Fiche proposée par le GTSO Données Couperin - mai 2025

Comment gérer des données de recherche volumineuses ?

Cette fiche s'adresse aux personnels d'appui amenés à accompagner les chercheurs sur les questions de données volumineuses.

1. Définition

Il est difficile de définir cette notion tant elle dépend de la discipline et des types de données produits. Nous pouvons néanmoins convenir qu'il s'agit d'un ensemble de données qui peut atteindre ou dépasser la capacité des moyens traditionnels de gestion et d'analyse des données (espace de stockage, bande passante, etc.). Nous pouvons nous baser sur les critères suivants :

- la quantité de données allant de plusieurs dizaines de giga-octets aux téraoctets et au-delà ;
- le temps important de téléchargement.

Les données peuvent parfois provenir de sources multiples et se présenter sous différents formats (textes, images, vidéos, données structurées et non structurées).

2. Comment partager un jeu de données volumineux ?

La mise à disposition d'un **jeu de données volumineux** peut s'avérer difficile du fait de sa spécificité et des contraintes techniques imposées par des solutions de partage. Cette liste de conseils non exhaustive a pour but de faciliter le processus d'ouverture d'un jeu de données de taille très importante (~ 500 Go à plusieurs To et plus) :

- définir de quel(s) **type(s) de données** il s'agit et respecter les contraintes liées à la typologie (voir la question *quels types de données est-ce que je produis ?* dans la FAQ)
- **trier des données**, définir des **critères de sélection** pertinents au contexte comme par exemple le coût de leur obtention, la facilité de leur reproduction, en particulier pour les types de données les plus volumineux (vidéos, images, …)
- vérifier si un entrepôt thématique de confiance existe et s'il accepte des jeux de données volumineux
- envisager l'option Recherche Data Gouv si aucun entrepôt thématique ne correspond à votre domaine de recherche (voir partie 3 : *Spécificité du dépôt dans Recherche Data Gouv*).
- s'adresser à l'atelier de la donnée de proximité pour étudier la possibilité de création du lien avec l'un des mésocentres
- vérifier **l'offre de service** de sa Direction des Systèmes d'Information pour identifier des outils efficaces et sécurisés pour le transfert de données volumineuses entre machines
  - p.ex. : service Renater (max 100 Go par dépôt)
  - s'adresser aux services de la DSI en cas de besoin de solution personnalisée
- vérifier et budgétiser le coût des espaces de partage, certains entrepôts peuvent facturer la prestation de mise à disposition d'un jeu de données en fonction de sa taille comme p.ex. : https://publicneuro.eu/upload.html

3. Spécificité du dépôt dans Recherche Data Gouv

La taille

- 50 Go par fichier. Le nombre de fichiers qu'il est possible de téléverser via l'interface utilisateur est limité à 1000 fichiers par téléversement. Au-delà, il faudra utiliser une API Dataverse ou l'outil DVUploader.
- la taille de l'espace institutionnel allouée par défaut : 5To max mais négociable et modifiable si besoin auprès des administrateurs de Recherche Data Gouv.
- pas de modèle économique connu à ce jour.
- exemple d'un jeu de données de ~ 1To disponible sur Recherche Data Gouv : https://doi.org/10.57745/XWDCT4.

Condition émise par Recherche Data Gouv : si un jeu de données est très volumineux, il faut contacter les administrateurs de votre espace institutionnel pour un accompagnement personnalisé.

4. Comment je stocke et je partage un jeu de données volumineux ?

Étapes	Questions à se poser	Recommandations ~ Ressources utiles
PLANIFIER	- Quelle est la volumétrie envisagée des données qui seront partagées à la fin du projet (fourchette ou volume initial avec taux de croissance estimé) ? - De quels outils vais je avoir besoin ? - Quels sont les outils d'accès facilité pour le téléchargement et la mise à disposition au sein du projet ? - De quel budget vais je avoir besoin pour gérer, stocker et partager un important volume de données ? - Comment estimer le coût de stockage et éventuellement le coût de l'archivage pérenne (données précieuses, patrimoniales… cf. archiver) ? > Estimation en fonction de la durée d'archivage.	Penser à vous référer à un projet précédent similaire s'il y en a un. Formaliser les réponses à ces questions au sein du plan de gestion de données..

COLLECTER ~ PRODUIRE	- Est-il possible de réutiliser les données qui existent plutôt que de collecter ou de créer des nouvelles données de forte volumétrie ?
	- Lors de la collecte et du stockage pensez aux bonnes pratiques pour réduire la taille des jeux de données dans la mesure du possible :
	  - Quelle est dans votre cas la résolution la plus optimale pour les images ?
	  - Parmi les formats ouverts, quels sont ceux qui sont nativement compressés ?
		Ressources utiles :
		Rechercher des données (fiche DoRANum)
		ou
		Rechercher des jeux de données (fiche Cirad)
		Pensez à utiliser des API pour accéder aux données de forte volumétrie.

STOCKER	Pour un gros volume de données : - À qui dois-je donner accès (consultation, modification) aux données en cours du projet ? - Certaines données seront-elles sensibles ? en accès restreint ? - Quelles solutions pour stocker mes données (bases de données, stockage déporté, cloud…) ? - À quelle fréquence ai-je besoin d'avoir accès aux données (disponibilité) ? - Comment concilier la règle de sauvegarde 3.2.1 et de sobriété numérique ?	Ressources utiles : Liste des mésocentres Fr (calcul et stockage) Consulter la direction informatique de votre établissement ou l'atelier de la donnée en proximité pour connaître les solutions proposées. Se renseigner auprès des centres de référence thématiques sur les outils mis à disposition. Par exemple, Huma-Num Box d'Huma-Num

TRAITER ~ ANALYSER	- Existe-t-il une puissance de calcul et des méthodes d'analyse ou outils adéquats pour ces volumes de données ? - Quels sont les outils permettant de nettoyer ses données (open Refine, Archifiltre) etc. ?	Ressource utile : COSTANZO, Lucia, 2023. Le nettoyage de données dans le processus de gestion des données de recherche. [en ligne]. 4 décembre 2023. DOI 10.5206/RHBN7291. [Consulté le 3 avril 2025]. *in*: La gestion des données de recherche dans le contexte canadien
PARTAGER ~ OUVRIR	- Est-ce que les données sont soumises à une exception légale d'ouverture ? - Parmi toutes les données produites, quelles sont celles qui viennent valider les résultats de vos recherches ? - Est-il possible d'effectuer un nouveau tri et une sélection de données pour ne garder que la partie essentielle au partage ? - Quel entrepôt est adapté à ma discipline et à la volumétrie de mes données ? - Est-il possible de compresser les données sans perte (partielle ou totale) sur le long terme ? Sinon discuter au sein de votre communauté pour décider si la compression avec perte est acceptable. - Est-il possible de découper le *dataset* en sous-ensembles plus faciles à télécharger?	Ressources utiles : Liste Entrepôts thématiques Si aucun entrepôt thématique ne convient, il est possible soit : • de déposer votre jeu de données dans un espace institutionnel ou dans l'espace générique de Recherche Data Gouv (voir p. 1 « 3. Spécificité du dépôt dans Recherche Data Gouv ») • de créer une notice signalétique dans l'entrepôt Recherche Data Gouv et ajouter le lien vers le stockage sur un serveur interne mais ouvert à l'extérieur en consultation. Il convient de s'assurer de la durée de maintenance de cet espace. Vous pouvez également contacter l'Atelier de la donnée présent sur votre territoire qui vous aidera à trouver une solution. L'utilisation d'une partie d'espace proposé par les mésocentres est possible dans certains lieux.
ARCHIVER	- Toutes les données du projet ont-elles vocation à être archivées de manière pérenne ? - Faut-il conserver les données brutes ? De quel espace de stockage faut-il disposer sur le long terme ? Qui le fournira ? - Quel lieu d'archivage est susceptible d'accueillir les gros volumes de données ?	Ressource utile : https://doranum.fr/stockage-archivage/stockage-et-archivage-fiche-synthetique_10_13143_0c4b-2743/ Fixer des durées de conservation en lien avec les archivistes. Il est également conseillé de mettre en place à l'échelle de l'unité ou du projet un « référentiel d'archivage » également appelé « tableau de gestion » (cf. *Référentiel de gestion des archives de la recherche*, section Aurore de l'AAF).
RÉUTILISER ~ DÉCOUVRIR	- Quelle partie des données (dérivées, brutes, code logiciel) sera rendue accessible ? - Comment un nombre potentiellement important de résultats sera-t-il affiché ? - Quels seront les outils mis à disposition pour faciliter la réutilisation des données ?