

Le stockage numérique

Objectif de la fiche : cette fiche a pour objectif :

- De donner des points de repères, bonnes pratiques
- De lutter contre les idées reçues
- De renvoyer vers des experts

Version : juillet 2020

Auteurs / contributeurs : Mikael Loaec, Sébastien Cat, Patrick Moreau, Eric Cahuzac, Lina Sbeih, Fanny Dedet, Jacques Foury, Alexandre Dehne-Garcia, Dimitri Szabo, Esther Dzalé-Yeumo

Définition : qu'est-ce que le stockage numérique ?

Le stockage correspond à l'ensemble des techniques et méthodes qui permettent de déposer et de conserver des informations de manière organisée sur un support numérique. La principale confusion faite, réside souvent dans l'amalgame entre les notions de stockage, de sauvegarde et d'archivage. Il s'agit cependant de termes distincts :

- L'archivage est une démarche de gestion et de conservation des informations à long terme.
- La sauvegarde est la duplication des informations pour prévenir toute disparition accidentelle.

Enjeux, Risques et Bénéfices

De manière générale, le stockage doit permettre de conserver de manière sécurisée la mémoire de l'information produite et fournir un accès facile à cette ressource à court ou moyen terme. Il doit principalement protéger des- risques de perte de cette information en garantissant son intégrité. La perte d'information peut être causée par une destruction (intentionnelle ou pas), un vol de matériel ou une dénaturation de l'objet numérique. La perte d'information peut aussi être causée par une difficulté d'accès à la ressource car l'identification de son existence a disparu ou la documentation afférente est insuffisante et ne permet plus sa compréhension/réutilisation. Pour prévenir ces risques, le stockage est recommandé tout au long du cycle de vie de la donnée et du projet de recherche et doit s'accompagner d'une documentation suffisante de la donnée. Ainsi, le stockage contribue à la "FAIRisation" de l'information, les métadonnées pouvant faciliter sa recherche (F), son accessibilité (A) et sa réutilisation (R).

Les enjeux actuels sont de taille, puisque le stockage fait face désormais à une problématique, résumée en "5V" :

- un **V**olume de données à traiter conséquent et croissant
- une **V**ariété d'informations importante (dans sa nature et sa structure)
- une **V**élocité dans la fréquence d'acquisition et de mise à jour de cette information.
- la garantie de la **V**aleur et de la **V**éracité de l'information

A celle-ci s'ajoute la question de la sensibilité de l'information stockée. En effet, les outils et méthodes de stockage doivent par exemple garantir la confidentialité des données personnelles (RGPD), et de toute- information [sensible](#) qu'elle provienne de recherches biomédicales, ou de santé animale. Par ailleurs, un stockage inadapté (exemple : accès trop lent pour les usages cibles) ou des coûts

disproportionnés– (financiers ou environnementaux) peuvent avoir un impact sur l'efficacité de la recherche.

Contexte / État des lieux

Les agents de l'institut ont à leur disposition plusieurs types d'offres. Elles peuvent être proposées à INRAE par des unités, des plateformes, la DipSO ou la DSI et à l'extérieur de l'institut par d'autres EPST ou établissements européens (voir tableau ci-dessous). Ces offres peuvent être payantes ou gratuites pour l'unité de l'agent. Ces offres ont des niveaux de technicité différents. Certaines sont directement utilisables par les agents, d'autres nécessitent l'accompagnement d'un informaticien pour leur mise en place. C'est pourquoi avant de faire le choix d'une solution de stockage, chaque agent est invité à solliciter les informaticiens en proximité afin de trouver la solution la plus adaptée à leur cas d'usage.

Liste des offres de stockage disponibles à INRAE selon leur usage :

Cette liste n'est pas exhaustive, d'autres solutions peuvent exister dans vos unités, notamment sur les plateformes ou en lien avec les tutelles associées. Pour celles-ci, il est conseillé de prendre contact directement avec leurs responsables.

Usage Type	ouverture extérieur/accès partenaires	Coût pour l'utilisateur final	Nom de l'Offre	Nécessite un accompagnement	opéré par
Partage/édition de documents	navigateur web	gratuit	Sharepoint	Formation nécessaire	DSI - Demande de service
Partage/édition de documents	navigateur web	gratuit	OneDrive	Non	DSI
Partage/édition de documents / Portail d'accès unifié	navigateur web et clients de synchronisation	gratuit	Nextcloud	Non	DSI - Demande de service
Partage/édition de documents	navigateur web et clients de synchronisation	gratuit	Stratus	Non	DSI
Base de données	clients de bases de données	gratuit	Bases de données "libres" à la demande	Nécessite la mise en place par un informaticien sur le serveur	DSI - Demande de service
Stockage/partage de fichiers	lecteur réseaux sur poste de travail	gratuit [1]	Partage CIFS	Nécessite la mise en place par un informaticien sur le poste de travail	DSI - Demande de service
Stockage/partage de fichiers	offre dédiée aux serveurs, montage de partition.	payant	Partage NFS	Nécessite la mise en place par un informaticien sur le serveur	DSI - Demande de service
Stockage/partage de fichiers	Clients lourds / navigateur web avec Nextcloud	payant	Stockage Objet	Nécessite la mise en place par un informaticien sur le serveur ou poste de travail	DSI - Demande de service
Partage Publication /	navigateur web	gratuit	Data INRAE	Formation utile	DipSO- Num4sci
Partage Publication /	navigateur web	gratuit	HAL INRAE	Formation utile	DipSO- Num4sci
Stockage de fichiers	Clients lourds	payant [2]	AgroDataRing	Nécessite la mise en place par un informaticien sur le serveur ou poste de travail	Groupeement d'unités
Stockage/partage code source et outils d'intégration continue	navigateur web	gratuit	ForgeMIA (GitLab)	Offre spécifique dédiée aux informaticiens	Unité MIATOneDrive

Partage/édition de documents	navigateur web	gratuit [3]	B2DROP (NextCloud)	Non	EUDAT
Partage de documents	navigateur web	gratuit	B2SHARE	Non	EUDAT
Stockage de code source	navigateur web	gratuit	SourceSup	Offre spécifique dédiée aux informaticiens	RENATER
Stockage de fichiers	Clients lourds	payant [2] au delà de 50 Go	FG-iRods France Grilles	Nécessite la mise en place par un informaticien sur le serveur ou poste de travail	France Grilles

[1] Payant au-delà de 15Go/ agent pour des espaces personnels, et nombre d'agents dans l'un ité X 5Go pour les espaces collectifs

[2] Participation aux investissements et support matériels

[3] Gratuit pour des espaces jusqu'à 20Go puis offre premium payante.

[*] pour les offres des plateformes INRAE se rapprocher directement des différentes plateformes.

Il faut aussi être conscient des caractéristiques de stockage qu'offrent les supports habituellement utilisés et des risques auxquels on s'expose. Le tableau suivant rappelle les principaux.

Principales caractéristiques des supports de stockage utilisés :

Support de stockage	Risques	Accès (interne/externe, vitesse)	Partage	Capacité de stockage	Terme (long/court)	Préconisations pour limiter les risques
Ordinateur professionnel	Sujet au piratage, au vol, à la détérioration et panne	Pas accessible à distance Pas limité par le réseau	Pas adapté au partage, nécessite l'utilisation d'un support externe ou d'internet (mail, cloud, ...)	Limitée	Court	Si données confidentielles ou restreintes , les chiffrer+ Sauvegarder en plus sur un autre support
Supports externes (disques, clés)	Sujet au vol, à la perte Durée de vie limitée	Pas accessible à distance Pas limité par le réseau	Facilement transportable	Limitée	Court	Si données confidentielles ou restreintes , alors stocker le disque externe dans une armoire fermant à clé ou dans un local protégé (bureau fermé à clé en l'absence de surveillance, zone à accès contrôlé, ...) Sauvegarder en plus sur un autre support
Serveurs institutionnels ou inter-EPST	Stockage fiable, durable et sécurisé	Accessible à distance Vitesse du réseau	Dépend des politiques liées à la sécurité de ou des établissements concerné(s)	Élevée, dépend du service négocié	Moyen /Long	Si confidentiel , chiffrement de bout en bout obligatoire. Si restreint , exigence label SecNumCloud délivré par l'ANSSI. Chiffrement obligatoire avec un logiciel qualifié par l'ANSSI
Cloud institutionnel ou inter-EPST	Stockage fiable, durable et sécurisé	Accessible à distance vitesse du réseau	Partage à priori facilité	Élevée, dépend du service négocié	Moyen /Long	Si confidentiel , chiffrement de bout en bout obligatoire. Si restreint , exigence label SecNumCloud délivré par l'ANSSI. Chiffrement obligatoire avec un logiciel qualifié par l'ANSSI

Notes:

Confidentiel: information dont la cible n'est pas nominative mais précise les structures ou entités destinataires de l'information, des personnes es-qualité et qui doit être diffusée via un canal dont l'accès est strictement contrôlé.

Restreint: information dont la cible est nominative ou précise les structures ou entités destinataires de l'information, des personnes es-qualité ayant besoin d'en connaître et qui doit être diffusée via un canal dont l'accès est strictement contrôlé.

Stockage et calcul/traitement de la donnée

Souvent les offres de calcul/traitement fournissent un stockage temporaire et de capacité limitée. Ils vous faudra donc transférer vos données d'entrées ainsi que vos résultats vers un stockage pérenne. Ces transferts peuvent ralentir et/ou complexifier le parcours de la donnée. Le risque lié à cette complexité est la possible perte de données (exemple oubli de transfert des résultats sur un support pérenne).

Quelques leviers qui peuvent aider à la réflexion sont les suivants :

- Rapprocher physiquement le stockage de vos données du calcul, par exemple dans le même bâtiment ou sur le même réseau local
- Utiliser un protocole de transfert robuste et adapté au réseau interne et internet. Par exemple préférer iRods, Swift, S3 au CIFS (partage Windows) ou au NFS (partage Unix/Linux).
- Utiliser une méthode de vérification de l'intégrité des données transférées.
- La manière dont les données sont structurées (en base de données SQL, NoSQL ou en fichiers) peut avoir un impact sur la vitesse.
- Adapter vos calculs aux capacités de transfert réseaux et aux protocoles utilisés : par exemple découper vos données et adapter vos algorithmes avant le traitement si besoin. Autre exemple, certains protocoles sont plus adaptés à l'envoi de gros fichiers, dans ce cas faire une archive de type zip ou tar.gz).
- Privilégier certains logiciels permettant des calculs directement depuis le stockage distant en chargeant les données en mémoire (ex: Spark et Stockage Objet) ce qui améliore grandement les temps de traitement.

Chaque plateforme a ses propres règles. Il est donc essentiel de prendre contact avec les gestionnaires des plateformes ciblées pour évoquer ces leviers avant de les mettre en œuvre. Il est nécessaire de vous assurer de la compatibilité des choix que vous retiendrez par rapport à leurs offres. Ce sont des acteurs essentiels pour la réussite de vos traitements.

Recommandations / Bonnes Pratiques

Il est nécessaire d'identifier la nature des données et l'usage qui en sera fait pour effectuer un choix adapté d'espace de données cible.

Le choix du stockage peut en effet varier en fonction de l'étape du cycle de vie du projet de recherche, des données et de nombreux autres critères. Par exemple, le niveau de sécurisation, le protocole d'accès aux données, la performance, le lieu de stockage et la connectivité au réseau du site de production des données, la volumétrie, le niveau de sensibilité des données (cf. fiche données sensibles), la nature structurée ou non des données, les collaborations internes/externes souhaitées aux différentes étapes (collecte, traitement/analyse, publication), etc.

Pour bien aborder les questions de stockage nous recommandons avant tout d'anticiper cette question dès la phase de conception du projet en rédigeant un plan de gestion des données (PGD). Ceci facilitera les recommandations suivantes :

1. Prévoir de quelle manière les données seront décrites (avec quelles métadonnées et quels vocabulaires : cf. fiche métadonnées), structurées, et enfin organisées (cf. fiche organisation et nommage des fichiers de données)

2. Toujours stocker les données accompagnées des métadonnées. Permettant ainsi de les identifier, archiver, interpréter et réutiliser ultérieurement. Les métadonnées qui accompagnent les données doivent aussi permettre à tout moment d'en gérer la fin de vie.
3. Recourir le plus tôt possible à un identifiant unique et pérenne qui renvoie vers l'ensemble métadonnées et données associées (notion de dataset par exemple dans Data INRAE <https://data.inrae.fr>).
4. Prendre en compte les coûts de stockage des données dans le plan de financement de vos projets de recherche
5. Identifier la/les personnes responsable(s) des données.

Qui Contacter / Ressources pour aller plus loin

Groupe de travail stockage et archivage INRAE :

[chantier-stockage-archivage-inra@inrae.fr](chantier-stockage-archivage-inra@inrae.fr)

Offre institutionnelle DSI :

[Formulaire en ligne pour contacter le groupe stockage DSI.](#)

Offre institutionnelle Agrodataring :

<https://www.ingenum.inra.fr/adr/contact>

Offre institutionnelle Data INRAE (<https://data.inrae.fr>) : datainrae@inrae.fr