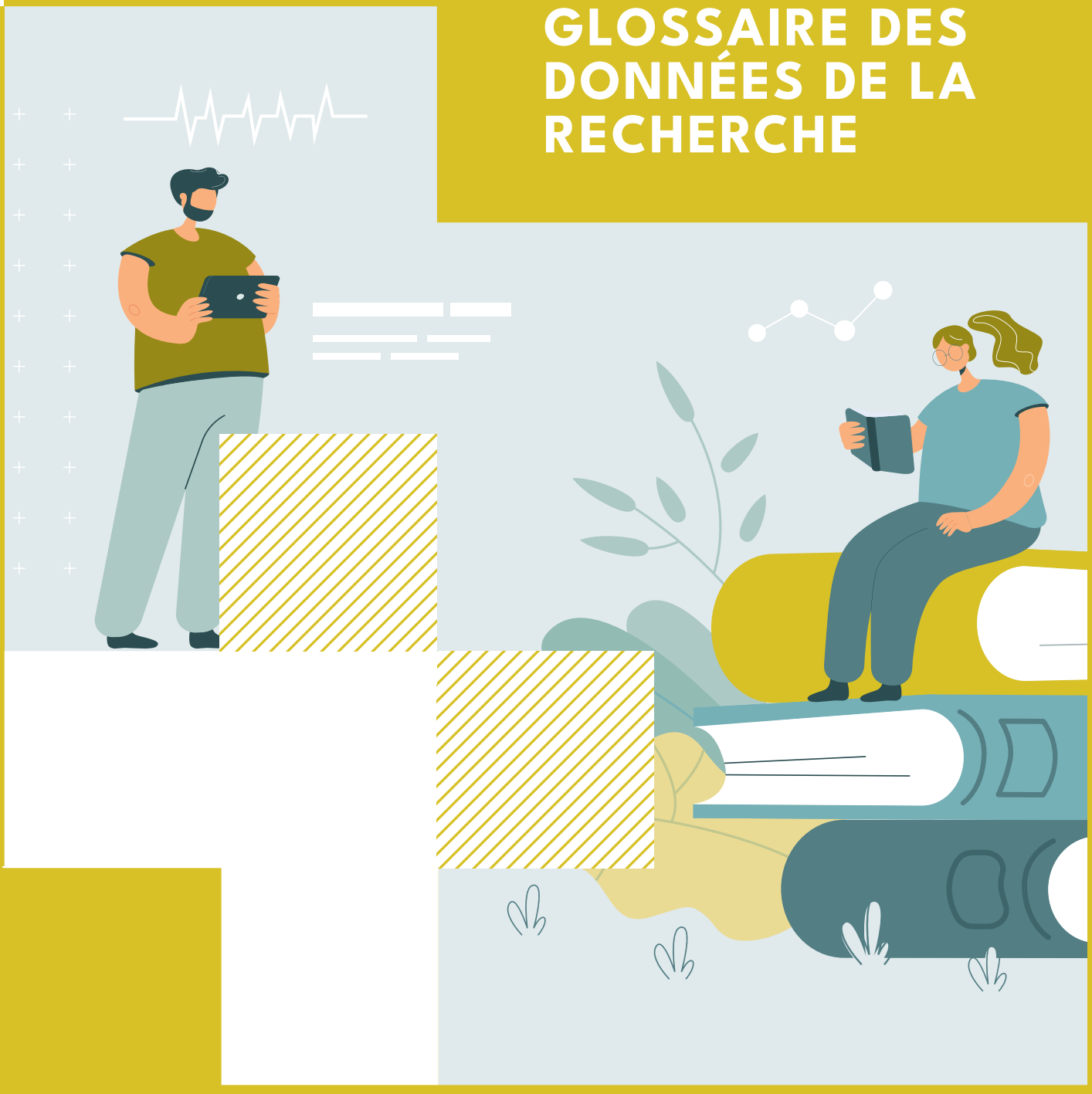
GROUPE INSA

CONSCIENCE COLLECTIVE

RECHERCHE INSA

GLOSSAIRE DES DONNÉES DE LA RECHERCHE



GROUPE INSA

GLOSSAIRE

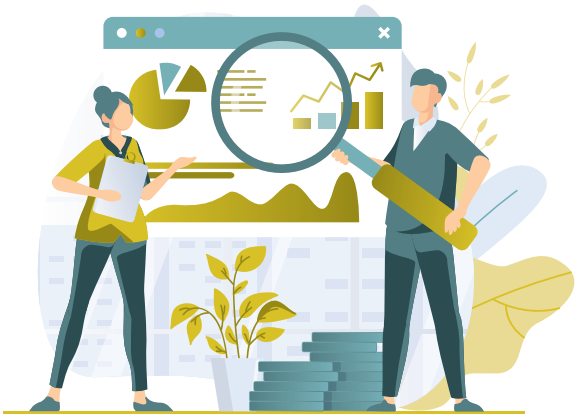
À PROPOS

En 2021, le Groupe INSA publiait son **Guide de la science ouverte**, avec l'ambition de sensibiliser et de former les personnels des écoles du Groupe INSA aux enjeux de ce mouvement. Au cours des dernières années, la science ouverte est en effet devenue un sujet crucial dans le monde de la recherche.

Maximiser la diffusion sans entrave des publications et des données de la recherche : l'enjeu est de taille, dans un contexte où l'adage - pour les données - « aussi ouvert que possible, aussi fermé que nécessaire » guide les travaux et réflexions de nos communautés. En effet, toutes les données de la recherche n'ont pas vocation à être ouvertes, pour des raisons de protection des données personnelles, de concurrence industrielle ou encore pour des intérêts fondamentaux ou réglementaires des États.

La question des données dans la recherche est complexe et nécessite de disposer d'une information éclairée sur le sujet. C'est pourquoi, les acteurs science ouverte du Groupe INSA ont souhaité poursuivre leur travail de sensibilisation entamé avec **le Guide de la science ouverte pour mettre à la disposition de toute la communauté recherche INSA un document ressource unique, pour enrichir les connaissances de toutes et tous sur le sujet et poursuivre ainsi la valorisation de la recherche du collectif INSA.**

Une recherche responsable et engagée, en phase avec les grands défis environnementaux, écologiques et sociétaux.



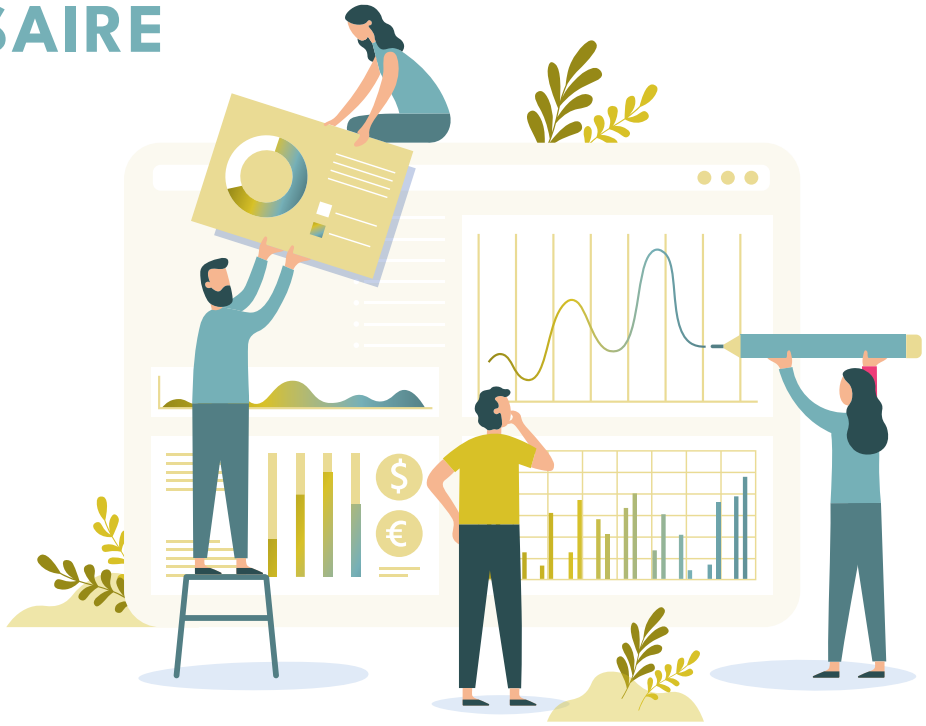
SOURCES UTILISÉES POUR LA RÉDACTION DU GLOSSAIRE :

- **RechercheDataGouv :**
  https://recherche.data.gouv.fr/fr/glossaire
- **DoraNum :**
  https://doranum.fr/glossaire-donnees-recherche
- **Ouvrir la science ! :**
  https://www.ouvrirlascience.fr/category/science-ouverte/glossaire
- **Plan "Données de la recherche" (CNRS) :**
  https://lstu.fr/plan-donnees-cnrs
- **Glossaire DMP OPIDoR :**
  https://dmp.opidor.fr/static/glossary

LES RÉFÉRENTS SCIENCE OUVERTE DU GROUPE INSA

- **INSA Centre Val de Loire :**
  hal@insa-cvl.fr
- **INSA Hauts-de-France :**
  servicechercheurs-bu@insa-hdf.fr
- **INSA Lyon :**
  hal@insa-lyon .fr
- **INSA Rennes :**
  bibliotheque@insa-rennes.fr
- **INSA Rouen Normandie :**
  biblio-hal@insa-rouen.fr
- **INSA Strasbourg :**
  bibliotheque@insa-strasbourg.fr
- **INSA Toulouse :**
  hal-insa@insa-toulouse.fr

GLOSSAIRE



ARCHIVER

L'archivage consiste à ranger un document dans un lieu où il sera conservé pendant une période plus ou moins longue et d'y associer les moyens pour réutiliser les données : la réutilisation se faisant en ajoutant de l'intelligence à la sauvegarde. Le contenu des documents archivés n'est pas modifiable. En revanche, le contenant (format) des documents archivés peut être modifié, pour éviter l'obsolescence logicielle.

ANONYMISATION DES DONNÉES

L'anonymisation est un 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. Contrairement à la pseudonymisation, l'anonymisation est une opération irréversible.

CODE SOURCE

Ensemble d'instructions composant un programme informatique dans un langage de programmation. Le code source se matérialise généralement sous la forme d'un ensemble de fichiers textes lisibles par son utilisateur et exécutables par une machine. Le code source est la représentation d'un logiciel pour qu'un utilisateur puisse y apporter des modifications.

CONVENTION DE NOMMAGE

Règles formalisées adoptées pour nommer et classer les fichiers ou documents d'un projet.

CURATION

Vérification des métadonnées et des fichiers de données déposés dans l'entrepôt dans le but de proposer des modifications pour améliorer la qualité des jeux de données, dans le respect de la charte des curateurs*.

*https://lstu.fr/charte-curateurs

DATA JOURNAL

Journal (toujours en libre accès) qui publie des articles de données, appelés data papers. Il fournit habituellement des modèles de description des données et guide les chercheurs sur les lieux de dépôt et sur la façon de décrire et de présenter leurs données.

DATA PAPER

Publication qui décrit un jeu de données scientifique, notamment à l'aide d'informations structurées appelées métadonnées. Le data paper présente le contexte de production des données mais, contrairement à un article de recherche classique, ne mentionne pas les hypothèses testées ou les analyses réalisées.

DATAVERSE*

Logiciel libre utilisé pour le développement de la plateforme Recherche Data Gouv qui permet de partager, préserver, citer, explorer et analyser des données de recherche. Il facilite la mise à disposition des données aux autres.

* https://dataverse.org/

DÉPOSANT

Toute personne inscrite dans l'entrepôt et autorisée à y accéder afin de mettre à disposition un jeu de données dont la publication présente un intérêt public.

DONNÉES DE LA RECHERCHE

Enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont reconnus par la communauté scientifique comme nécessaires pour la validation des résultats.

Données

Faits ou chiffres collectés pour un projet de recherche (statistiques, mesures, observations, résultats d'enquêtes, textes, images, sons), disponibles sous forme numérique, utilisés comme sources principales pour la recherche scientifique. Les données sont généralement reconnues par la communauté scientifique comme nécessaires pour valider les résultats de la recherche.

Données achevées

Données non préparatoires produites par les établissements de recherche et d'enseignement dans le cadre de leur mission de service public. Il peut s'agir de données brutes, de données élaborées ou de métadonnées. Elles sont qualifiées de documents administratifs et sont donc communicables à toute personne qui en fait la demande, sauf exceptions légales.

Données brutes

Les données brutes sont des données issues d'une expérimentation, d'un procédé, d'une enquête, etc. : il peut s'agir de données de recherche communicables.

Données environnementales

Les données environnementales regroupent toutes les informations relatives à :

- L'état des éléments de l'environnement, notamment l'air, l'atmosphère, l'eau, le sol, les terres, les paysages, les sites naturels, les zones côtières ou marines et la diversité biologique, ainsi que les interactions entre ces éléments ;
- Les décisions, les activités et les facteurs, notamment les substances, l'énergie, le bruit, les rayonnements, les déchets, les émissions, les déversements et autres rejets, susceptibles d'avoir des incidences sur l'état des éléments visés au premier point ;
- L'état de la santé humaine, la sécurité et les conditions de vie des personnes, les constructions et le patrimoine culturel, dans la mesure où ils sont ou peuvent être altérés par des éléments de l'environnement, des décisions, des activités ou des facteurs susmentionnés ;
- Les analyses des coûts et avantages ainsi que les hypothèses économiques utilisées dans le cadre des décisions et activités visées au deuxième point ;
- Les rapports établis par les autorités publiques ou pour leur compte sur l'application des dispositions législatives et réglementaires relatives à l'environnement ;

Les dispositions concernant ces données sont issues de la Convention d'Aarhus. Les données environnementales devant être diffusées sont celles relatives à des zones sur lesquelles la France détient ou exerce sa compétence.

Données géographiques

Données faisant directement ou indirectement référence à un lieu ou une zone géographique spécifique. Les thèmes des données géographiques devant être diffusées sont listés aux Annexes I, II et III de la directive INSPIRE*. Ce sont les données géographiques relatives à une zone sur laquelle la France détient ou exerce sa compétence qui doivent être diffusées.

* https://lstu.fr/directive

Données intégrées

Données figurant dans une publication scientifique.

Données personnelles

Constitue des données à caractère personnel toutes les informations relatives à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres.

Données présentant des risques pour la protection du potentiel scientifique ou technique de la nation.

Savoirs, savoir-faire et technologies dont le détournement ou la captation pourrait porter atteinte aux intérêts économiques de la nation, renforcer des arsenaux militaires étrangers ou affaiblir les capacités de défense de la France, contribuer à la prolifération des armes de destruction massive et de leurs vecteurs ou favoriser les actions malveillantes sur le territoire national ou à l'étranger.

Données préliminaires

Les données préliminaires sont des données préparatoires, préalables, nécessaires à la mise en place d'une expérimentation, d'un procédé, d'une enquête, etc. Il ne s'agit pas de données de recherche.

Données publiques

Données incluses dans les documents produits ou reçus dans le cadre de leur mission de service public par l'État, les collectivités territoriales, les autres personnes de droit public ainsi que les personnes de droit privé chargées d'une mission de service public. Des exemples de données publiques sont donnés aux articles L.300-2 et L.312-1-1 du Code des relations entre le public et l'administration.

Données sensibles

Informations réglementées par la loi en raison d'un risque possible pour la faune, la flore, les êtres humains ou les communautés, ainsi que pour les organisations publiques et privées. Les données à caractère personnel sensibles comprennent les informations relatives à l'origine raciale ou ethnique, aux opinions politiques, aux croyances religieuses ou philosophiques, à l'appartenance syndicale et à la santé ou la vie sexuelle d'un individu. Ces données pourraient être identifiables et leur divulgation pourrait par conséquent causer préjudice. Pour les autorités locales et gouvernementales, les données sensibles sont liées à la sécurité (données politiques, diplomatiques ou militaires, risques biologiques, etc.), aux risques environnementaux (installations nucléaires et autres installations sensibles) ou à la protection de l'environnement (habitat naturel, faune ou flore protégée). Les données sensibles des organismes privés concernent en particulier des éléments stratégiques ou susceptibles de compromettre leur compétitivité.

Données sous-jacentes

Données nécessaires à la validation des résultats présentés dans les publications scientifiques.

Données statistiques

Données collectées par voie d'enquête statistique ou transmises au service statistique public à des fins d'établissement des statistiques.

GLOSSAIRE (suite)

ENTREPÔT DE DONNÉES

Un entrepôt de données de recherche (Research Data Repository ou Data Repository) est une base de données destinée à accueillir, conserver, rendre visibles et accessibles des données de recherche. Son rôle est de permettre le dépôt ou la collecte de données, leur description, leur accès, et leur partage en vue de leur réutilisation. Chaque entrepôt dispose généralement d'une politique de dépôt, de description et de diffusion des données.

ENTREPÔT DE DONNÉES DE CONFIANCE

Entrepôt de données qui respecte les critères de confiance définis dans le guide Criteria for the Selection of Trustworthy Repositories*. Ces critères visent à promouvoir des entrepôts de données fiables et durables. De nombreux entrepôts n'ont pas de certification mais sont cependant largement reconnus par la communauté scientifique et offrent des garanties de conservation à long terme.

* https://lstu.fr/trustworthy-repositories

ENTREPÔT MIXTE

Entrepôt contenant à la fois des publications scientifiques et des jeux de données.

ESPACE INSTITUTIONNEL

Espace créé dans la plateforme Recherche Data Gouv à la demande d'un établissement de recherche française et dont l'administration et la curation lui sont déléguées.

FAIR (PRINCIPES)

La notion de FAIR (Findable, Accessible, Interoperable, Reusable) data recouvre les manières de construire, stocker, présenter ou publier des données de manière à faire en sorte que la donnée soit facile à trouver, accessible, interopérable et réutilisable.

FICHIER README

Les fichiers README sont des guides, habituellement en texte clair, qui maximisent la stabilité et la préservation à long terme des données. Ils ont pour but d'aider les chercheurs à comprendre ces ensembles de données, ainsi que leur contenu, provenance, licence et manière de les utiliser. Ces fichiers sont habituellement nommés README, readme.txt ou readme.md.

INTEROPÉRABLE

Principe FAIR* qui peut se décomposer en : téléchargeable, utilisable, intelligible, et combinable avec d'autres données, par des humains et des machines. Une mise en œuvre courante de ce principe consiste à utiliser les technologies du Web sémantique (RDF, OWL, SKOS) pour représenter et lier les données et les métadonnées.

* https://lstu.fr/principes-fair

JEU DE DONNÉES

Agrégation, sous une forme lisible, de données brutes ou dérivées présentant une certaine "unité", rassemblées pour former un "ensemble cohérent". Un jeu de données est un ensemble de ressources qui forme une unité cohérente du point de vue du contenu. Il est important de bien réfléchir à la granularité du jeu de données. Attention, dans le cas des logiciels, un jeu de données peut être le code source ainsi que la documentation associée.

MÉTADONNÉE

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

OPIDOR

OPIDoR* (Optimisation du partage et de l'interopérabilité des données de la recherche) est un portail mis en place et hébergé par l'Inist-CNRS. Il met à disposition de la communauté de l'enseignement supérieur et de la recherche un ensemble d'outils et de services facilitant la gestion et la valorisation des données afin de répondre aux critères d'intégrité, de reproductibilité et aux principes FAIR.

* https://dmp.opidor.fr

ORCID

ORCID ID (Open Researcher and Contributor Identifier) est un identifiant numérique pérenne qui permet d'identifier de manière univoque un chercheur donné et qui référence et regroupe les travaux scientifiques de ce chercheur issus de différentes plateformes de dépôt telles que HAL ou site de publication. La mention d'un ORCID est fortement recommandée voire indispensable pour soumettre une publication chez la plupart des éditeurs scientifiques (Elsevier, Wiley, Springer...). Il est de plus en plus utilisé par les organismes de financement.

PLAN DE GESTION DES DONNÉES (PGD ou DMP pour Data Management Plan)

Plan évolutif, rédigé en début d'un projet de recherche, qui précise les modalités de la gestion des données (collecte, documentation, stockage, gestion des données sensibles, conditions d'ouverture ou de partage, etc.).

RE3DATA

Le Re3data* (Registry of Research Data Repositories) est un annuaire mondial d'entrepôts de données de recherche couvrant différentes disciplines scientifiques. Il présente des entrepôts pour le stockage permanent et l'accès aux jeux de données destinés aux chercheurs, aux organismes de financement, aux éditeurs et aux établissements d'enseignement.

* https://www.re3data.org

REPRODUCTIBILITÉ

Capacité pour un autre chercheur d'obtenir les mêmes résultats en utilisant les mêmes méthodes et données (met en lumière l'importance des méthodes de production des résultats).

SAUVEGARDER

La sauvegarde consiste à dupliquer les données sur un support numérique externe à celui où elles sont stockées. L'objectif est de pouvoir les retrouver en cas de perte ou de dégradation de l'organe de stockage. Il s'agit d'une sauvegarde octet par octet dans une perspective de court ou de moyen terme. La recherche de la préservation de l'intelligibilité des données n'est pas un élément pris en compte.

STANDARD DE MÉTADONNÉES

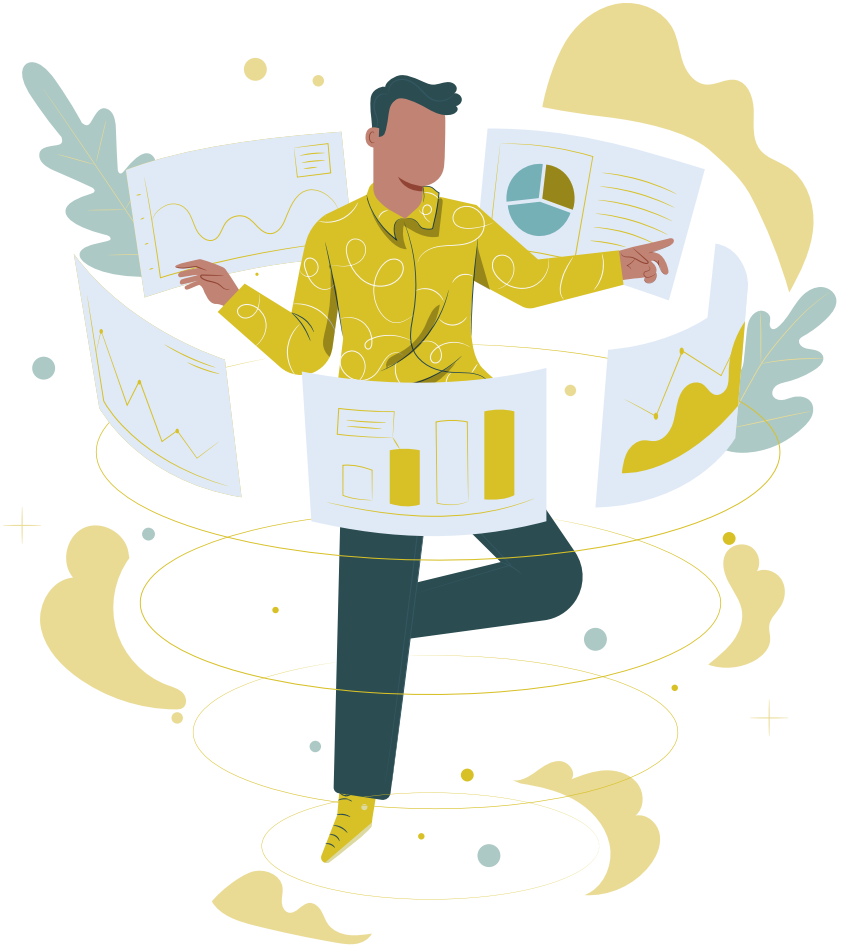
Un standard de métadonnées (ou schéma de métadonnées) est un modèle qui précise toutes les métadonnées nécessaires pour décrire un certain type de données. Utiliser un standard de métadonnées permet de décrire les données de façon riche et précise, en utilisant le même vocabulaire que votre communauté (interopérabilité sémantique).

Exemples : Dublin Core, DDI (Data Documentation Initiative), EML (Ecological Metadata Language), Metadata Standards Catalog *.

* https://lstu.fr/metadata-standards-catalog

STOCKER

Étape qui consiste à déposer les données sur un support numérique pour les rendre accessibles. Cela peut être un ordinateur personnel, un disque partagé ou tout autre organe de dépôt. Le stockage permet d'assurer la continuité de l'exploitation sur du court terme. À ce stade, la donnée n'est ni sauvegardée ni sécurisée.



GROUPE INSA

INSA

CENTRE VAL DE LOIRE
HAUTS-DE-FRANCE
LYON
RENNES
ROUEN NORMANDIE
STRASBOURG
TOULOUSE

En savoir plus :
www.groupe-insa.fr

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Cti Commission des Titres d'Ingénieur

DIPLÔME D'INGÉNIEUR CONTRÔLÉ PAR L'ÉTAT

European Accreditation of Engineering Programmes EUR-ACE®

Création : www.lamatiererose.com | Crédits illustrations : freepik.com | Novembre 2023