

***Département du Système d’Information***

|  |  |  |
| --- | --- | --- |
| ***CONTEXTE*** |  | **Entrepôt de la Donnée Inserm (EDI)** |
| ***SUJET*** |  | **Description du service** |
| ***mis à jour le*** |  | 06/05/2024 |
| ***par*** |  | L’équipe EDI (DSI/SSDUN/DomIS)V9 |

SOMMAIRE

1 Introduction 3

2 Contexte du projet et du service 3

2.1 Contexte général 3

2.2 Périmètre du projet et mode opératoire 4

2.3 Positionnement de l’entreposage des données dans le cycle de vie de la donnée 4

2.4 Distinguer dépôt et référencement 5

3 Caractéristiques de la plate-forme EDI 5

3.1 Eligibilité 6

3.1.1 Eligibilité des jeux de données 6

3.1.2 Eligibilité des déposants 6

3.2 Organisation des collections 7

3.2.1 Sur la plate-forme Recherche Data Gouv (RDG) 7

3.2.2 Sur EDI 7

3.3 La curation des jeux de données 8

3.3.1 Mode de curation 8

3.3.2 Schémas de métadonnées 8

3.3.3 Améliorer la réutilisabilité des données 8

3.4 La gestion des habilitations 9

3.4.1 Habilitation des déposants 9

3.4.2 Habilitation des consultants 9

3.4.3 Responsabilités 9

3.5 L’ouverture des données 9

3.6 Divers 10

3.6.1 Les modes de connexion 10

3.6.2 Les identifiants uniques 10

4 Utilisation d’EDI 10

4.1 Les prérequis au dépôt 10

4.2 Les principales étapes du dépôt 11

4.3 Suites au dépôt, 11

4.3.1 Curation par l’administrateur de la collection 11

4.3.2 Cas de retrait d'un jeu de données publié 11

4.4 Assistance – Accompagnement 11

 Données personnelles 12

 Données de santé 12

 Données sensibles 12

 Données personnelles sensibles 13

 Données à caractère personnel concernant la santé 13

 Données de recherche 13

Annexes

Annexe 1 : Définitions

Annexe 2 : Conditions générales de Recherche Data Gouv (extraits)

Annexe 3 : Abréviations

Annexe 4 : Charte de la curation

Annexe 5 : Les schémas de métadonnées de RDG

***Attention*** *: la CNIL utilise la notion d' "entrepôt de données de santé" (EDS, cf.* [*https://www.cnil.fr/fr/la-cnil-adopte-un-referentiel-sur-les-entrepots-de-donnees-de-sante*](https://www.cnil.fr/fr/la-cnil-adopte-un-referentiel-sur-les-entrepots-de-donnees-de-sante)), *décrite ainsi :*

*"L’entrepôt vise à collecter des données massives sur une période prolongée pour une réutilisation des données dans plusieurs projets de recherche. La recherche, qui prévoit une collecte de données particulières, répond à un objectif scientifique précis et est limitée dans le temps."*

*EDI n'est pas un "entrepôt de données de santé", pour plusieurs raisons : il ne peut pas contenir de données personnelles, de données "chaudes" ou de données massives et il n'est pas réservé à un seul responsable de traitement.*

# Introduction

Le projet EDI fait partie de la politique nationale et du plan stratégique de l’Inserm. Il s’inscrit dans le contexte général de l’ouverture des données de la recherche et s’adressera aux différentes structures labellisées Inserm. Il a pour but de mettre en valeur et d’assurer la visibilité ainsi que la sécurité des données de la recherche à l’Inserm. Le projet EDI s’inscrit dans l’action 48 du COMP (Contrat d’objectifs, de moyens et de performance 2021-2025) : « Concevoir un environnement numérique complet au plus près du cycle de vie de la donnée ».

Le présent document décrit le service apporté par l’entrepôt de la Donnée Inserm. Il est complété par d’autres documents de référence :

* « EDI - Guide du déposant »
* « EDI - Guide de l’administration et de la curation »

Les annexes 1 et 3 présentent la définition de termes utilisés dans le présent document ainsi que des abréviations.

# Contexte du projet et du service

## Contexte général

 L’entrepôt de la données Inserm (EDI) a pour objectif principal de fournir à la communauté de recherche de l’Inserm un espace de publication, référencement et mise à disposition de ses données de recherche. Il s’inscrit ainsi dans une démarche visant à augmenter la dimension FAIR (« Trouvable, accessible, interopérable et réutilisable ») des données de recherche produites au sein de l’institut.

Si de nombreux jeux de données seront amenés à être déposés sur l’outil, l’un des objectifs d’EDI est également de permettre le simple référencement de données pouvant être physiquement stockées sur une autre infrastructure, permettant ainsi de rendre « trouvables » des données nécessitant une protection particulière et ne pouvant être totalement ouvertes. En ce sens, dans la terminologie choisie, « entrepôt » correspond à la définition anglaise de *Data Repository* et non de *Data Warehouse.*

 La phase d’étude du projet a été menée au courant de l’année 2021. Cette phase a permis d’identifier les besoins majeurs de l’Inserm en termes d’entrepôt, ainsi que les principales pistes de déploiement et de mise en œuvre. Cette étude a été conclue par la remise d’un rapport en avril 2022[[1]](#footnote-1) dont la préconisation principale est d’envisager la mise en place de EDI comme entrepôt institutionnel au sein de la plateforme ministérielle Recherche Data Gouv (RDG).

L’entrepôt de la donnée Inserm est amené à faire partie, à terme, de l’environnement numérique du chercheur (ENC), tel que défini dans l’action 48 (objectif 8) du Contrat d’Objectifs, de Moyens et de Performances entre l’Inserm et l’état 2021-2025. A ce titre, il sera en interaction avec plusieurs de ses autres composantes : cahier de laboratoire électronique (CLE), offre de stockage, PGD (plan de gestion de données), etc.

## Périmètre du projet et mode opératoire

La cible initiale du projet est constituée de l’ensemble des personnels scientifiques des structures Inserm (chercheurs et ITA), quel que soit l’organisme qui les rémunère ; pour ce qui est des sources de données, on visera en priorité celles issues de projets de recherche dont l'Inserm est le financeur ou le promoteur (dans le cas de recherches impliquant la personne humaine).

Cette cible peut évoluer et le comité de pilotage devra préciser ces règles au cours de la phase pilote du projet.

EDI est un projet conçu pour mieux répondre aux besoins des chercheurs Inserm en terme de publication et d’accès aux différents jeux de données.

Ne font pas partie du projet : le stockage chaud pour les données en début de cycle de vie, l’archivage des données froides, la gestion des données sensibles.

La mise à disposition technique d’EDI a été accompagnée d’une phase pilote. Durant cette phase, l’outil était disponible et accessible à un nombre restreint d’utilisateurs volontaires. L’objectif de cette phase était de valider et améliorer l’ensemble des règles, propositions et contraintes fonctionnelles décrites dans ce document.

## Positionnement de l’entreposage des données dans le cycle de vie de la donnée



## Distinguer dépôt et référencement

L’offre RDG comprendra à terme 5 services, dont un catalogue, annoncé pour 2024 :



(cf. <https://recherche.data.gouv.fr/fr/page/kit-communication-recherche-data-gouv>)

En attendant sa mise à disposition ou celle d’un catalogue propre à l’Inserm (projet InsermData), il est possible de créer un jeu de données dans EDI sans y déposer de fichier mais en y indiquant comme les données peuvent être accédées. On parle alors de « référencement » et non de « dépôt ».

# Caractéristiques de la plate-forme EDI

L’Entrepôt de la Donnée Inserm est hébergé, comme les entrepôts institutionnels d’autres entités du domaine enseignement supérieur et recherche, par la plateforme ministérielle Recherche Data Gouv (RDG). Celle-ci repose sur la technologie ouverte [Dataverse](https://dataverse.org/) et est opérée par l’INRAe.

## Eligibilité

### Eligibilité des jeux de données

Pour pouvoir y créer son entrepôt institutionnel, l’Inserm a pris une série d’engagements vis-à-vis de la plate-forme RDG :

* Ne déposer que des données qui ne peuvent pas bénéficier d’un entrepôt disciplinaire (thématique) (cf. <https://doranum.fr/depot-entrepots/depot-et-entrepots-fiche-synthetique_10_13143_a3d4-7553/>)
* Ne déposer que des données qui ont fait l’objet d’un soin particulier (cf. <https://recherche.data.gouv.fr/fr/categorie/9/guide/avant-de-deposer>)
* Respecter les règles établies par les [conditions générales d’utilisation](https://recherche.data.gouv.fr/fr/page/conditions-generales-dutilisation)de l'entrepôt Recherche Data Gouv
* En particulier, ne pas déposer de [données sensibles](https://www.cnil.fr/fr/definition/donnee-sensible) (notamment des données à caractère personnel), à moins que leur sensibilité n’ait été modérée par des procédés du type anonymisation (cf. annexes 1 et 2)

Déposer des données dans EDI suppose donc de se plier à ces obligations.

On vérifiera également que les données peuvent être partagées en prenant en compte les aspects contractuels ou réglementaires (existence de co-auteurs, par exemple).

EDI recevra des données issues de projets de recherche dont l'Inserm est le financeur ou le promoteur (dans le cas de recherches impliquant la personne humaine), si celles-ci ne peuvent pas être déposées dans un entrepôt disciplinaire (ou thématique).

Ce service ne prend pas en compte : le stockage chaud pour les données en début de cycle de vie (cf. schéma du paragraphe 2.3), l’archivage des données froides, la gestion des données sensibles. Les données sensibles peuvent néanmoins être référencées dans l’entrepôt ; c’est-à-dire qu’on crée le jeu de données dans EDI, qu’on n’y dépose pas de données mais qu’on y indique leur emplacement.

Deux types particuliers de fichiers n’ont *a priori* pas leur place dans un dépôt dans RDG :

* les plans de gestion de données (PGD), qui sont à déposer dans HAL-Inserm,
* les codes, scripts ou logiciels, qui sont néanmoins "tolérés" dans EDI en attendant la mise en place par l’Inserm d’une offre *ad hoc* (projet « Usine logicielle »).

Les dépôts disciplinaires (ou thématiques)

Y faire appel permettra de limiter le nombre de jeux de données déposés dans EDI, tout en respectant un engagement pris vis-à-vis de RDG. Pour trouver un dépôt adapté, RDG fait référence au site <https://www.re3data.org/> : « registry of research data repositories »).

Les instituts thématiques devraient pouvoir aider à trouver ceux adaptés à leur domaine.

On peut utilement se référer aux pages suivantes :

<https://www.ouvrirlascience.fr/selectionner-un-entrepot-thematique-de-confiance-pour-le-depot-de-donnees-methodologie-et-analyse-de-loffre-existante/>

<https://doranum.fr/depot-entrepots/depot-et-entrepots-fiche-synthetique_10_13143_a3d4-7553/>

### Eligibilité des déposants

EDI est destiné à l’ensemble des personnels scientifiques des structures Inserm (chercheurs et ITA), quel que soit l’organisme qui les rémunère.

C’est au PI du projet de recherche ou au responsable de la structure de faire le choix du déposant, en s’aidant des éléments de réflexion suivants :

* se référer, le cas échéant, à ce qui est indiqué à ce sujet dans le PGD du projet.
* choisir de préférence un statutaire (chercheur ou ITA), pour pouvoir garder la trace des interventions sur les données.
* choisir la personne qui connait le mieux les données, puisque c'est à elle que seront soumises les éventuelles demandes d'amélioration (pendant la phase de curation).
* prendre en compte le sens de la propriété des données, pour que le principal auteur des données ne s’en sente pas dépossédé.

## Organisation des collections

### Sur la plate-forme Recherche Data Gouv (RDG)

*RDG repose en premier lieu sur les principes de Dataverse* :

* L’organisation est hiérarchique. Au sommet de l'arborescence se trouve un "dataverse".
* Sous tout dataverse, on peut avoir entre 0 et n "dataverses" et entre 0 et n "datasets".
* Un « dataset » doit être rattaché directement à un « dataverse » et un seul ; établir des liens avec d’autres « dataverses », sans duplication des données, est possible.

*Mais également sur des choix de l'équipe gérant RDG* :

* Un dataverse est appelé une **collection** ; un dataset est appelé un **jeu de données**.
* Sous le sommet de l'arborescence sont créées des collections constituant les entrepôts institutionnels ; il n'y a pas de jeux de données à ce niveau. Est créée également une collection « Entrepôt Générique », pour les institutions ne disposant pas d’un entrepôt institutionnel ; cette collection ne contient pas de sous-collection.
* Chaque administrateur d'un entrepôt institutionnel peut choisir l'organisation de celui-ci.
* Il n’a pas été mis de limite à la profondeur de l’arborescence.

### Sur EDI

Un premier niveau de collections a été mis en place, par thématique scientifique. Ceci doit aider le "producteur de données" à trouver le bon point de dépôt et le "consultant de l'entrepôt" à trouver le jeu de données qui l'intéresse.

Les intitulés des instituts thématiques de l’Inserm ont été repris :

* Bases moléculaires et structurales du vivant
* Biologie cellulaire, développement et évolution
* Cancer
* Génétique, génomique et bioinformatique
* Immunologie, inflammation, infectiologie et microbiologie
* Neurosciences, sciences cognitives, neurologie, psychiatrie
* Physiopathologie, métabolisme, nutrition
* Santé publique
* Technologies pour la santé

Comme pour RDG, on dispose, en complément, d’une « Collection générique Inserm », pour les jeux de données ne trouvant pas leur place dans une collection par IT. A l’image de la collection générique de l’entrepôt RDG, cette collection ne contiendra pas de sous-collections.

Sous le niveau parent des collections thématiques, les utilisateurs peuvent demander la création de sous-collections. Le demandeur d’une sous-collection en deviendra le curateur.

*Procédure* : demander, à l’adresse support-edi.dsi@inserm.fr, la création de la sous-collection, en en précisant la destination, le nom souhaité et la collection thématique de rattachement. L’administrateur de la collection concernée instruira la demande.

Le schéma ci-dessous représente l’architecture mise en place :



## La curation des jeux de données

### Mode de curation

L’Inserm a décidé de suivre les recommandations de RDG en la matière : le déposant effectue la mise en qualité (curation) des données qu’il souhaite publier puis l’organisme propriétaire de l’entrepôt effectue un contrôle *a minima*.

Les métadonnées et les fichiers répondent aux normes de qualité.

La curation *a posteriori* (par le curateur de la collection*)* interviendra au fil de l’eau, au fur et à mesure de l’arrivée des notifications de publication des jeux de données.

La Charte de curation des métadonnées d’EDI, présentée en annexe 4, traite de la définition, des niveaux de la curation et du processus en jeu.  Elle se termine par les règles de saisie des métadonnées, sur lesquelles portera le contrôle effectué par le curateur.

### Schémas de métadonnées

RDG propose un certain nombre de champs de métadonnées disciplinaires ; cf. annexe 5. L’Inserm a choisi de rendre accessible celui relatif aux sciences de la vie (sans que son renseignement soit obligatoire).

D’autres champs disciplinaires, présents dans RDG, peuvent être activés ; de nouveaux champs pourraient éventuellement être ajoutés à RDG pour le bénéfice de nos collections. Ceci suppose que des demandes précises soient faites par différents déposants.

### Améliorer la réutilisabilité des données

Plusieurs méthodes ou outils peuvent aider à rendre les dépôts plus interopérables et réutilisables :

* remplir les métadonnées au-delà des champs obligatoires, on facilite la découverte des données, ce qui les rend plus interopérables et réutilisables.
* utiliser le modèle de jeu de données proposé par défaut, dans lequel des champs sont pré-remplis. D’autres modèle peuvent être créés pour répondre à des besoins spécifiques et récurrents (envoyer un mail à support-edi@inserm.fr).
* décrire en détail les fichiers, les données qu’ils contiennent et la manière dont elles ont été produites ; ceci peut se faire à partir de modèles mis à disposition depuis l’intranet de l’Inserm (<https://pro.inserm.fr/rubriques/support-a-la-recherche/informatique-scientifique/lentrepot-de-donnees-inserm-edi>). Ce type de fichier de description est appelé « readme » (lisez-moi), par analogie avec ce qui se fait dans le développement informatique.

## La gestion des habilitations

La gestion des habilitations (qui a un compte et quels droits sont attachés à ce compte) concerne 2 temps de la vie du jeu de données (jdd) dans l'entrepôt :

- avant la publication d'un jdd : droit pour les déposants de soumettre pour publication un jdd dans une collection donnée.

- après publication : droit pour les "consultants" de consulter et télécharger un jdd, avec ou sans authentification.

### Habilitation des déposants

Le droit "créateur de jeu de données", qui donne la possibilité de créer un jdd puis de le publier, est octroyé par défaut à tout détenteur d’une adresse de messagerie @inserm.fr

### Habilitation des consultants

C'est le déposant qui décide du niveau d'ouverture du jeu de données qu'il publie : cf. paragraphe 3.5

### Responsabilités

Le DSI est responsable du bon fonctionnement de l’entrepôt.

C’est le Département de la Science Ouverte (DSO) qui assure la curation des jeux de données.

## L’ouverture des données

Comme pour le mode de curation, l’Inserm suit la recommandation de RDG :

* en utilisant les licences Creative Commons (<https://creativecommons.org/>, ONG fournissant le moyen de libérer ses œuvres des droits de propriété intellectuelle standard de son pays),
* en proposant par défaut la licence « etalab 2.0 » (<https://www.etalab.gouv.fr/licence-ouverte-open-licence/>, qui correspond à la licence CC-BY (attribution).

Un décret liste les types de licence pour la réutilisation des données des administrations : cf. <https://www.data.gouv.fr/fr/pages/legal/licences/> (il est question d’un ajout prochain à cette liste des licences Creative Commons).

Il appartient au déposant d’évaluer le degré d’ouverture qu’il entend donner à son jeu de données. Il peut ainsi détailler les "Conditions d'utilisation" de son jeu de données, au moyen des rubriques suivantes :

* Déclaration de confidentialité,
* Autorisations spéciales,
* Restrictions,
* Exigences de citation,
* Exigences du déposant,
* Conditions,
* Avis de non-responsabilité

et/ou déclarer ses "Fichiers en accès réservé + Conditions d'accès " :

* Conditions d'accès,
* Autoriser la demande d'accès,
* Emplacement des données,
* Dépôt original,
* État de disponibilité,
* Contact pour les demandes d'accès,
* Taille de la collection de données,
* Éléments achevés de l'étude.

Il est ainsi possible d’assujettir l’accès à un jeu de données à une demande d’accès.

## Divers

### Les modes de connexion

RDG propose 3 modes de connexion :

1) avec l'adresse de messagerie @inserm.fr

2) avec un ORCID (identifiant du chercheur).

3) avec un compte propre à RDG.

Seule l'utilisation du premier mode est prise en compte pour l’accès à EDI

*URL de connexion* : <https://entrepot.recherche.data.gouv.fr/loginpage.xhtml?redirectPage=dataverse.xhtml>

Cliquer sur le lien "Se connecter" en haut à droite.

### Les identifiants uniques

Comme représenté dans le schéma ci-dessous, RDG attribue un identifiant unique à tout jeu de données qui y est publié, que celui-ci contienne des fichiers ou non. C’est le [système DOI](https://www.doi.org/), largement utilisé dans l’ESR, qui a été choisi à cet effet.

# Utilisation d’EDI

Les procédures sont détaillées dans le document « EDI – Guide du déposant ».

## Les prérequis au dépôt

A faire avant de considérer le dépôt :

* S’assurer que les données n’ont pas été déposées dans un autre entrepôt.
* Vérifier que les données peuvent être partagées, en tenant compte :
* de l’accord des co-auteurs
* des contrats ou réglementations
* de la présence de données sensibles
* Considérer la sensibilité des données
* Envisager les solutions alternatives

A faire une fois le dépôt décidé :

* Choisir la collection ; on déposera par défaut dans la collection correspondant à la thématique principale de la structure de recherche
* Choisir la licence (conditions de réutilisation) ; par défaut : « licence ouverte » du MESR = CC BY
* Assurer la qualité des métadonnées et des fichiers (nommage, format, cohérence)

## Les principales étapes du dépôt

* Préparer le jeu de données : respect des règles, rassembler les métadonnées (citation, « life science »)
* Se connecter à RDG
* Sélectionner l’entrepôt Inserm puis la collection choisie
* Créer le jeu de données et le décrire, dans le respect des engagements de l’Inserm sur les métadonnées de citation et de fichiers
* Le cas échéant (dépôt et non référencement), télécharger les données proprement dites :
	+ avec la méthode classique via le navigateur
	+ en ligne de commande, pour un nombre de fichiers important
* Publier

Un guide pratique, détaillé, est disponible : « EDI – Guide du déposant ».

## Suites au dépôt,

### Curation par l’administrateur de la collection

L’administrateur de la collection (au DSO) est prévenu par mail de la publication d’un jeu de données ; il doit alors assurer la curation de celui-ci.

### Cas de retrait d'un jeu de données publié

Voici quelques raisons possibles du retrait, temporaire ou définitif, d’un jeu de données par l’administrateur de la collection :

* Eligibilité : il n'est pas issu d'une recherche financée par l'Inserm ou menée dans une structure labellisée Inserm.
* Qualité : il ne répond pas aux normes de qualité imposées par RDG et reprises par l'Inserm.
* Condition de réutilisation : le niveau d'accès demandé ne correspond pas à la sensibilité des données.

En cas de retrait de la publication :

* temporaire : on indiquera les changements à apporter au jdd
* définitif : on orientera vers une autre collection pour le dépôt (dans EDI, dans RDG, dans un entrepôt disciplinaire ou autre)

Les jeux de données retirés de la publication ne sont pas effacés de la plate-forme mais rendus invisibles par l’internaute.

## Assistance – Accompagnement

1. Consulter la documentation mise à disposition par la plate-forme RDG : <https://recherche.data.gouv.fr/fr/guide/ou-deposer-et-publier-ses-donnees>

ou par l’Inserm sur son intranet : <https://pro.inserm.fr/rubriques/support-a-la-recherche/informatique-scientifique/lentrepot-de-donnees-inserm-edi>

1. A défaut, écrire à l’adresse support-edi.dsi@inserm.fr ; des échanges directs peuvent être proposés.

Annexes

Annexe 1 : Définitions

## Données personnelles

Toute information se rapportant à une personne physique identifiée ou identifiable. Mais, parce qu’elles concernent des personnes, celles-ci doivent en conserver la maîtrise.

Une personne physique peut être identifiée :

⦁ directement (exemple : nom et prénom) ;

⦁ indirectement (exemple : par un numéro de téléphone ou de plaque d’immatriculation, un identifiant tel que le numéro de sécurité sociale, une adresse postale ou courriel, mais aussi la voix ou l’image).

L’identification d’une personne physique peut être réalisée :

⦁ à partir d’une seule donnée (exemple : nom) ;

⦁ à partir du croisement d’un ensemble de données (exemple : une femme vivant à telle adresse, née tel jour et membre dans telle association).

Par contre, des coordonnées d’entreprises (par exemple, l’entreprise « Compagnie A » avec son adresse postale, le numéro de téléphone de son standard et un courriel de contact générique « compagnie1[@]email.fr ») ne sont pas, en principe, des données personnelles.

*Source : CNIL,* [*https://www.cnil.fr/fr/definition/donnee-personnelle*](https://www.cnil.fr/fr/definition/donnee-personnelle)

## Données de santé

Trois catégories de données :

* celles qui sont des données de santé par nature : antécédents médicaux, maladies, prestations de soins réalisés, résultats d’examens, traitements, handicap, etc.
* celles, qui du fait de leur croisement avec d’autres données, deviennent des données de santé en ce qu’elles permettent de tirer une conclusion sur l’état de santé ou le risque pour la santé d’une personne : croisement d’une mesure de poids avec d’autres données (nombre de pas, mesure des apports caloriques…), croisement de la tension avec la mesure de l’effort, etc.
* celles qui deviennent des données de santé en raison de leur destination, c’est-à-dire de l’utilisation qui en est faite au plan médical.

*A noter : la loi ne s’applique pas aux traitements qui comporteraient des données de santé à l’usage exclusif de la personne. A titre d’exemple, la loi ne s’applique pas aux applications mobiles en santé qui proposent dans leurs fonctionnalités, la collecte, l’enregistrement ou la conservation de données à condition que ces opérations s’effectuent localement sur un ordinateur, un ordiphone ou une tablette, sans connexion extérieure et à des fins exclusivement personnelles.*

N’entrent pas dans la notion de données de santé celles à partir desquelles aucune conséquence ne peut être tirée au regard de l’état de santé de la personne concernée (ex : une application collectant un nombre de pas au cours d’une promenade sans croisement de ces données avec d’autres).

*Source : CNIL,* [*https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante*](https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante)

## Données sensibles

L’instruction interministérielle n°901/SGDSN/ANSSI du 28 janvier 2015 définit un système d’information sensible comme traitant d'informations dont la divulgation à des personnes non autorisées, l'altération ou l'indisponibilité sont de nature à porter atteinte à la réalisation des objectifs des entités qui le mettent en œuvre.

*Source : ANSSI,* [*https://cyber.gouv.fr/instruction-interministerielle-n901*](https://cyber.gouv.fr/instruction-interministerielle-n901)

## Données personnelles sensibles

Informations qui révèlent la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique.

Le règlement européen interdit de recueillir ou d’utiliser ces données, sauf, notamment, dans les cas suivants :

* si la personne concernée a donné son consentement exprès (démarche active, explicite et de préférence écrite, qui doit être libre, spécifique, et informée) ;
* si les informations sont manifestement rendues publiques par la personne concernée ;
* si elles sont nécessaires à la sauvegarde de la vie humaine ;
* si leur utilisation est justifiée par l'intérêt public et autorisé par la CNIL ;
* si elles concernent les membres ou adhérents d'une association ou d'une organisation politique, religieuse, philosophique, politique ou syndicale.

*Source : CNIL,* [*https://www.cnil.fr/fr/definition/donnee-sensible*](https://www.cnil.fr/fr/definition/donnee-sensible)

## Données à caractère personnel concernant la santé

Données relatives à la santé physique ou mentale, passée, présente ou future, d’une personne physique (y compris la prestation de services de soins de santé) qui révèlent des informations sur l’état de santé de cette personne.

Cette définition comprend donc par exemple :

1. les informations relatives à une personne physique collectées lors de son inscription en vue de bénéficier de services de soins de santé ou lors de la prestation de ces services : un numéro, un symbole ou un élément spécifique attribué à une personne physique pour l’identifier de manière unique à des fins de santé ;

2. les informations obtenues lors du test ou de l’examen d’une partie du corps ou d’une substance corporelle, y compris à partir des données génétiques et d’échantillons biologiques ;

3. les informations concernant une maladie, un handicap, un risque de maladie, les antécédents médicaux, un traitement clinique ou l’état physiologique ou biomédical de la personne concernée (indépendamment de sa source, qu’elle provienne par exemple d’un médecin ou d’un autre professionnel de santé, d’un hôpital, d’un dispositif médical ou d’un test de diagnostic in vitro).

Cette définition permet d’englober certaines données de mesure à partir desquelles il est possible de déduire une information sur l’état de santé de la personne.

*Source : CNIL,* [*https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante*](https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante)

## Données de recherche

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 les résultats de la recherche.

*Source : OCDE,* [*https://www.oecd.org/fr/science/inno/38500823.pdf*](https://www.oecd.org/fr/science/inno/38500823.pdf)

Research data are the evidence that underpins the answer to the research question, and can be used to validate findings regardless of its form (e.g. print, digital, or physical). These might be quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, modelling, interview or other methods, or information derived from existing evidence. Data may be raw or primary (e.g. direct from measurement or collection) or derived from primary data for subsequent analysis or interpretation (e.g. cleaned up or as an extract from a larger data set), or derived from existing sources where the rights may be held by others. Data may be defined as ‘relational’ or ‘functional’ components of research, thus signalling that their identification and value lies in whether and how researchers use them as evidence for claims. They may include, for example, statistics, collections of digital images, sound recordings, transcripts of interviews, survey data and fieldwork observations with appropriate annotations, an interpretation, an artwork, archives, found objects, published texts or a manuscript.

*Source : OpenAIRE,* [*https://www.openaire.eu/how-do-i-know-if-my-research-data-is-protected*](https://www.openaire.eu/how-do-i-know-if-my-research-data-is-protected)

* **Anonymisation**

Traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible.

Pour construire un processus d’anonymisation pertinent, il est ainsi conseillé :

* d’identifier les informations à conserver selon leur pertinence.
* de supprimer les éléments d’identification directe ainsi que les valeurs rares qui pourraient permettre un ré-identification aisée des personnes (par exemple, la présence de l’âge des individus peut permettre de ré-identifier très facilement les personnes centenaires) ;
* de distinguer les informations importantes des informations secondaires ou inutiles (c’est-à-dire supprimables) ;
* de définir la finesse idéale et acceptable pour chaque information conservée.

Ce prérequis permet de déterminer le procédé d’anonymisation à appliquer, c’est-à-dire l’enchaînement des techniques d’anonymisation à mettre en place. Celles-ci peuvent être regroupées en deux familles : la randomisation et la généralisation.

La randomisation consiste à modifier les attributs dans un jeu de données de telle sorte qu'elles soient moins précises, tout en conservant la répartition globale. Cette technique permet de protéger le jeu de données du risque d’inférence (cf. ci-dessous).

Exemple : il est possible de permuter les données relatives à la date de naissance des individus de manière à altérer la véracité des informations contenues dans une base de données.

La généralisation consiste à modifier 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. Cette technique permet d’éviter l’individualisation d’un jeu de données. Elle limite également les possibles corrélations du jeu de données avec d’autres (cf. ci-dessous).

Il s’agit d’une solution, parmi d’autres, pour pouvoir exploiter des données personnelles dans le respect des droits et libertés des personnes.

*Source : CNIL,* [*https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles*](https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles)

* **Pseudonymisation**

Traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans information supplémentaire.

En pratique, la pseudonymisation consiste à remplacer les données directement identifiantes (nom, prénom, etc.) d’un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.).

La pseudonymisation permet ainsi de traiter les données d’individus sans pouvoir identifier ceux-ci de façon directe. En pratique, il est toutefois bien souvent possible de retrouver l’identité de ceux-ci grâce à des données tierces : les données concernées conservent donc un caractère personnel. L’opération de pseudonymisation est également réversible, contrairement à l’anonymisation.

La pseudonymisation constitue une des mesures recommandées par le RGPD pour limiter les risques liés au traitement de données personnelles.

*Source : CNIL,* [*https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles*](https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles)

* **Données à fort risque** : Données dont la divulgation à une personne non autorisée a un impact élevé sur la vie privée.
* **Données à faible risque** : Données dont la divulgation à une personne non autorisée a un impact limité sur la vie privée.

*Source :* [*SNDS - Guide\_accompagnement.pdf*](https://www.snds.gouv.fr/download/Guide_accompagnement.pdf)*, p.4*

* **Méta-données** :

Ensemble d’informations structurées qui décrit, explicite, localise une ressource informationnelle dans le but d’en faciliter la recherche, l’usage et la gestion.

*Source : NISO. Understanding Metadata, (2017) :* [*http://groups.niso.org/higherlogic/ws/public/download/17446/Understanding%20Metadata.pdf*](http://groups.niso.org/higherlogic/ws/public/download/17446/Understanding%20Metadata.pdf)

* **Jeu de données** (ou ensemble de données ou **dataset**)

Ensemble cohérent de données, brutes ou dérivées, produites dans le cadre d’un même projet, sur un même objet d’étude et/ou recueillies sur un même lieu.

*Un jeu de données défini ainsi est créé par une personne unique et est géré de la même façon (même méthode de traitement, de stockage, de partage, etc.)*

*Toutes les données d’un dataset peuvent donc être décrites avec une majorité de métadonnées communes.
Il se présente sous la forme d’un ou plusieurs fichiers électroniques, téléchargeables, citables (notamment par l’intermédiaire d’un DOI) et intelligibles – ce jeu étant accompagné des métadonnées descriptives suffisantes.*

*Un jeu de données peut être croisé avec d'autres jeux de données permettant la création de nouvelles connaissances.*

*Sources :
1)* [*https://data.ird.fr/definitions/*](https://data.ird.fr/definitions/) *;
2) W3C :* [*https://www.w3.org/TR/vocab-dcat-2/*](https://www.w3.org/TR/vocab-dcat-2/) *3) modèle de DMP de l'institut Pasteur
4) Gaillard R. De l’open data à l’open research data: quelle(s) politique(s) pour les données de la recherche. Janvier 2014. (Mémoire d’étude pour le diplôme de conservateur de bibliothèque))*

Annexe 2 : Conditions générales de Recherche Data Gouv (extraits)

Code de conduite et responsabilités des Utilisateurs

 Publication d’un jeu de données

 Règles générales

"[...] Les Contributeurs sont seuls responsables des données, métadonnées ou contenus qu’ils publient sur la plateforme Recherche Data Gouv."

 (Jeux de) Données à caractère personnel

Les Jeux de données contenant des données à caractère personnel, c’est-à-dire des données, y compris non nominatives, permettant la ré-identification de personnes physiques, ne peuvent pas être diffusés par la plateforme Recherche Data Gouv sauf si les personnes concernées ont donné leur accord ou si une disposition législative ou le décret n° 2018-1117 du 10 décembre 2018 à l’article L. 312-1-2 du CRPA le permet.

Le Contributeur publiant le Jeu de données est responsable des jeux de données qu’il publie sur la plateforme Recherche Data Gouv et s’assure que cette publication est conforme à la législation en vigueur. Il prend toute mesure nécessaire à cette fin et Entrepôt-Catalogue Recherche Data Gouv l’encourage à mentionner la présence de données à caractère personnel dans la documentation accompagnant le Jeu de données et préciser, lorsqu’elles existent, les restrictions juridiques à la Réutilisation.

Le Réutilisateur de telles données doit se conformer à la législation relative à la protection des données à caractère personnel en vigueur dans son territoire de résidence et respecter les stipulations pertinentes de la licence attachée à ces données.

 Données couvertes par des droits de propriété intellectuelle

Lorsque le Contributeur publie un jeu de données comprenant des droits de propriété intellectuelle qui font obstacle à la Réutilisation, il mentionne leur présence dans la documentation accompagnant le Jeu de données. Il indique l’identité de la personne physique ou morale titulaire de ces droits ou, si celle-ci n’est pas connue, l’identité de la personne auprès de laquelle l’information en cause a été obtenue.

Annexe 3 : Abréviations

|  |  |  |
| --- | --- | --- |
| **Terme** | **Signification** | **Définition** |
| CC | Creative Commons | association fournissant le moyen de libérer ses œuvres des droits de propriété intellectuelle standard de son pays |
| COMP | Contrat d’objectifs, de moyens et de performance | cadre de développement de l'Inserm entre 2021 et 2025 |
| DOI | Digital object identifier | identifiant unique et persistant d’objet numérique |
| DomIS | Domaine Informatique Scientifique | composant du SSDUN |
| DSO | Département de la Science Ouverte | composant de la direction de l'Inserm |
| DSI | Département du Système d'Information | composant de la direction de l'Inserm |
| EDI | Entrepôt de la Donnée Inserm | projet du DSI de l'Inserm |
| ENC | environnement numérique du chercheur  | projet du DSI de l'Inserm |
| EPST | Etablissement Public à caractère Scientifique et Technique | catégorie d'établissements dépendant de l'administration centrale |
| ESR | Enseignement Supérieur et de la Recherche | domaine d'activité |
| Inist | Institut National de l’Information Scientifique et Technique (rattaché au CNRS) | EPST |
| INRAe | Institut National de Recherche pour l’Agriculture, l’alimentation et l’environnement | EPST |
| IT | Institut thématique | unité organisationnelle de l'Inserm |
| ITA | Ingénieurs, Techniciens, Administratifs | catégorie de personnel des établissements publiques de recherche |
| ORCID | Open Researcher and Contributor ID | identifiant numérique unique et persistent de personne |
| PGD | Plan de gestion de données (= DMP = data management plan) | document décrivant la manière dont les données sont gérées au cours d'un projet de recherche |
| RDG | Recherche Data Gouv | projet du Ministère de l'Enseignement Supérieur et de la Recherche |
| Renater | Réseau national de télécommunications pour la technologie, l'enseignement et la recherche | groupement d’intérêt public qui assure la maîtrise d’ouvrage du réseau national de communications électroniques pour la technologie, l’enseignement et la recherche. |
| SSDUN | Service Solutions et Développement des Usages Numériques | composant du DSI de l'Inserm |

Annexe 4 : Charte de curation

**1) Que désigne-t-on par Curation des données**

La curation est le processus visant à vérifier si l'ensemble de données est conforme aux règles établies par l'administrateur de la collection. Elle a pour but de mettre en évidence la valeur des données publiées, afin qu'elles soient trouvables, accessibles, interopérables et réutilisables (FAIR).

 Les principales actions de curation des données consistent principalement à

* Vérifier l'exhaustivité des métadonnées,
* Analyser et valider la documentation complémentaire qui sert à décrire le jeu de données publié
* Analyser la licence et les précisions sur les conditions d’utilisation des données.

Le curateur s’engage àassurer la curation de jeux de données en effectuant une vérification des métadonnées et des fichiers de données déposés dans la plateforme EDI. Il peut proposer des modifications pour améliorer la qualité documentaire des jeux de données, dans le respect de la charte des curateurs (cf. <https://recherche.data.gouv.fr/fr/categorie/15/guide/charte-des-curateurs>).

Les curateurs d’espaces institutionnels s’engagent pour chaque collection à :

* Veiller au respect de l’application des règles établies pour la gestion de l’entrepôt EDI
* Assurer la curation conformément au niveau défini dans la curation
* Traiter les jeux de données envoyés en révision dans les meilleurs délais possibles et assurer le suivi des modifications en lien avec le déposant jusqu’à la publication
* Ne pas modifier un jeu de données, en dehors de corrections mineures (syntaxe et forme), sans l’accord du déposant
* Traiter rapidement les demandes d’accès aux données
* Contribuer, avec l’administrateur, au repérage des jeux de données non publiés au-delà d’un délai de 12 mois afin d’éviter l’utilisation de la plateforme comme un espace de stockage et non de partage des données de la recherche
* Participer au réseau des curateurs de la plateforme Recherche Data Gouv afin d’harmoniser les pratiques et d’améliorer la qualité des données.

**2) Processus de curation**

La curation est réalisée par des curateurs désignés pour la collection. Elle peut avoir lieu avant ou après la publication du jeu de données en fonction du rôle attribué par défaut au créateur de jeu de données sur son jeu de données.

* Si un **rôle de collaborateur par défaut** a été attribué au créateur du jeu de données, celui-ci ne peut pas publier le jeu de données directement et doit le soumettre à un curateur de la collection pour validation et publication. La curation a lieu **avant la publication**.
* Si un **rôle de curateur par défaut** a été attribué au créateur du jeu de données, celui-ci peut publier le jeu de données directement. La curation par un curateur de la collection a lieu **après la publication**.

Pour les espaces institutionnels qui choisissent la méthode de curation avant publication, Le curateur effectue les corrections mineures (forme, syntaxe, orthographe) et suggère au déposant des améliorations à apporter au jeu de données. Plusieurs échanges avec le déposant peuvent être nécessaires avant la validation du jeu de données.

Un [rapport de curation](http://recherche.data.gouv.fr/uploads/general/Documents/Modele_RapportCuration.pdf) est transmis au déposant afin d’indiquer et suivre les modifications et suggestions du curateur.

**3) Les droits sur un jeu de données associés au rôle de curateur**

Un curateur possède différents droits sur un jeu de données, on cite

* Télécharger un fichier
* Consulter un jeu de données non publié
* Consulter la version provisoire d’un jeu de données
* Supprimer la version provisoire d’un jeu de données
* Modifier un jeu de données (modifications mineures)
* Publier un jeu de données
* Retirer un jeu de données publié de la diffusion
* Gérer les autorisations pour un jeu de données ou un fichier
* Générer une URL privée

**4) Délai de réalisation de la curation**

Quel que soit le processus opéré, la curation est réalisée dans les meilleurs délais. Cependant, lorsque la curation avant publication est choisie, le curateur veille à ce que le déposant procède aux modifications recommandées afin de publier le jeu de données dans un délai maximum d’un an.

Au-delà de ce délai, le jeu de données pourrait être supprimé par le curateur ou l’administrateur.

* 1. **La curation d’un jeu de données avant publication**

Lorsqu’un déposant soumet un jeu de données pour révision, le jeu de données à réviser apparaît dans la collection avec l’étiquette « En révision ». Le curateur est notifié par courriel et dans son [espace personnel](https://recherche.data.gouv.fr/fr/categorie/9/guide/la-creation-dun-compte#Espace+Personnel) dans l’entrepôt.

Le curateur procède aux vérifications des données selon le niveau de curation choisi. Il peut alors :

1. Soit réaliser des corrections mineures si nécessaire, publier le jeu de données et envoyer par courriel au déposant le rapport de curation mentionnant les corrections effectuées
2. Soit retourner le jeu de données au déposant via l’interface et suggérer des améliorations par le biais du rapport de curation envoyé par courriel.

Une fois le jeu de données publié, toute nouvelle modification apportée par le déposant devra à nouveau être validée par le curateur.

* 1. **La curation d’un jeu de données après publication**

La curation après publication porte sur des jeux de données publiés, elle nécessite donc une veille active de la part des curateurs. Les notifications reçues par mail aident à repérer les jeux de données publiés. Le suivi de la curation peut être assuré au moyen d’un fichier commun à la collection.

Une fois les jeux de données identifiés, le curateur peut :

1. Soit réaliser des corrections mineures si nécessaire, publier à nouveau le jeu de données et envoyer par courriel au déposant le rapport de curation mentionnant les corrections effectuées
2. Soit suggérer des améliorations par le biais du rapport de curation envoyé par courriel

**5) Les niveaux de curation**

L’organisation internationale [CoreTrustSeal](https://www.coretrustseal.org/) a défini 4 niveaux de curation :

* **Niveau A : Contenu publié tel que déposé**

Ce niveau, sans curation avant publication, n’est pas accepté dans l’entrepôt Recherche Data Gouv / EDI

* **Niveau B : Curation de base**

Présence de métadonnées et de documentation suffisantes pour comprendre le contexte de production des données.

* **Niveau C : Curation avancée**

Conversion vers des formats ouverts ou largement utilisés si besoin

Description précise de la production, de l’organisation et de la structuration des données :

* Dictionnaire de données, README, plan de gestion de données, etc., documentant les schémas et vocabulaires contrôlés utilisés pour les variables et valeurs ;
* Informations sur la provenance des données.
* **Niveau D : Curation au niveau des données**

Annotation et édition des données : requiert une expertise dans le domaine scientifique.

Un tel niveau de curation requiert une large équipe de curateurs possédant de l’expertise dans différents domaines.

* Le choix du niveau de curation relève de l’administrateur de la collection
* Tous les jeux de données déposés dans l’entrepôt Recherche Data Gouv/ EDI font l’objet d’une curation. Cette curation porte sur la qualité documentaire et garantit au minimum un niveau B.
* Les chercheurs/ auteurs sont seuls responsables de la qualité scientifique des données et du [respect du droit à diffuser ces données](https://hal.inrae.fr/hal-02791224).

**6) Curation des métadonnées**

Le curateur doit s’assurer de la complétude des métadonnées renseignées par les déposants. Pour y parvenir un ensemble de règles de saisie est établi. Ces règles de saisie des métadonnées vont servir à la rédaction du rapport de curation. Il existe deux types de règles de saisie :

* 1. **Règles de saisie des métadonnées de citation**

|  |  |  |
| --- | --- | --- |
| **Métadonnées** | **Explication** | **Ordre d’importance** |
| title | Titre complet le plus informatif possible.**Conseil :** privilégier la langue du public cible |  Obligatoire |
| subtitle | Titre secondaire utilisé pour compléter le titre principal. | Optionnel |
| Alternative title | Titre sous lequel le travail est communément appelé ou abréviation du titre. | OptionnelRecommandé : Si le titre principal est en français il est recommandé de le traduire ici en anglais  |
| Link to data | Adresse URL pour les données accessibles et dont le fichier de données n’est pas associé au jeu de données créé.  | OptionnelObligatoire cette information est obligatoire si les données ne sont pas déposées dans cet entrepôt |
| Other ID | Tout autre identifiant unique pour ce jeu de données | OptionnelExemple de champs : Agency, identifier |
| Contact | Contact(s) pour ce jeu de donnéesNB : Il faut cliquer sur le + pour ajouter un contact | ObligatoireExemple de champs : Name, Affiliation, E-mail |
| Author | Personne(s) physique(s) ou morale(s) responsable(s) du jeu de donnéesNB : Il faut cliquer sur le + pour ajouter un contact | ObligatoireName : ObligatoireAffiliation, identifier, identifiant ORCID: Obligatoire |
| Contributor | Personne ou organisme ayant contribué à la collecte, la gestion, etc. des données.NB : Il faut cliquer sur le + pour ajouter un contact | RecommandéExemple de champs : Type, Name, Contributor, Contributor identifier |
| Producer | La partie(organisme/personne) qui détient la responsabilité financière ou administrative de ce jeu de données.NB : Cliquer sur le + pour ajouter un producteur (dans le cas d’une UMR par exemple, indiquer chaque institution) | RecommandéExemple de champs: Name, Affiliation RecommandéAbréviation, URL, Logo URL :Optionnel |
| Production Date | Date à laquelle la collecte des données a été produiteNB : Différente des dates de distribution, publication ou archivage | Recommandé Format : AAAA-MM-JJ. |
| Production Place | Lieu de la collecte des données. | Recommandé |
| Distributor | Organisme désigné par l’auteur ou le producteur pour la diffusion des données.NB : cliquer sur le + pour ajouter un distributeur. | RecommandéName : Nom du diffuseur(Recommandé)Affiliation, Abbreviation, URL, Logo URL : Optionnel  |
| Distribution Date | Date à laquelle les travaux ont été diffusés, selon la convention ISO | OptionnelFormat AAAA-MM-JJ |
| Description | Résumé décrivant le jeu de données : intérêt, objectifs, contexte de production et son potentiel de réutilisation. | Text : ObligatoireLanguage : Recommandé  |
| Language | Langue(s) des données | Recommandé  |
| Subject | Domaine(s) dont relève le jeu de données.NB : sélectionner plusieurs domainesdans la liste déroulante**.** | Obligatoire |
| Keyword | Mots-clés (4 ou 5) décrivant les aspects importants du jeu de données.NB : Il est conseillé d’utiliser un vocabulaire de référence dans la discipline concernée (ex. PACTOL, Thésaurus INRAE, etc.). | Recommandé  / Obligatoire  |
| Topic classification  | Grands thèmes couverts par les données. | Optionnel |
| Kind of data  | Type(s) de données. | Obligatoire |
| Other Kind of Data | À remplir si "Other" est sélectionné pour Kind of Data. | Recommandé si applicable  |
| Data Origin  | Origine des données. Exemple : données expérimentales | Recommandé  |
| Data sources | Liste des livres, articles, séries ou fichiers de données qui ont servi de sources pour la collecte des données. | Optionnel |
| Origin of sources  | Pour les documents historiques : renseignements sur l'origine des sources, règles suivies dans l'établissement des sources.Exemple : Institut Géographique National | Optionnel |
| Characteristic of sources Noted | Évaluation et caractéristiques des sources de données. | Optionnel |
| Documentation and Access to Sources | Niveau de documentation des sources. | Optionnel |
| Software | Renseignements sur le logiciel utilisé pour générer le jeu de données. | OptionnelName, Version : Recommandé  |
| Series | Renseignements sur la série de jeux de données | Optionnel |
| Life cycle step | Événement(s) du cycle de vie des données considéré comme suffisamment important pour être documenté. | Optionnel |
| Notes | Autres renseignements importants sur le jeu de données. | Optionnel |
| Related publication  | Publication(s) utilisant les données de ce jeu de données. Seule la première sera visible sous la citation, la liste complète sera affichée dans l’onglet Métadonnées. | Obligatoire |
| Related material | Informations, telles qu’un ID pérenne ou une citation, sur le contenu lié au jeu de données : annexe, informations d’échantillonnage disponibles en dehors du jeu de données, etc. | Recommandé  |
| Related datasets | Informations, telles qu’un ID pérenne ou une citation, sur un autre jeu de données, produit lors de recherches précédentes sur le sujet du jeu de données | Recommandé  |
| Other References | Toute référence qui servira de contexte ou d'aide pour ce jeu de données. | Optionnel |
| Grant information | Renseignements sur les subventions obtenues | Recommandé  |
| Project information | Projet à l’origine de la production des données. | Recommandé  |
| Time Period Covered | Période couverte par les données. Différent des dates de codage, création des fichiers, ou de collecte des données. | Recommandé Start : date de début au format : AAAA-MM-JJEnd : date de fin au format : AAAA-MM-JJ |
| Date of Collection  | Période de collecte des données | Recommandé Start : date de début au format : AAAA-MM-JJEnd : date de fin au format : AAAA-MM-JJ |

* 1. **Règles de saisie des métadonnées de fichiers**

|  |  |  |
| --- | --- | --- |
| **Métadonnées** | **Explication** | **Ordre d’importance** |
| Nom du fichier | Nom du fichier respectant la convention de nommage si elle a été établie | Obligatoire |
| Chemin d’accès au fichier  | Chemin hiérarchique de la structure des répertoires utilisé pour afficher l’arborescence des fichiers. | Optionnel |
| Description | Résumé décrivant le fichier et ses variables. | Recommandé  |
| Fichier de provenance | Fichier issu d’un outil de capture de provenance (tels que provR, RDataTracker, NoWorkFlow, recordr ou CamFlow) et qui permet de générer un graphique de l’origine des données et des transformations subies. | Optionnel**NB :** Le fichier doit être au format JSON et être conforme aux normes du W3C. Cette information est pour le moment uniquement visible aux personnes qui ont des droits d'édition sur le jeu de données. |
| Description de la provenance | Informations documentant l'historique du fichier de données, y compris la manière dont il a été créé, comment il a été modifié et qui y a travaillé. | Optionnel**NB :** Cette information est pour le moment uniquement visible aux personnes qui ont des droits d'édition sur le jeu de données. |
| Libellés de fichier | Libellé décrivant un fichier. Les valeurs proposées dans la liste sont : Documentation, Données, Code. Il est possible de créer de nouveaux libellés | Recommandé **NB :** Les libellés ne doivent pas remplacer d’autres métadonnées spécifiques, telles que Keyword ou Kind of data. |
| Libellés des données tabulaires | Libellé qualifiant le type de données. Les valeurs proposées dans la liste sont : Survey, Time Series, Panel, Event, Genomics, Network, Geospatial. | Recommandé **NB :** Les libellés ne doivent pas remplacer d’autres métadonnées spécifiques, telles que Keyword ou Kind of data. |

Annexe 5 : Les schémas de métadonnées de RDG

Voici le nom des champs de métadonnées proposés par RDG :



Le champ « Citation » est obligatoire au niveau de RDG donc d’EDI.

Le champ « Life science » est affiché par défaut dans EDI, avec certains champs à saisie obligatoire.

En voici le contenu détaillé :



RDG propose [ici](https://recherche.data.gouv.fr/fr/categorie/39/guide/guide-de-saisie-des-metadonnees-de-citation) un guide de remplissage du champ « Citation ».

Les champs suivants sont dits « métadonnées de fichiers » ; leur remplissage est décrit [ici](https://recherche.data.gouv.fr/fr/categorie/39/guide/guide-de-saisie-des-metadonnees-de-fichiers).











|  |
| --- |
|  |

|  |
| --- |
|  |

1. Fichier « EDI\_Rapport-et-conclusions.docx » disponible sur Resana : «https://resana.numerique.gouv.fr/public/perimetre/consulter/146874?information=2058787 ». [↑](#footnote-ref-1)