

Choisir un entrepôt de données généraliste

Candice HECTOR
Alicia LEÓN Y BARELLA
Dimitri SZABO



Le Printemps
de la Donnée

Programme

Durée 2 h

Première partie

- 1 — Les entrepôts de données
- 2 — Présentation de Zenodo
- 3 — Présentation de Recherche Data Gouv

Deuxième partie

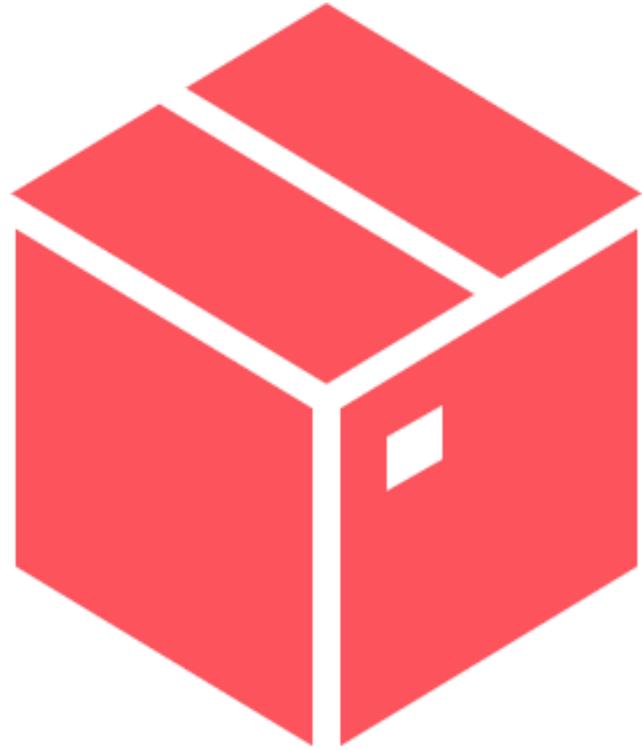
- 4 — Préparer et documenter ses jeux de données
- 5 — Déposer son jeu de données

Objectifs de l'atelier

À l'issue de l'atelier, vous devriez pouvoir :

- Identifier des entrepôts généralistes
- Trouver des entrepôts
- Préparer et documenter vos jeux de données
- Déposer vos données dans un entrepôt généraliste

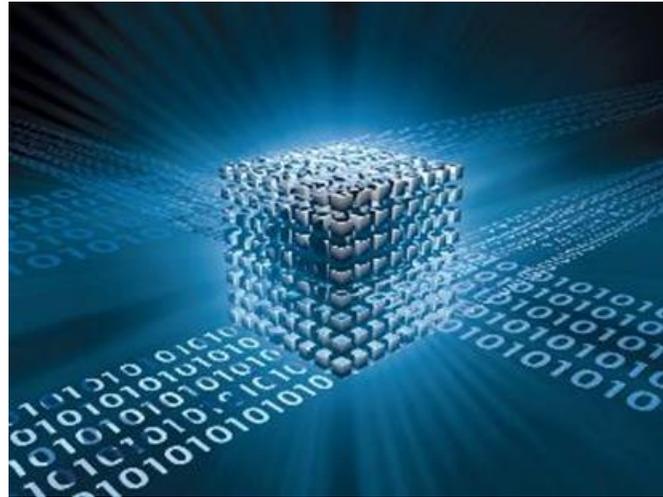




1 Les entrepôts de données

Les entrepôts de données

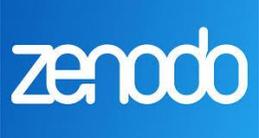
Un entrepôt de données est une base de données dans laquelle un chercheur peut déposer des jeux de données afin d'en permettre le stockage, l'accès, la diffusion et la réutilisation auprès d'une communauté et/ou du grand public



Il existe de nombreux entrepôts répartis en plusieurs types : disciplinaires, généralistes (multidisciplinaires), propres à un éditeur, institutionnels, spécifiques à un projet de recherche

Où diffuser vos données? Quelques exemples

Entrepôts généralistes



La plateforme nationale
fédérée des données
de la recherche
LANCEMENT
Printemps 2022

[Recherche Data Gouv](#)

Entrepôts disciplinaires



Santé

Systeme Terre
et
Environnement



DATA
TERRA



SHS

Code source
des logiciels



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Comment choisir son entrepôt ?

Un entrepôt est-il préconisé par :

- Le financeur ?
- La discipline ?
- L'institution de recherche ou les partenaires ?

Trouver un entrepôt de données grâce à des annuaires spécialisés :

- Re3data : <https://www.re3data.org/>
moteur de recherche des entrepôts de données qui affine les résultats en fonction des besoins
- CatOpidor : <https://cat.opidor.fr/index.php/Conservation>
permet d'identifier les services nationaux dédiés aux données de recherche

Comment choisir son entrepôt ?

Critères pour choisir un entrepôt :

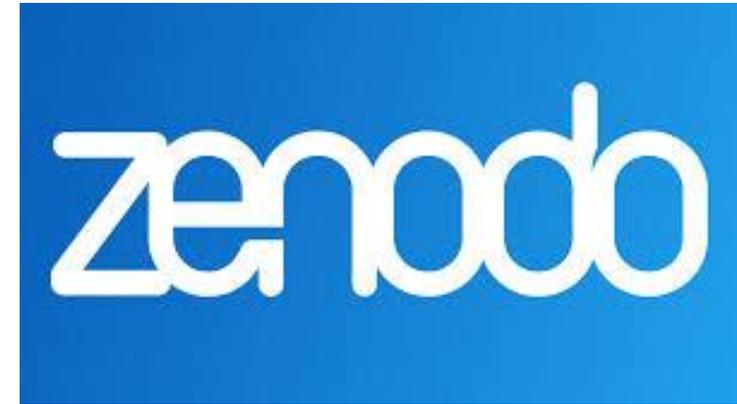
- Par discipline
- Gestion des droits d'accès (accès restreints, embargos, licences d'utilisation...)
- Identifiant pérenne (DOI, PURL, etc.)
- Gestion des versions déposées
- Par certification (CoreTrustSeal : certifie la qualité et la fiabilité d'un entrepôt)
- Métadonnées adaptées (schéma, standard...)
- Formats de fichiers autorisés
- Propriété des données conservée
- Liaison entre données et publications scientifiques
- Limite de volumétrie
- Lieu physique de stockage (politique de l'entrepôt et du pays en vigueur)
- En fonction des coûts et des statistiques d'utilisations
- Durée de conservation proposée...



2 Présentation de Zenodo

Présentation de Zenodo

- Zenodo est un entrepôt de données multidisciplinaire
- Financé par la Commission européenne
- Piloté par le CERN (European Organization for Nuclear Research)
- En collaboration avec OpenAIRE (Open Access Infrastructure for Research in Europe)



Avantages et inconvénients



Avantages

- Tous domaines
- Tout type de données
- Mixte : accepte les données et les publications scientifiques
- 50 gigas par jeux de données
- Attribution automatique de DOI
- Gestion des accès (accès ouvert, accès restreint, embargo, accès fermé)
- Visibilité des dépôts immédiate
- Communauté U-Lille et possibilité d'en rejoindre d'autres
- Notification automatique auprès du financeur pour les projets financés OpenAIRE et les organismes associés
- Gratuit

Inconvénients

Si les articles ne sont pas liés aux données, privilégier le dépôt de ces derniers dans des archives ouvertes (Lilloa ou HAL)

Pas de modération de la part de Zenodo : pas de vérification sur le contenu des dépôts, la fiabilité de ces derniers et sur les droits associés

Démonstration

The image shows a screenshot of the Zenodo website interface. At the top, there is a navigation bar with icons for 'Inscription Connexion', 'Téléchargement', 'Formulaire', and 'Dépôt'. Below this is the Zenodo logo and a search bar. A blue arrow points to the 'Upload' button in the top navigation bar, labeled '1. Cliquez sur « Upload »'. Below the search bar, there is a 'New Upload' button, which is highlighted by a blue arrow labeled '2. Cliquez sur « New Upload »'. In the center of the page, there is a large area with the text 'Drag and drop files here' and a 'Choose files' button, which is highlighted by a blue arrow labeled '3. Faites glisser vos fichiers ou cliquez sur « Choose files »'. A file explorer window is open in the bottom left corner, showing a directory structure with various files and folders.

12

DoRANum. Données de la recherche : apprentissage numérique [En ligne]. France : Doranum; 2021. Dépôt et Entrepôts : Déposer ses données de recherche dans Zenodo. 25 mars 2021; Disponible sur : <https://doranum.fr>. <https://doranum.fr/depot-entrepots/depot-donnees-recherche-zenodo/>

02/05/2022

Démonstration

Inscription Connexion

Téléchargement

Formulaire
Champs à renseigner

Dépôt

1. Type de publication : 5. Accès
2. Communauté : 6. Financement
3. DOI : 7. Travaux associés
4. Champs de base : 8. Champs optionnels

Upload type required

Publication Poster Presentation Dataset Image Video/Audio Software Lesson Physical object Other

Publication type: Journal article

Journal article
Annotation collection
Book
Book section
Conference paper
Data management plan
Journal article
Patent
Preprint
Project deliverable
Project milestone
Proposal
Report
Software documentation
Taxonomic treatment
Technical note
Thesis
Working paper
Other

DoRANum. Données de la recherche : apprentissage numérique [En ligne]. France : Doranum; 2021. Dépôt et Entrepôts : Déposer ses données de recherche dans Zenodo. 25 mars 2021; Disponible sur : <https://doranum.fr>.
<https://doranum.fr/depot-entrepots/depot-donnees-recherche-zenodo/>

Démonstration

Inscription
Connexion

Téléchargement

Formulaire

Dépôt

DÉPÔT

Une fois le formulaire rempli, vous pouvez l'enregistrer ou déposer votre jeu de données.

Save Publish

*Votre jeu de données est publié.
Vous pouvez alors le visualiser.*

Important : une fois soumis, il est possible de modifier le formulaire mais impossible de modifier ou de supprimer le dépôt.

zenodo Research. Shared.

Search Communities Browse Upload Get started

Home / Search results

Source code and datasets used to link new waves of plague outbreaks in medieval Europe to climate fluctuations affecting

Showing results 1 to 1 out of 1 results

Any Collection Datasets (1)

Any Author Branwen, Barbara (1) Brangen, M (1) Caenning, W Ryan (1) Goulet, Christian (1) Schmid, Sara M (1) More...

Any Access right Open (1)

11 February 2021 **New** **Open access**

Source code and datasets used to link new waves of plague outbreaks in medieval Europe to climate fluctuations affecting the reservoirs of the disease in Asia.

Schmid, Sara M., Brangen, M., Caenning, W., Ryan, B., Goulet, Christian, et al.

The profile contains the project directory which includes the source code and datasets used in the paper on Climate driven introduction of the Black Death and successive plague reintroductions into Europe, as published in Proceedings of the National ...



3

Présentation de Recherche Data Gouv

Pourquoi Recherche Data Gouv ?

Il existe des entrepôts génériques, mais il reste des problèmes...

- de curation et d'administration
- de souveraineté
- d'accompagnement



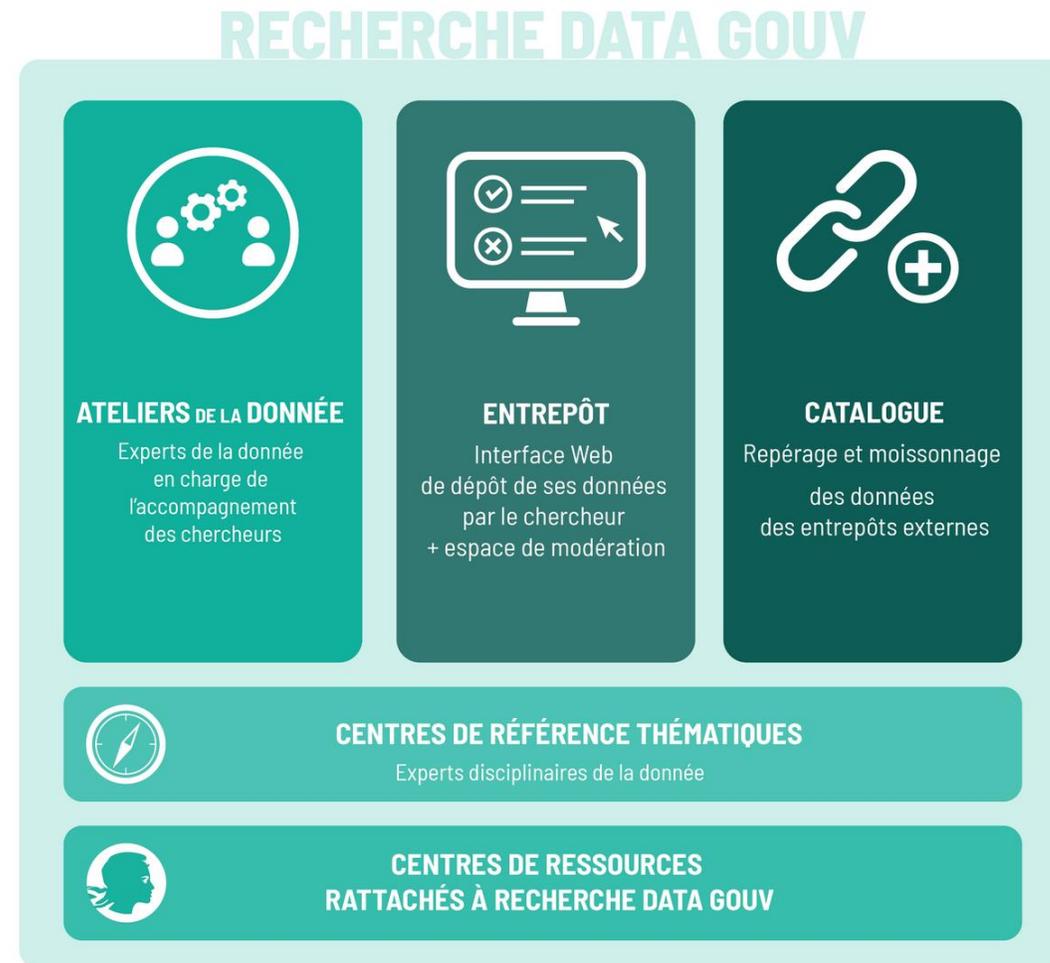
**RÉPUBLIQUE
FRANÇAISE**

*Liberté
Égalité
Fraternité*

Le projet

- 2 Phases : Béta et Cible pour pouvoir :
 - ouvrir rapidement un service
 - coconstruire la cible
- Première version s'appuyant sur un entrepôt existant, Data INRAE...
- ...Mais avec un projet impliquant différents instituts
- En savoir plus : <https://projet-recherchedatagv.ouvrirlascience.fr/>

Les différents modules



Utilisation de l'entrepôt

- Espace institutionnel « Université de Lille »
- Dépôt en deux étapes :
 - enregistrement du jeu de données
 - complétion des métadonnées, organisées en blocs
- Modalités spécifiques de l'espace à voir avec le SCD
- Guides existants :
 - <https://ist.blogs.inrae.fr/datainrae-guide/>
 - <https://guides.dataverse.org/en/5.3/>



4 Préparer et documenter ses jeux de données : quelques bonnes pratiques

Documenter ses données

La documentation des données est essentielle avant tout dépôt dans un entrepôt.

Cette étape rend les données compréhensibles et réutilisables par vous-mêmes et par d'autres.

Le déposant pourra par exemple produire des fichiers de description

- Fichiers Readme
- Dictionnaires de données,
- Dictionnaire de codes...



Nommer et organiser les fichiers

Nommage

Le nom des dossiers doit permettre d'identifier le projet, les chercheurs, la date de constitution des données, le type de données et/ou les conditions de collecte

Le nom des fichiers doit explicitement présenter leur contenu – sans répéter le nom d'un dossier :

- Abréviation normalisée
- Format de dates normalisé YYYY-MM-JJ
- Version
- Majuscules, tirets ou underscores
- Pas de caractères spéciaux, pas d'accents, pas d'espaces
- Traduire les chiffres : 1 à 10 = 01-10...

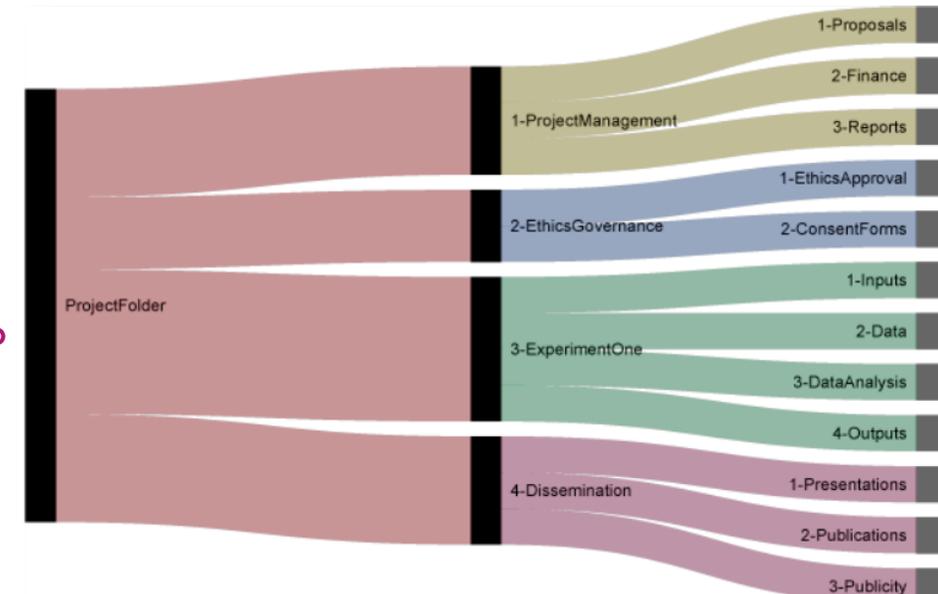
Organisation

Limiter le nombre de sous-dossiers pour limiter le nombre de clics

Ranger par thématiques : Gouvernance, expérimentation...

Séparer les données brutes des données traitées

Figure 1
proposée par
Nikola Vukovic,
UC San Francisco

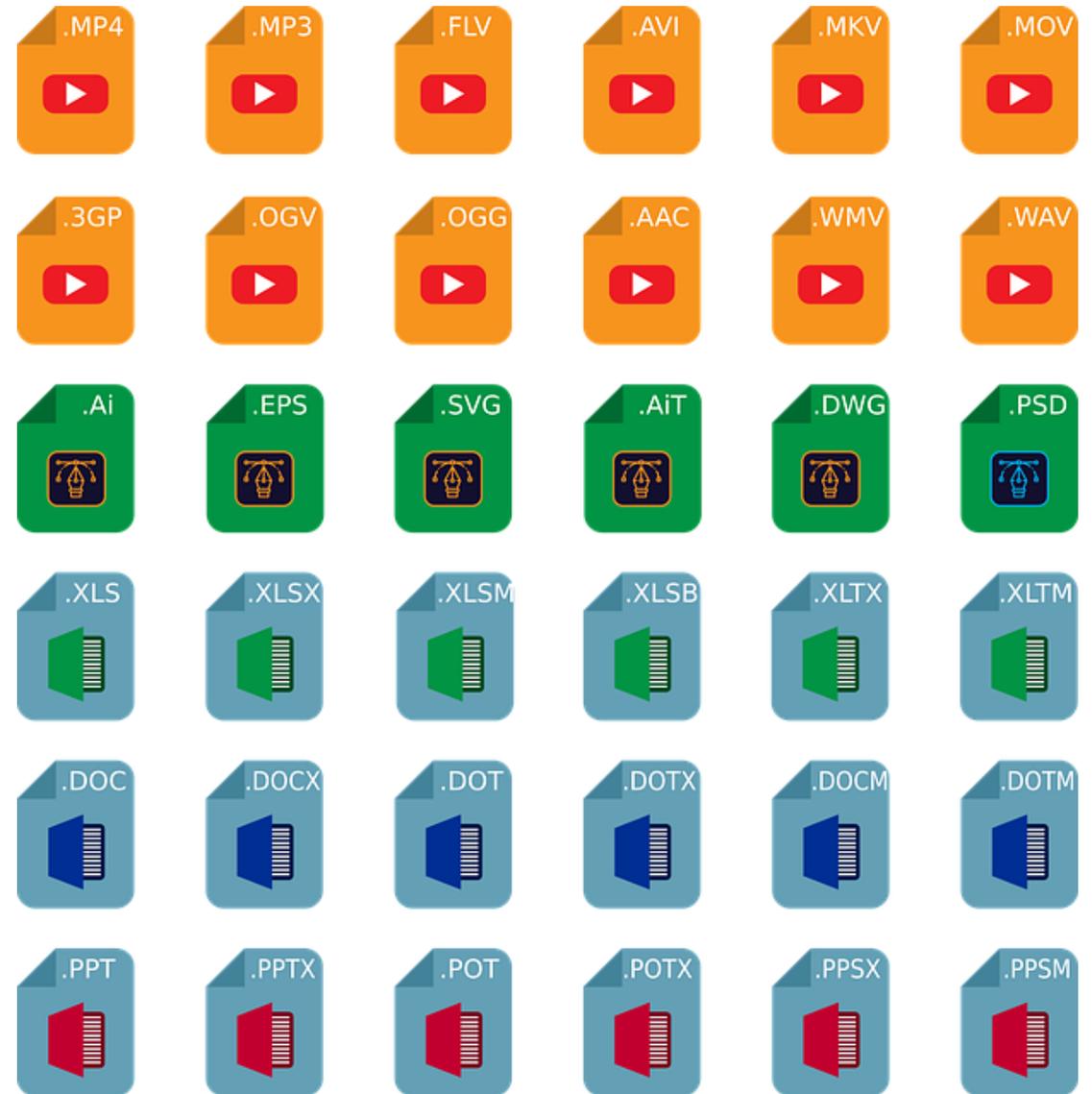


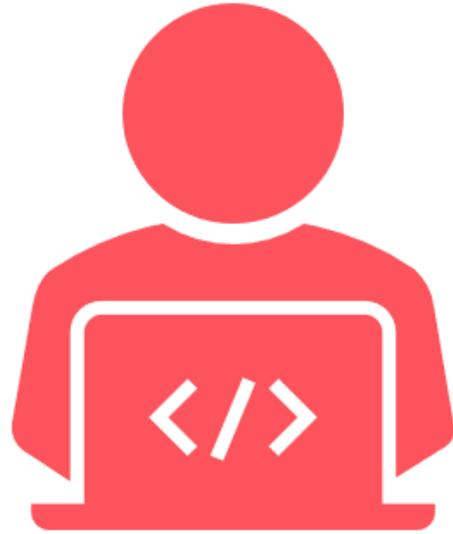
Formats de fichiers

Quand cela est possible, les formats préconisés sont les formats ouverts, non-propriétaires.

Ils garantissent l'interopérabilité, l'accessibilité et la modification du contenu indépendamment du logiciel utilisé.

Par exemple, pour l'utilisation de tableurs, utilisez le format ouvert CSV plutôt que le format propriétaire XLS d'Excel.





5

Déposer son jeu de données

Place à la pratique !

Pour choisir un jeu de données
Rendez-vous sur : <https://urlz.fr/i6Ei>



Pour l'ensemble des problématiques liées aux données de la recherche :



Règlementation

- [Logigramme de l'institut Pasteur](#) – Questions juridiques liées à la diffusion des données

Anonymisation des données

- <https://amnesia.openaire.eu/>

Description des données & fichier Read me :

- [Guide de Cornell](#)

Nommage :

- [Bulk Rename Utility](#) : un outil pour renommer automatiquement les fichiers

Formats de fichier

- [Liste indicative de formats ouverts et fermés](#)
- [L'outil FACILE du CINES](#), pour vérifier si les formats de vos fichiers sont pérennes

Merci de votre attention!



Candice HECTOR
Alicia LEÓN Y BARELLA

scd-aap@univ-lille.fr

Dimitri SZABO

dimitri.szabo@inrae.fr