1,2,3,4} Amy L. McGuire,⁵ David Golan,⁶ Eran Halperin,^{7,8,9} Yaniv Erlich^{1*}

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

GYMREK Melissa, McGuire Amy L., GOLAN David, HALPERIN Eran, ERLICH Yaniv. Identifying Personal Genomes by Surname Inference. Science (2013).

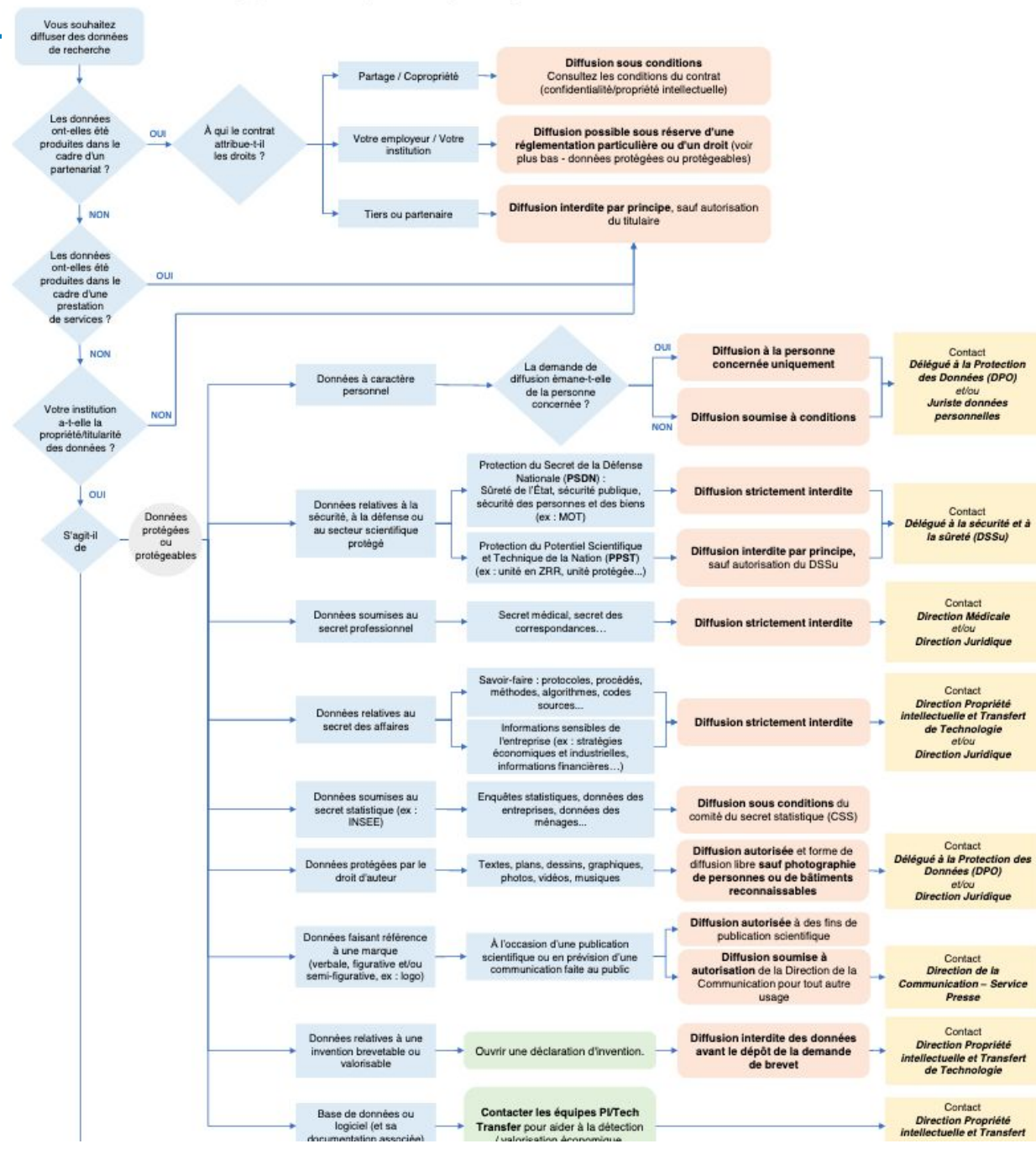
Homer N, Szeling S, Redman M, et al. () *Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays*. PLoS Genet. 2008;4:E1000167



Attention au Le Cloud Act (Clarifying Lawful Overseas Use of Data Act) :

- Concerne les entreprises américaines
- Votée en mars 2018, cette loi permet aux États-Unis d'accéder plus facilement aux données stockées sur des serveurs de société américaines situés hors des États-Unis.

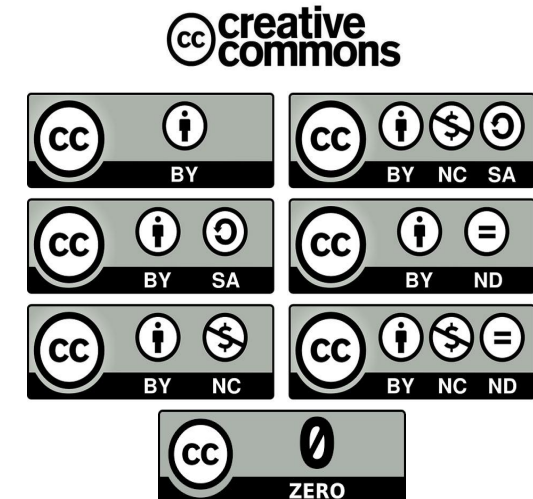
- [INRAe](#)
- [CIRAD](#)
- [Institut Pasteur](#)
- [Ponts et Chaussées](#)
- [Ouvrir la science](#)
- un outil : [DAISY](#) (ELIXIR-LU) – pour les données soumises au RGPD



- Les licences pour encadrer le partage et la réutilisation des données
 - Permet d'accorder à l'avance aux utilisateurs certains **droits d'utilisation**
 - Peut comporter des **restrictions d'usage**
 - Il est fortement recommandé d'en utiliser une dans tous les cas **pour clairement afficher les droits afférents**

- Exemple : licences Creative Commons
 - toutes n'ont pas de jurisprudence spécifique en France mais leur validité n'est pas remise en cause

- LPRN : une liste de licences [Licences - data.gouv.fr](https://data.gouv.fr/licences)





Le procès « *Havasupai Tribe versus Arizona State University Board of regents* »

- 1989 – Etude menée auprès de la Tribu Havasupai sur le diabète: les chercheurs collectèrent des échantillons de sang et des données
- 2003 – une membre de la Tribu assiste à un cours à l'Université d'Arizona, où elle découvre que les échantillons Havasupai ont aussi été utilisés pour analyser :
 - ✓ la prédisposition de la tribu aux maladies psychiatriques (schizophrénie)
 - ✓ son degré de consanguinité
 - ✓ Le parcours migratoire de la tribu
- Les havasupai disent qu'ils ont été mal informés, et demandent le retour de leurs échantillons □ -> refus de l'équipe -> lancement de 2 actions en justice
- 2010 : accord financier entre la tribu et l'Université

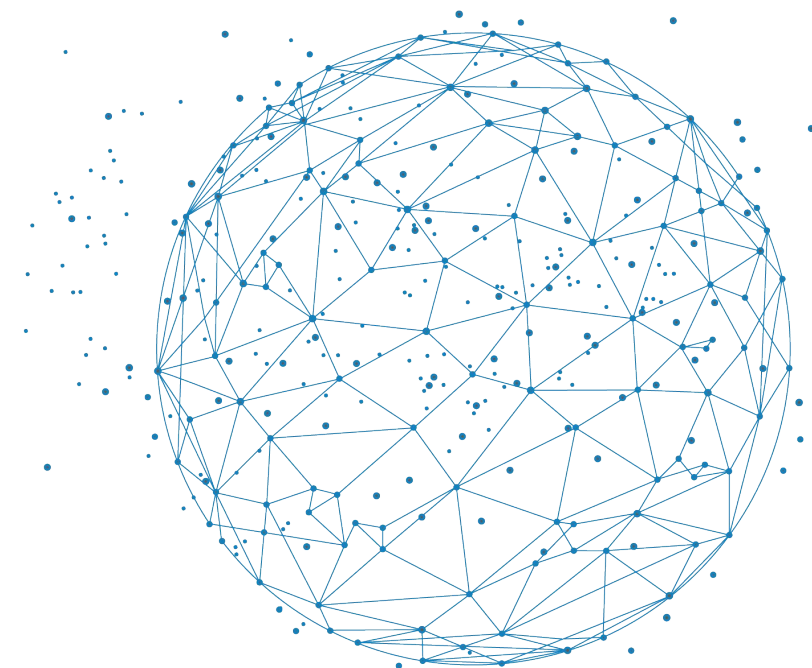


- en France
 - Loi Lemaire - LPRN 2016 (*ouverture des données*)
 - Loi Valter 2015 (*gratuité*)
 - Loi Informatique et Liberté 1978 (mise à jour en 2019 pour intégrer le RGPD)
 - [Décret 2021-1572 sur l'intégrité scientifique](#)
- en Europe
 - [Directive on open data and the re-use of public sector information](#), Open Data Directive, 16 July 2019
 - [Regulation on a framework for the free flow of non-personal data in the EU](#) (EU) 2018/1807 (*free movement of non-personal data*)
 - RGPD 2018



- [Aspects juridiques et éthiques – DoRANum](#)
- [Ouverture des données de la recherche : de quoi parle-t-on ?](#)
- [À qui appartiennent les données ?](#)
- [Les principes FAIR – DoRANum](#)
- [Ouverture des données de la recherche. Guide d'analyse du cadre juridique en France](#)
- [En route vers la science ouverte : Gérer les données de sa recherche](#)
- [OCDE : Recommandation du Conseil concernant l'accès aux données de la recherche financée sur fonds publics](#)
- [Guide d'application de la loi pour une République numérique](#)
- [Droit d'auteur et fouille de texte](#)
- [Déclaration de l'AMM sur les considération éthiques concernant les BDD de santé et les biobanques](#)

Modalités de partage





➔ Se rappeler les questions du PGD

- Quels jeux de données sont partagés ? ➔
- Quel est le potentiel de réutilisation ? ➔
- Quand ? ➔
- Où ? dans quel entrepôt de données ? ➔
- Comment ? ➔
 - quelles modalités ?
 - quelles licences ?
 - pour quels publics ?

Considérer :

Restrictions au partage ?

présence de données sensibles, personnelles, de santé, issues de partenariats avec le privé...

Embargo ?

pour publication, dépôt de brevet, exploitation

Limites de réutilisation ?

Utilisation commerciale, accès sur demande seulement...



Ceci doit être renseigné et justifié dans le PGD



- Données issues de partenariats privés
- Données sensibles
 - concernant des espèces protégées ou envahissantes
 - données cliniques, issues d'expérimentations animales
 - données personnelles (soumises au RGPD)
 - issues de ressources biologiques du Sud (Protocole de Nagoya)
- Données stratégiques que vous souhaitez exploiter
 - identification de marqueurs génétiques, d'arômes
 - création d'une appli, d'une base de données originale
- Jeux de données contenant des données préexistantes
 - produites par d'autres
 - sous licences non ouvertes

***Se référer aux
logigrammes / outils
d'aide à la décision***



A priori toutes les données non listées dans les données à ne PAS partager

Mais en particulier :

- Données ayant un potentiel de réutilisation
 - Nouvelles analyses, nouvelles questions de recherche
 - Méta-analyses, nourrir des modèles
 - etc.
- Données utiles
 - jeux de données contrôles, lots témoins
- Données présentant un intérêt pour certains publics (ex: société civile, pays du Sud)
- Données dont le coût d'obtention est élevé

Et aussi : des données FAIR

- Correctement décrites, contextualisées □ réutilisables
- Dans un format interopérable

Exemples :

- données environnementales, climat,
- gestion des territoires
- santé publique
- genre
- favorisant la participation citoyenne



➔ estimer la valeur de ses données

- Données rares ou uniques
 - expérimentation impossible à répéter
 - groupes difficilement accessibles
 - phénomènes rares
- Données à forte valeur scientifique
 - données de référence
 - reproduction difficile ou coûteuse
 - ayant un grand intérêt pour certains publics (ex: société civile, pays du Sud)
- Données ayant une valeur économique
 - perspectives d'application, développement commercial
- Données ayant une valeur environnementale



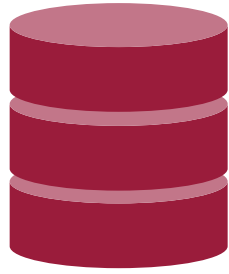
- Communautés scientifiques
- Enseignants
- Décideurs
- Secteur privé, créateurs de start-up
- ONGs, associations internationales influentes
- Journalistes
- Grand public



- Dépend
 - du bailleur (Commission Européenne, ANR, fondation Gates, ...)
 - de politiques institutionnelles ou nationales
 - de politiques de certains partenaires
 - de la revue de publication
- Après avoir
 - exploité vos données
 - publié vos résultats de recherche
 - mis en forme vos données et métadonnées
 - anonymisé vos données
 - obtenu l'accord de tous vos partenaires

Exemple : indiquer dans le PGD :

- Les données seront déposées dans Zenodo et seront accessibles après un embargo d'un an (pour publier 2 articles)
- Les données seront disponibles sur demande dès l'année 3 puis seront accessibles sur ENA dès la fin du projet

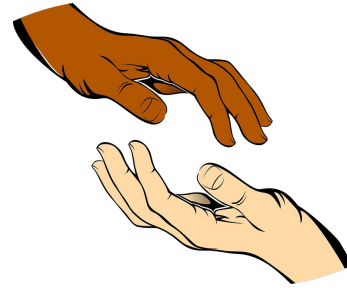


Entrepôt de données



Recommandé par les financeurs

- **visibilité, citabilité, préservation**
- **optimise les possibilités de réutilisation**



Sur demande



Supplementary data dans les articles



Préférer plutôt : indiquer dans l'article le lien vers un jeu de données dans un entrepôt



Services en ligne pour la collecte, la description, la préservation, la découverte et la diffusion de données

Stocker des jeux de données et leurs métadonnées



Trouver les jeux de données et les réutiliser

- Environ 1500 entrepôts de données en sciences de la vie



Entrepôts ouverts



Entrepôts généralistes



Entrepôts institutionnels



Entrepôts éditeurs



Entrepôts avec accès contrôlé



Entrepôts spécifiques d'un domaine



Entrepôt national (à venir)

RECHERCHE DATA GOUV

La plateforme nationale fédérée des données de la recherche



- **Institutionnels** : Dataverse



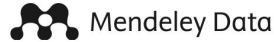
- **Europe** : Zenodo, B2Share



- **Généralistes** : Figshare, Dryad



- **Editeurs** : Oxford Univ Press (GigaDB) ; Ubiquity Press (Dataverse) ; Elsevier (Mendeley Data)



- **Thématiques**

- GBIF (Global Biodiversity Information Facility)



- KNB (Knowledge Network for biocomplexity), EDI (Environmental Data Initiative)

- Pangaea, SEANOE



- Movebank, WormBase, ViPR, MycoBank, ComBase, FLOW



- GenBank, Barcode of Life Data Systems, UniProt, Intact



- Entrepôt financé par la **Commission Européenne**
- Hébergé au **CERN** (European Organization for Nuclear Research)
- Issu d'une collaboration avec **OpenAire** (Open Access Infrastructure for Research in Europe)
- Objectif : partage des **résultats de la recherche** pour lesquels on ne dispose pas d'entrepôt institutionnel, disciplinaire ou thématique.

<https://zenodo.org/>

The screenshot shows the Zenodo website interface. At the top, there is a blue navigation bar with the Zenodo logo on the left, a search bar in the center, and 'Upload' and 'Communities' links on the right. Further right are 'Log in' and 'Sign up' buttons. Below the navigation bar, the main content area shows search results. On the left, there are filters for 'All versions', 'Access Right' (with options for Open, Closed, Restricted, and Embargoed), and 'File Type'. The main results area displays 'Found 138455 results.' with a pagination bar showing page 1 of 9. A specific result is highlighted: 'PSSH2 - database of protein sequence-to-structure homologies (including Sars-CoV-2 structures)'. This result is dated 'February 9, 2022 (2021-11)', is a 'Dataset', and is 'Open Access'. It lists authors: Andrea Schafferhans, Sean O'Donoghue, Neblina Sikta, and Sandeep Kaur. The description states: 'Protein sequence and structure data This data set contains data from Uniprot (in the files called protein_sequence, protein_synonyms, protein_names, organism_synonyms) and PDB (in the files called PDB and PDB_chain) as used by the Aquaria web resource at the time of download (2022-02-08)'. It also notes 'Uploaded on February 10, 2022' and '3 more version(s) exist for this record'. A 'View' button is visible next to the result.

- Se conformer aux **exigences de certains financeurs et éditeurs**

Horizon Europe – Grant Agreement art. 17 :

The beneficiaries must manage the digital research data generated in the action ('data') responsibly, in line with the FAIR principles and by taking all of the following actions:

- establish a data management plan ('DMP') (and regularly update it)
- as soon as possible and within the deadlines set out in the DMP, deposit the data in a **trusted repository**; if required in the call conditions, this repository must be federated in the EOSC in compliance with EOSC requirements
- as soon as possible and within the deadlines set out in the DMP, ensure open access — via the repository — to the deposited data, under the latest available version of the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a licence with equivalent rights, following the principle 'as open as possible as closed as necessary', unless providing open access would in particular:

Pourquoi déposer ses données dans un entrepôt ?

- Se conformer aux **exigences de certains financeurs et éditeurs**
- Rendre ses données **faciles à trouver** et **accessibles** sur le long terme



Les articles pour lesquels les données sont disponibles publiquement sont davantage cités (Piowar et al. 2013)

OpenAIRE | EXPLORE

SEARCH DEPOSIT LINK CONTENT PROVIDERS

Research outcomes Search in OpenAIRE for scholarly works SEARCH

Open Access

Include: Publications Research data Software Other research products

Advanced Search

856K datasets interlinked with publications

Scholixplorer zenodo DataCite PANGAEA figshare protocols.io

OpenTrials kaggle reactome EASY DRYAD

OpenAire Explore : <https://explore.openaire.eu/>

DataCite
FIND, ACCESS, AND REUSE DATA

Search for work Search



- Se conformer aux **exigences de certains financeurs et éditeurs**
- Rendre ses données **faciles à trouver et accessibles** sur le long terme
- Rendre ses données **compréhensibles et interprétables**

Summary

Title
Quantitative proteomic dataset from oro- and naso-pharyngeal swabs used for COVID-19 diagnosis: detection of viral proteins and host's biological processes altered by the infection

Description
We recovered proteins from SARS-CoV-2 positive and negative oro/naso-pharyngeal swabs, obtained from diagnostic center localized at Institut Pasteur de Montevideo, and performed comparative quantitative proteomic analysis.

Sample Processing Protocol
Oro- and nasopharyngeal swabs were inactivated with 2% SDS. Protein extraction was performed by vigorous agitation and supernatant was recovered. Proteins were first separated shortly (1 cm) in a SDS-PAGE. Gels slices were destained, reduced, alkylated and digested with trypsin as usual. Peptide extraction from gel pieces was performed. Samples were analyzed by nano-LC MS/MS using a shotgun strat...

Read more

Data Processing Protocol
PatternLab for Proteomics (version 4.0) software (<http://www.patternlabforproteomics.org>) was employed to generate a target-reverse database using sequences from Homo sapiens and Severe acute respiratory syndrome coronavirus 2, both downloaded from Uniprot consortium in June, 2020 (<http://www.uniprot.org>). In addition, 127 common mass spectrometry contaminants were incorporated. Thermo raw files w...

Read more

Contact
Anaía Lima, Institut Pasteur de Montevideo, Proteomic and Analytical Biochemistry Unit
Rosario Durán, Analytical Biochemistry and Proteomics Unit, Institut Pasteur de Montevideo and Instituto de Investigaciones Biológicas Clemente Estable, Montevideo, Uruguay (lab head)

Submission Date
16/07/2020

Publication Date
28/07/2020

Properties

Organism
Sars bat coronavirus
Homo sapiens (human)

Organism part
Pharyngeal mucosa

Diseases
Covid-19

Modification
No PTMs are included in the dataset

Instrument
Q Exactive

Software
Unknown

Experiment Type
Sars-cov-2
Covid-19

Quantification
Spectrum counting

Dataset reuses
Not available

Similar Studies

[Embryonic developmental arrest in the annual killifish Austrolebias charrua: a proteomic approach to diapause III.](#)
2021-04-06

Exemple de description d'un jeu de données sur PRIDE (PRoteomics IDentifications Database)

<https://www.ebi.ac.uk/pride/archive/projects/PXD020394>



- Se conformer aux **exigences de certains financeurs et éditeurs**
- Rendre ses données **faciles à trouver et accessibles** sur le long terme
- Rendre ses données **compréhensibles et interprétables**
- **Contrôler l'accès** à ses données ou attribuer une **licence de diffusion**

Ex : données accessibles après validation par un comité scientifique

Dataset

[Browse files](#)

16S-based fecal microbiota composition

Dataset ID	Technology	Samples
EGAD00001004979	N/A	1311

Dataset Description

The dataset reports the 16S rRNA gene sequencing of the fecal microbiota of donors from the Milieu Intérieur Cohort. The Milieu Intérieur cohort includes a total of 1,000 healthy individuals of western European ancestry, recruited in France as part of the Milieu Intérieur project. To assess their fecal microbiota composition, 16S rRNA profiles were generated from stool samples of 863 of the 1,000 donors. Human stool samples were produced at home no more than 24 hours before the scheduled medical visit and collected in a double-lined sealable bag maintaining strict anaerobic conditions. Upon reception at the clinical site, the fresh stool samples were aliquoted and stored immediately at -80°C. DNA was extracted from stool and barcoding PCR was carried out using indexed primers targeting the V3-V5 region of the 16S rRNA gene. Equal volumes of normalized PCR reaction were pooled and thoroughly mixed. The amplicon libraries were sequenced on Illumina MiSeq.


Data Use Conditions

See further information on [Data Use Conditions](#)

Label	Code	Version	Modifier
general research use	DUO:0000042	2019-01-07	

Who controls access to this dataset

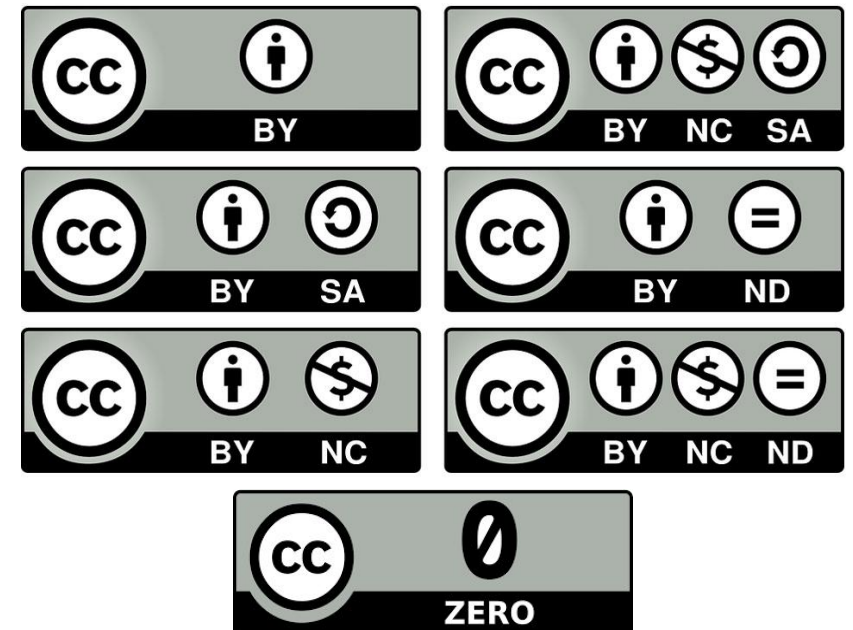
For each dataset that requires controlled access, there is a corresponding Data Access Committee (DAC) who determine access permissions. Access to actual data files is not managed by the EGA. If you need to request access to this data set, please contact:

 **Milieu Intérieur Data Access Committee**
Contact person: Darragh Duffy
Email: milieuinterieudac [at] pasteur [dot] fr
More details: **EGAC00001001785**

Exemple de description d'un jeu de données sur EGA (European Genome-Phenome Archive) : <https://ega-archive.org/datasets/EGAD00001004979>



- Privilégier une licence **largement utilisée** et compatible avec les autres licences existantes, afin de faciliter la compilation de vos données avec d'autres données mises à disposition sous d'autres licences
- Tenir compte du potentiel de vos données et des restrictions appliquées
 - **Etalab** : plutôt pour une distribution en France
 - **Creative Commons** : pour l'international
 - CC-BY et CC-BY-SA les plus utilisées
 - Option ND (sans modification) à éviter
 - CC-0 à éviter si vous souhaitez être cité



<https://creativecommons.org/choose/>

<https://datapartage.inrae.fr/Partager-Publier/Choisir-une-licence>

LPRN : <https://www.data.gouv.fr/fr/pages/legal/licences/>





- Chacun dépose un jeu de données dans l'espace de test de Zenodo :
<https://sandbox.zenodo.org/>

- Nommage des jeux de données pour pouvoir les retrouver facilement :

Formation IFB Mars 2022

- Chacun choisit des conditions d'accès différentes :

Access right *

-  Open Access
-  Embargoed Access
-  Restricted Access
-  Closed Access



Exemple : <https://doi.org/10.15454/IASSTN>

A multi-site experiment in a network of European fields for assessing the maize yield response to environmental scenarios

Version 2.0



Millet, Emilie J.; Pommier, Cyril; Buy, Mélanie; Nagel, Axel; Kruijjer, Willem; Welz-Bolduan, Therese; Lopez, Jeremy; Richard, Cécile; Racz, Ferenc; Tanzi, Franco; Spitkot, Tamas; Canè, Maria-Angela; Negro, Sandra S.; Coupel-Ledru, Aude; Nicolas, Stéphane D.; Palaffre, Carine; Bauland, Cyril; Praud, Sébastien; Ranc, Nicolas; Presterl, Thomas; Bedo, Zoltan; Tuberosa, Roberto; Usadel, Björn; Charcosset, Alain; van Eeuwijk, Fred A.; Draye, Xavier; Tardieu, François; Welcker, Claude, 2019, "A multi-site experiment in a network of European fields for assessing the maize yield response to environmental scenarios", <https://doi.org/10.15454/IASSTN>, Portail Data INRAE, V2, UNF:6:zF9w0A2f+MHeW7maeeXJWA== [fileUNF]

Citer le dataset ▾

Pour en apprendre davantage sur le sujet, consulter le document [Data Citation Standards \[en\]](#).

Modalités d'accès au dataset ▾

Contact

Partager

Statistiques d'utilisation sur les datasets



6 834 consultations ?

2 783 téléchargements ?

3 citations ?



Exemple : <https://doi.org/10.15454/IASSTN>
Generic metadata

Common metadata

Description ?

Subject ?

Mot-clé ?

Related Publication ?

Identifiant pérenne du dataset ?

Date de publication ?

Title ?

Link to data ?

Contact ?

Author ?

Contributor ?

Producer ?

Distributor ?

Kind of Data ?

Data Origin ?

Life cycle step ?

Related Publication ?

Grant Information ?

Project Information ?

Time Period Covered ?

Date of Collection ?

Depositor ?

Deposit Date ?

Data:

external

Link to data ?

<https://urgi.versailles.inra.fr/ephep/ephep/viewer.do#>

internal

1 à 10 de 11 Fichiers

0_Data_outline.pdf
application/pdf - 1.0 Mo - 27 mars 2019 - 395 téléchargements
MD5: 1a11142041623c0364c90acae645ae78
This file contains the outline of the files organization and content. It describes the different measurement scales available: Weather and soil water data at daily time step (1). Phenotypic data at the plot level in each experiment (2a) and at the genotypic level in each experiments (2b and 3). Variable calculated at the genotype level (4). Phenotypic data at the plant level in each experiment for the reference genotype (5 and 6). Genotyping data available for each genotype (7a and 7b). Description of the genotypic material used (8). The map of the experimental sites has been drawn with the R package ggmap (Kahle and Wickham, 2013, doi:10.32614/RJ-2013-014). The maize 3D canopy is adapted from Pradal et al., 2008 (<https://doi.org/10.1071/FP08084>). The maize 3D plant is adapted from Fournier and Andrieu, 1999 (<http://dx.doi.org/10.1051/agro:19990311>).

1-Env_variables_daily-1.tab
Données tabulaires - 687.6 Ko - 5 nov. 2019 - 351 téléchargements
30 Variables, 2633 Observations - UNF:6:ubHZzvd3YZcP64UGsyUaww==
This file contains the environmental characterization of each Location x year combination. Based on the recorded weather, daily environmental variables were calculated. This file contains 30 columns: «year», «Site», «Env»: experiments ID with year of experiment («year»), location («Site»). In «Env», experiments are described by the three first letters of the city's name followed by the year of experiment. «Date»: Calendar date in the crop cycle. «D20.air», «D20cum.air»: thermal time calculated



Exemple : <https://doi.org/10.15454/IASSTN>
Generic Metadata

[Fichiers](#) [Métadonnées](#) [Conditions](#) [Versions](#)

Conditions d'utilisation ^

Licence accordée ?

Les [normes de la communauté Dataverse](#) de même que les bonnes pratiques scientifiques exigent que toute source utilisée soit citée correctement. Veuillez utiliser la référence bibliographique ci-dessus générée par Dataverse.

Aucune licence n'a été sélectionnée pour ce dataset.

Conditions d'utilisation ?



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

Registre des visiteurs ^

Registre des visiteurs ?

Aucun guestbook n'est associé à ce dataset donc aucun renseignement ne vous sera demandé concernant le téléchargement des fichiers.



Détails des différences de version

A multi-site experiment in a network of European fields for assessing the maize yield response to environmental scenarios

Version : 1.1
Dernière mise à jour: 28 mars 2019 15:43:55 CET

Version : 2.0
Dernière mise à jour: 5 nov. 2019 14:24:11 CET

Citation Metadata

Link to data

<https://urgi.versailles.inra.fr/ephep/ephep/view/trialSetIds=42> <https://urgi.versailles.inra.fr/viewer.do#dataResults/trialSetIds=42>

Topic Classification

Climate

Plant Breeding and Plant Products

Topic Classification

Genetic variability of grain yield under changing climate

Plant Breeding and Plant Products

Related Publication

doi

Genome-Wide Analysis of Yield in Europe: Allelic Effects Va
Heat Scenarios. 2016. Emilie J. Millet, Claude Welcker, Will
Negro, Aude Coupel-Ledru, Stéphane D. Nicolas, Jacques
Bauland, Sebastien Praud, Nicolas Ranc, Thomas Presterl,
Zoltan Bedo, Xavier Draye, Björn Usadel, Alain Charcosset,
& François Tardieu. 2016. Plant Physiology, 172 (2) 745
10.1104/pp.16.00621; <http://www.plantphysiol.org/content/172/2/745>
/749.long" rel="noopener"><http://www.plantphysiol.org/content/172/2/745>
/749.long; Genomic prediction of maize yield across E
environmental conditions. 2019. Emilie J. Millet, Willem Kru
Ledru, Santiago Alvarez Prado, Llorenç Cabrera-Bosquet, S
Alain Charcosset, Claude Welcker, Fred van Eeuwijk &
Nature Genetics, 51, 952–956; doi; <https://doi.org/10.1038/s41588-019-0411-4>
<https://www.nature.com/articles/s41588-019-0411-4>
rel="noopener"><https://www.nature.com/articles/s41588-019-0411-4>

Identifiant du fichier: 82962

Identifiant du fichier: 95808

Zenodo et Dataverse : points communs

Exemple : <https://doi.org/10.15454/IASSTN>
Specialized (Plant) Metadata

Citation Metadata ▾

Life Sciences Metadata ▲

Organism ? Zea mays



Social Science and Humanities Metadata ▾

Life Sciences Metadata ▲

Design Type ? Sélectionner...

Other Design Type ?

Factor Type ? Sélectionner...

Other Factor Type ?

Organism ? Zea mays ✕

Other Organism ?

Measurement Type ? Sélectionner...

Other Measurement Type ?

Technology Type ? Sélectionner...

Other Technology Type ?

Technology Platform ? Sélectionner...

Other Technology Platform ?

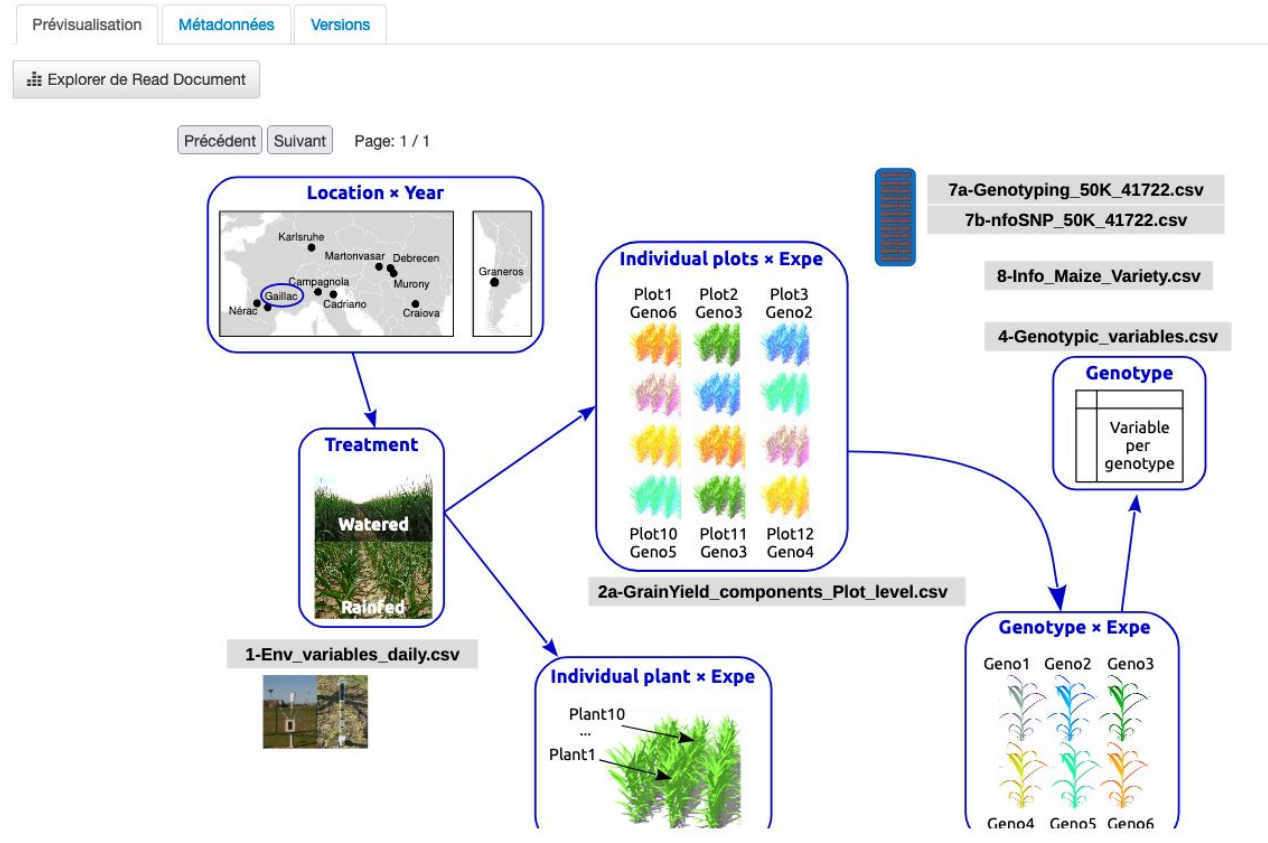
Cell Type ?

Sample type ?

Protocol type ?



Exemple : <https://doi.org/10.15454/IASSTN>
Specialized (Plant) Metadata
Metadata File : Simplified provenance





Exemple : <https://doi.org/10.15454/IASSTN>

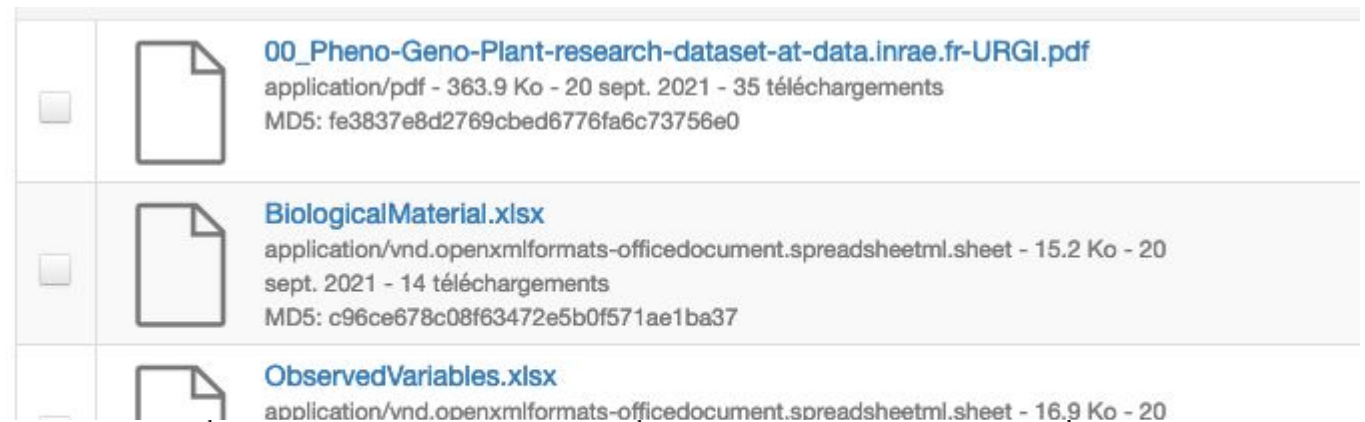
Specialized (Plant) Metadata

Metadata File : Biological Material, Metadata CSV/XLS template

Dedicated Guidelines

- <https://doi.org/10.15454/BWJVVG>

Metadata templates



	Accession_Number	accession_holding	Material source DOI	Material source ID (Holding institute/stock centre, accession)	Biological material ID*	Organism*
Definition			Digital Object Identifier (DOI) of the material source	An identifier for the source of the biological material, in the form of a key-value pair comprising the name/identifier of the repository from which the material was sourced plus the accession number of the repository for that material. Where an accession number has not been assigned, but the material has been derived from the crossing of known	Code used to identify the biological material in the data file . Should be unique within the Investigation. Can correspond to experimental plant ID, seed lot ID, etc... This material identification is different from a BiosampleID which corresponds to Observation Unit or Samples sections below.	An identifier for the organism at the species level. Use of the NCBI taxon ID is recommended.
Example			doi:10.15454/1.4658436467893904E12	INRA:W95115_inra ICNF:PNB-RPI	INRA:W95115_inra_2001; INRA:inra_kernel_2351; Rothamsted:res_GK090847	NCBITAXON:4577
Format			DOI	Unique identifier	Unique identifier	Unique identifier
	A3_H	inra		inra:A3	A3_H	NCBITAXON:4577
	A310_H	inra		inra:A310	A310_H	NCBITAXON:4577




Exemple : <https://doi.org/10.15454/IASSTN>

Specialized (Plant) Metadata

Metadata File : Biological Material, Metadata CSV/XLS template

11 à 11 de 11 Fichiers

 **8-Info_Maize_variety.tab**
Données tabulaires - 25.4 Ko - 27 mars 2019 - 338 téléchargements
13 Variables, 256 Observations - UNF:6:CvjCsNsZyIZz3bYvvbriA==
This file contains the description of the genotypes. Briefly, all studied hybrids result from a F1 cross between the donor dent lines (parent 1) and one flint tester (parent 2, UH007 from University of Hohenheim) (Negre et al. 2018, <https://doi.org/10.1101/476598>). The file cont

Prévisualisation **Métadonnées** Versions

Explorer de View Data

	Variety_ID	Accession_ID	accession_holding	parent1	parent1_synonym	parent1_holding	parent2	parent
1	11430	11430_H	inra	11430_usda	11430_usda	USDA	UH_007	UH
2	A3	A3_H	inra	A3_inra	A3_inra	INRA	UH_007	UH
3	A310	A310_H	inra	A310_inra	A310_inra	INRA	UH_007	UH
4	A347	A347_H	inra	A347_inra	A347_inra	INRA	UH_007	UH
5	A374	A374_H	inra	A374_inra	A374_inra	INRA	UH_007	UH
6	A375	A375_H	inra	A375_inra	A375_inra	INRA	UH_007	UH
7	A554	A554_H	inra	A554_inra	A554_inra	INRA	UH_007	UH
8	AS5707	AS5707_H	inra	AS5707_usda	AS5707_usda	USDA	UH_007	UH
9	B100	B100_H	inra	B100_usda	B100_usda	USDA	UH_007	UH



En accord

- Avec le bailleur, l'institution, les partenaires, la revue de publication Entrepôts certifiés ou recommandés

Adapté à ses besoins

- Reconnu dans sa discipline
- Volume de données
- Accessibilité des données (possibilité de contrôler l'accès aux données ou non)
- Budget (la plupart sont gratuits ou avec un coût raisonnable)

FAIR

- Délivrant un identifiant numérique unique et pérenne
- Permettant de choisir la licence de diffusion
- Métadonnées toujours accessibles publiquement



Cross-disciplinary repositories

- > [Dryad Digital Repository](#)
- > [figshare](#)
- > [Harvard Dataverse Network](#)
- > [Kaggle](#)
- > [Network Data Exchange \(NDEx\)](#)
- > [Open Science Framework](#)
- > [Zenodo](#)

Repositories by type

Biochemistry	Neuroscience	Social Sciences
Biomedical Sciences	Omics	Structural Databases
Marine Sciences	Physical Sciences	Taxonomic & Species Diversity
Model Organisms	Sequencing	Unstructured and/or Large Data

<https://journals.plos.org/plosone/s/recommended-repositories>



Data Repository Guidance

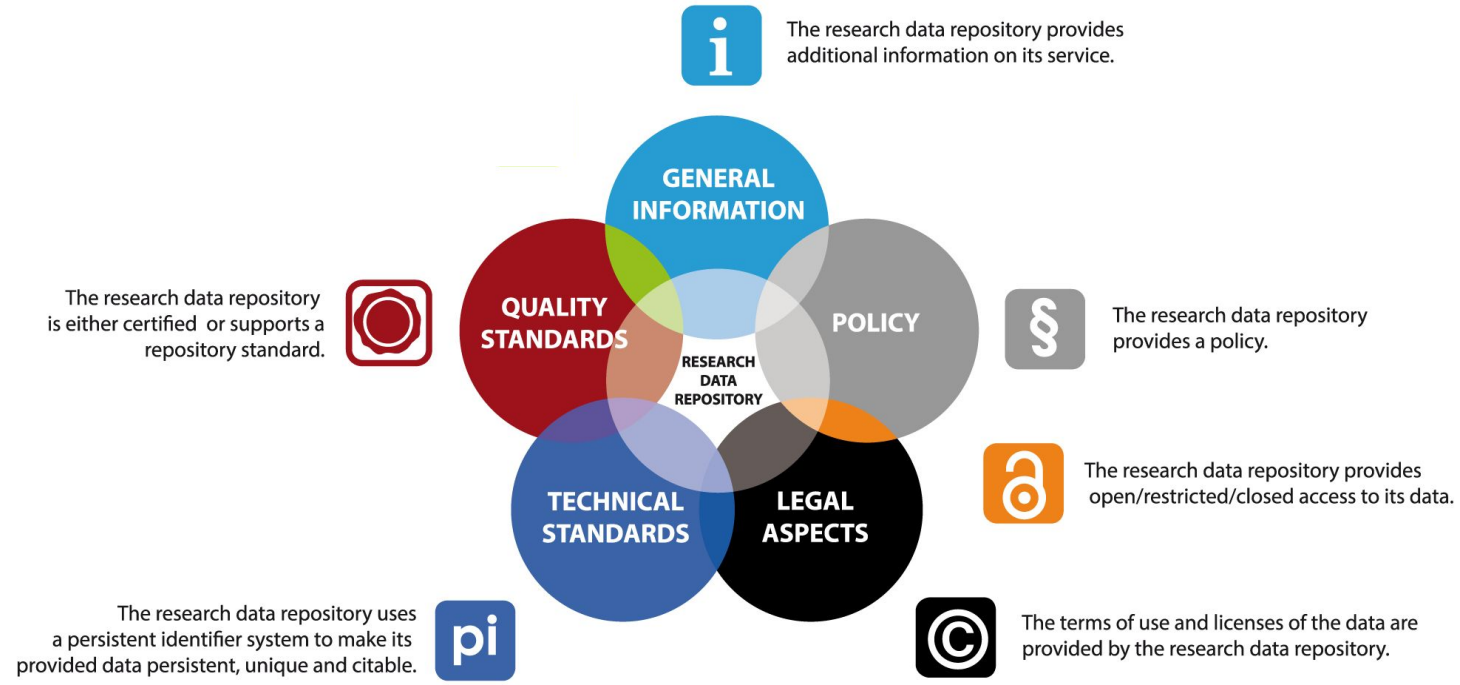
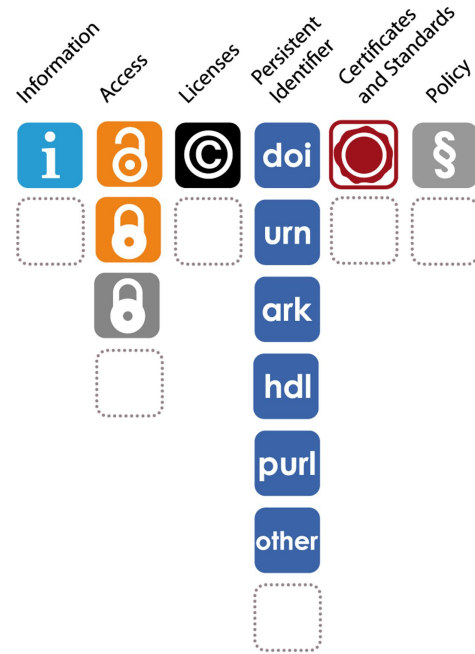
- **Biological sciences:** Nucleic acid sequence; Protein sequence; Molecular & supramolecular structure; Neuroscience; Omics; Taxonomy & species diversity; Mathematical & modelling resources; Cytometry and Immunology; Imaging; Organism-focused resources
- **Health sciences**
- **Chemistry and Chemical biology**
- **Earth, Environmental and Space sciences:** Broad scope Earth & environmental sciences; Astronomy & planetary sciences; Biogeochemistry and Geochemistry; Climate sciences; Ecology; Geomagnetism & Palaeomagnetism; Ocean sciences; Solid Earth sciences
- **Physics**

<https://www.nature.com/sdata/policies/repositories>

Deux répertoires d'entrepôts :



↳ Système d'icônes :

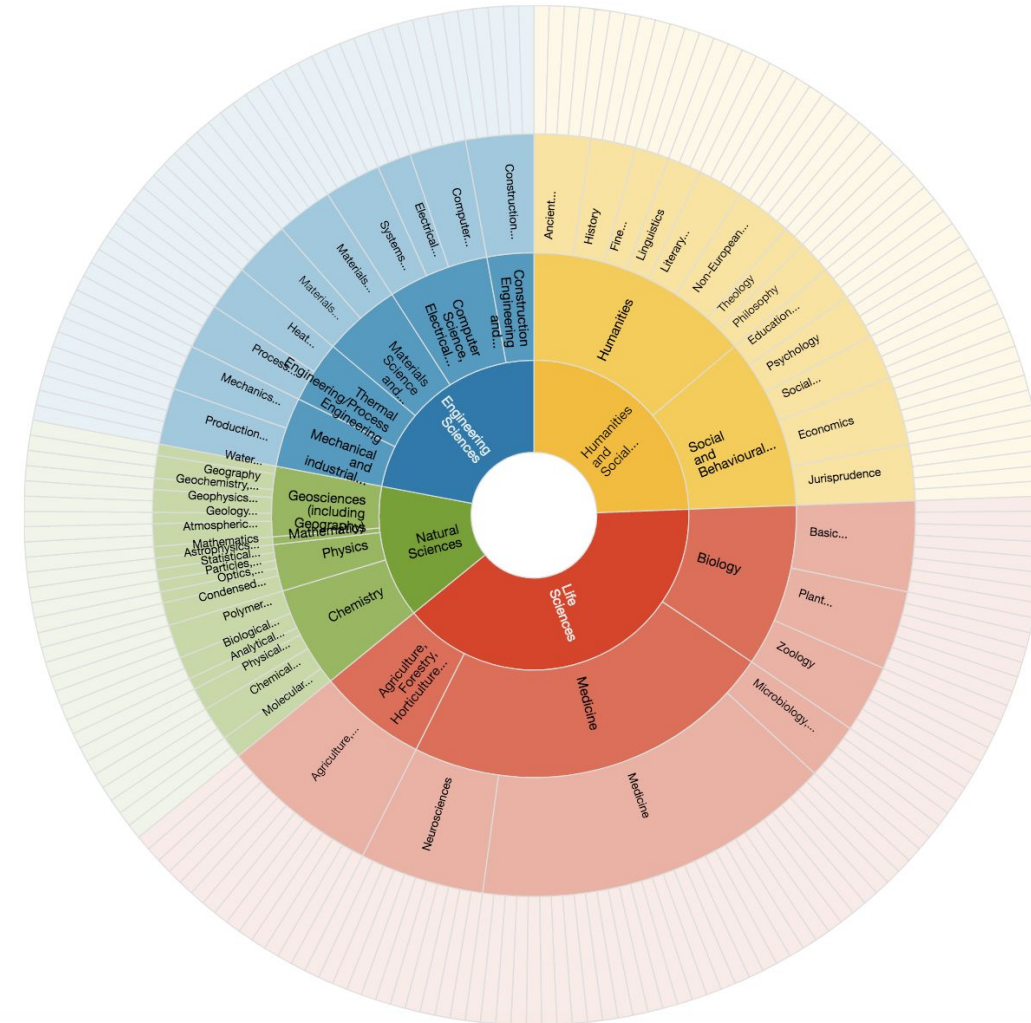


Images : <https://doi.org/10.1371/journal.pone.0078080>



Browse by subject

- Graphical
- Text

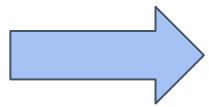




En fonction de vos jeux de données à partager

Recherchez un entrepôt, à partir de

- Re3data (<https://www.re3data.org/>)
- Revues scientifiques dans vos domaines



Essayez d'identifier si cet entrepôt correspond à vos besoins

- licences proposées
- modalités de dépôt (standard de métadonnées, format, taille des fichiers...)



- Publier un **data paper**
- Publier le **PGD**
- Publier un **article de recherche**
- Rédiger une brève pour un **magazine** spécialisé
- Contribuer à un **blog**,



Open Data Journal for Agricultural Research



[Data life cycle](#)[Your role](#)[Your domain](#)[Your tasks](#)[Tool assembly](#)[All tools and resources](#)[All training resources](#)

Are you working with data in the Life Sciences? Do you feel overwhelmed when you think about Research Data Management?

The ELIXIR Research Data Management Kit (RDMkit) is an online guide containing good data management practices applicable to research projects from the beginning to the end. Developed and managed by people who work every day with life science data, the RDMkit has guidelines, information, and pointers to help you with problems throughout the data's life cycle. RDMkit supports FAIR data — Findable, Accessible, Interoperable and Reusable — by-design, from the first steps of data management planning to the final steps of depositing data in public archives.

The RDMkit organises information into the six sections displayed below, which are interconnected but can be browsed independently.

Data life cycle

Start here to get an overview of research data management. Click on a section of the diagram below to get an introduction to that stage of the data management life cycle.





Merci !

Remerciements :

Toute l'équipe de formation @ IFB

ELIXIR-CONVERGE

