Testez vos connaissances sur les entrepôts de données

Bienvenue dans ce jeu où vous devrez répondre à 22 questions sur les entrepôts de données (questions générales, questions plus ciblées sur le choix d'un entrepôt, sur le dépôt lui-même, sur *Recherche DataGouv*).

Attention, certaines questions peuvent comporter plusieurs réponses, sachant que chaque bonne réponse compte et rapporte un point.

Le nombre de bonnes réponses attendues est précisé entre parenthèses.

À vous de jouer !

Question 01/22

Est-ce obligatoire de diffuser les données de recherche ? (1 réponse attendue)

- ☐ Oui, pour tous les projets
- ☐ Oui, dans le cas où l'activité de recherche est financée au moins à50% par des fonds publics
- ☐ Non, il n'y a aucune obligation
- ☐ Non, si vous avez une dispense signée de votre directeur de laboratoire

Correction :

Il est obligatoire de diffuser les données de recherche dans le cas où l'activité de recherche est financée au moins à 50% par des fonds publics, et l'idéal est de déposer ces données dans un entrepôt.

Depuis la loi pour une République numérique de 2016, les données de recherche "achevées" sont assimilées à des données administratives et font donc l'objet d'un principe d'ouverture "par défaut", c'est-à-dire qu'elles sont censées être publiées et rendues accessibles sur internet. Il existe toutefois des exceptions légales au partage et à la diffusion des données (données personnelles, sensibles, protégées par le droit de propriété intellectuelle...).

Dès 2018, le gouvernement invite dans son premier Plan national pour la science ouverte les chercheurs « à déposer les données dans des entrepôts de données certifiés, dont la gouvernance et les règles de propriété intellectuelle seront conformes aux bonnes pratiques ». En 2021, l'une des mesures du Deuxième Plan national pour la science ouverte est la mise en œuvre de l'obligation de diffusion des données de recherche financées au moins pour moitié par des fonds publics.

Question 02/22

À quelle étape du cycle de vie des données se situe le dépôt des données dans un entrepôt ? (1 réponse attendue)

- ☐ Stockage
- ☐ Traitement et analyse
- ☐ Archivage pérenne
- ☐ Accès et partage

Correction :

Le dépôt des données dans un entrepôt intervient le plus souvent à la fin du projet, après la sélection des jeux de données à partager.

Pour en savoir plus sur le cycle de vie des données en général, vous pouvez consulter cette ressource sur DoRANum .

Question 03/22

Qu'est-ce qu'un entrepôt de données ? (1 réponse attendue)

- ☐ Un lieu de stockage
- ☐ Une archive de conservation sur le long terme
- ☐ Un service en ligne de dépôt et de partage

Correction :

Un entrepôt est bien un **service en ligne** (et non un lieu) permettant de déposer et partager des données de recherche.

Il ne faut pas confondre l'**entrepôt** dont le but est le **partage des données (avec toute la communauté scientifique)** et l'archive pérenne destinée à conserver les données sur le long terme.

Attention, l'entrepôt n'est pas un lieu de stockage des données durant le projet. Généralement, des solutions de stockage existent au niveau de votre institution, notamment des plateformes collaboratives permettant de

partager les données uniquement avec des partenaires pendant la durée du projet.

Pour mieux comprendre la différence entre stockage, partage et archivage pérenne, vous pouvez consulter cette ressource sur DoRANum.

Question 04/22

Quels sont les intérêts de déposer ses données dans un entrepôt ? (2 réponses attendues)

- ☐ Valoriser ses travaux de recherche (validation par les pairs et citation)
- ☐ Rendre les données réutilisables
- ☐ Alléger l'espace disque de son ordinateur
- ☐ Archiver ses données à long terme

Correction :

Le but d'un entrepôt n'est pas d'archiver les données à long terme.

Le partage des données dans un entrepôt présente plusieurs intérêts pour l'équipe de recherche elle-même, comme la valorisation des travaux de recherche qui facilite la validation par les pairs et la citation.

Il permet d'autre part à la communauté de recherche de diminuer les coûts de reproduction des données et de trouver des données réutilisables pour de nouveaux projets.

Question 05/22

Dans cette liste, cherchez l'intrus (1 réponse attendue).

- ☐ Pangaea
- ☐ Nakala
- ☐ Dryad
- ☐ ResearchGate
- ☐ Ortolang

Correction :

ResearchGate est un réseau social scientifique. Il n'est pas conçu pour y déposer ses données.

Les autres propositions concernent bien des **entrepôts de données**.

NAKALA est un entrepôt français dédié aux sciences humaines et sociales.

ORTOLANG est un entrepôt national français en sciences du langage.

DRYAD est un entrepôt pluridisciplinaire international qui s'adresse plus aux sciences biologiques et médicales.

PANGAEA est un entrepôt en sciences de la Terre et de l'environnement.

Question 06/22

Quel type d'entrepôt privilégier ? (Classer les propositions par ordre de priorité)

Correction :

Il est conseillé de déposer vos données **en priorité dans un entrepôt disciplinaire** s'il en existe un pour votre communauté. Dans le cas contraire, vérifiez si votre institution a mis en place un entrepôt pour ses chercheurs.
Si aucune de ces deux solutions n'est possible, pour la communauté française il est recommandé de déposer ses données dans l'entrepôt national *Recherche Data Gouv*.

Question 07/22

Comment trouver un entrepôt adapté à ses données ? (4 réponses attendues)

- ☐ Je consulte des annuaires de référence
- ☐ Je me réfère à ce qui est préconisé par le financeur
- ☐ Je me renseigne auprès de mon éditeur
- ☐ Je discute avec mes collègues sur les pratiques de ma communauté
- ☐ Je prends contact avec l'atelier de la donnée local

Correction :

Il existe plusieurs solutions pour vous aider à trouver un entrepôt adapté à vos besoins :
- consulter des annuaires de référence (Cat OPIDoR, re3data, Fairsharing…),
- voir ce que préconise votre financeur
- discuter avec vos collègues sur les pratiques en usage dans votre communauté
- vous faire accompagner, par exemple, par un atelier de la donnée s'il en existe un proche de chez vous.

Par contre, il n'est pas recommandé de suivre les préconisations de votre éditeur car il peut s'agir d'un entrepôt qui déroge aux principes de partage gratuit et ouvert des données.
Voici une infographie interactive qui récapitule ce qu'il y a à savoir à ce sujet.

Question 08/22

Sur quels critères se baser pour choisir un entrepôt ? (2 réponses attendues)

- ☐ La durée de conservation des données garantie par l'entrepôt
- ☐ L'attribution d'un identifiant pérenne
- ☐ L'attribution d'une récompense à la fin du dépôt
- ☐ L'hébergement aux États-Unis
- ☐ L'attribution de licences permissives (CC0)

Correction :

Pour choisir un entrepôt, vous pouvez vous baser sur plusieurs critères, notamment :

- l'attribution d'un identifiant pérenne (exemples : DOI, ARK, SWHID...) : pour relier vos données à vos publications, permettre un accès sûr et pérenne aux données, en faciliter la réutilisation...
- la durée de conservation que garantit l'entrepôt : convient-elle à vos besoins ?

Attention au lieu d'hébergement de l'entrepôt choisi car ce sera **le droit du pays** qui s'appliquera.

Concernant la licence, **le droit français demande d'appliquer a minima une licence reconnaissant la paternité** (comme la Licence Ouverte Etalab ou la CC-By). Si vos données nécessitent une licence plus restrictive ou spécifique (comme ODbL ou GNU), il faut vérifier que l'entrepôt choisi la propose.

Pour en savoir plus, vous pouvez consulter cette ressource sur DoRANum.

Question 09/22

Dans quels types d'entrepôt est-il recommandé de déposer ses données ? (2 réponses attendues)

- ☐ Un entrepôt certifié
- ☐ Un entrepôt de confiance
- ☐ Une base de données accessible sur le site web du laboratoire
- ☐ Un entrepôt labellisé éco-responsable

Correction :

Il est recommandé de déposer ses données dans :

- un entrepôt certifié

- un entrepôt de confiance, c'est-à-dire dont la qualité et le sérieux sont validés par la communauté scientifique qui l'utilise.

Par contre, les bases de données sur un site web ne permettent pas l'interopérabilité et ne sont pas reconnues comme étant des entrepôts.

Question 10/22

Qu'est-ce que le CoreTrustSeal ? (2 réponses attendues)

- ☐ Un entrepôt en zoologie
- ☐ Un organisme de certification
- ☐ Un centre de télécommunication spatial facilitant l'interopérabilité
- ☐ Un label garantissant des critères de conformité et de fiabilité

Correction :

Eh non, CoreTrustSeal n'est pas un entrepôt de zoologie qui accueillerait les données sur les morses, phoques et otaries (*seal* en anglais).

CoreTrustSeal est un des organismes de certification des entrepôts de données les plus connus. Il garantit la fiabilité et la durabilité du dépôt de données ainsi que l'archivage et le partage à long terme. Ce label signifie que l'entrepôt remplit des critères de conformité.

Question 11/22

En quoi consiste la curation des jeux de données ? (2 réponses attendues)

- ☐ Vérifier la conformité et la complétude des métadonnées
- ☐ Supprimer les jeux de données et métadonnées obsolètes
- ☐ Nettoyer les échantillons collectés par votre laboratoire
- ☐ Vérifier la présence d'une licence

Correction :

La curation est une étape importante qui consiste notamment à :

- vérifier la conformité et la complétude des métadonnées
- vérifier la présence d'une licence.

Il revient au déposant de retirer les jeux de données obsolètes de l'entrepôt et de le signaler au niveau des métadonnées qui, elles, restent consultables. Il faut penser aussi à déclarer l'obsolescence au niveau de l'identifiant pérenne attribué aux jeux de données concernés.

Question 12/22

Quels jeux de données est-il recommandé de partager ? (3 réponses attendues)

- ☐ Les données coûteuses à produire
- ☐ Les échantillons biologiques
- ☐ Les données liées à une publication
- ☐ Le code source permettant de lire/traiter les données
- ☐ Toutes les données liées à un projet

Correction :

Il faut **sélectionner les jeux de données** à partager en fonction de leur valeur scientifique, de leur coût de production, de leur réutilisabilité.

D'autre part, les jeux de données nécessaires à la compréhension et à la reproductibilité des résultats de recherche publiés dans un article doivent être accessibles. Le cas échéant, le code source peut être un autre élément important à fournir.

Rappelons que déposer des données dans un entrepôt se fait selon le principe « aussi ouvert que possible, aussi fermé que nécessaire ».

Les échantillons biologiques eux-mêmes ne peuvent pas être partagés dans un entrepôt, par contre le descriptif numérique peut l'être. Il devra être suffisamment détaillé et référencé.

Question 13/22

Quels jeux de données peut-on diffuser librement ? (2 réponses attendues)

- ☐ Les données géographiques
- ☐ Les données personnelles après anonymisation
- ☐ Les données personnelles après pseudonymisation
- ☐ Les œuvres de l'esprit
- ☐ Toutes les données produites sur fonds publics

Correction :

Les données géographiques font partie, avec les données environnementales, des données dont la diffusion libre, gratuite et ouverte est obligatoire par principe.

Les données produites sur fonds publics peuvent être partagées librement si elles ne font pas l'objet de restrictions particulières (ZRR, secret défense, données confidentielles...).

Les données personnelles peuvent être partagées à condition d'être anonymisées. C'est un processus irréversible qui garantit la sécurité et la confidentialité.

Par contre, la pseudonymisation n'est pas suffisante : les données restent soumises au RGPD et à la loi Informatique et Libertés.

Pour en savoir plus sur les différences entre anonymisation et pseudonymisation, voir le site de la CNIL.

Les œuvres de l'esprit sont protégées par le droit de la propriété intellectuelle et ne peuvent être partagées sans le consentement de l'auteur.

Pour en savoir plus sur la communicabilité des données, vous pouvez consulter ce logigramme.

Pour en savoir plus sur les données géographiques, vous pouvez consulter ce module sur la Directive INSPIRE.

Question 14/22

Que faut-il faire pour préparer le dépôt des données ? (3 réponses attendues)

- ☐ Sélectionner les données à déposer
- ☐ Enrichir les métadonnées en fonction du standard
- ☐ Garder les formats de données d'origine
- ☐ Préparer si nécessaire les codes sources
- ☐ S'affilier à la Caisse des Dépôts

Correction :

La première chose à faire est de sélectionner les jeux de données à déposer dans l'entrepôt.

Au moment du dépôt, il faudra veiller à **renseigner au maximum les métadonnées**.

L'idéal est d'utiliser un **standard** de métadonnées disciplinaire adapté à vos besoins.

Si des **codes sources** sont nécessaires pour lire et traiter les données, il faudra aussi les partager en les déposant dans un entrepôt dédié comme Software Heritage, en faisant bien le lien avec les données via le PID.

Il faut privilégier des **formats ouverts** ou convertir, si possible, le format propriétaire en format ouvert.

Pour préparer au mieux vos données, vous pouvez consulter cette checklist.

Question 15/22

Voici une liste d'éléments conseillés pour compléter un dépôt. Quel est l'intrus ? (1 réponse attendue)

- ☐ Des métadonnées enrichies
- ☐ Un fichier Readme.txt
- ☐ Un dictionnaire de données
- ☐ Le curriculum vitae de l'auteur
- ☐ Les identifiants pérennes des publications associées

Correction :

Pour une compréhension et une réutilisation des données optimales, il est conseillé de compléter le dépôt par un certain nombre d'éléments :

- des métadonnées richement renseignées,
- un fichier Readme.txt
- un dictionnaire de données (se définit comme un référentiel de métadonnées qui renseigne sur le contexte d'une base de données et qui fournit les informations nécessaires pour pouvoir l'interpréter).
- PID des publications (articles, data papers, plan de gestion de données...) et logiciels associés.

Pensez également à renseigner votre identifiant ORCID, mais le CV n'est pas indispensable ;-)

Question 16/22

Renseigner les métadonnées obligatoires et facultatives est très important.

Pourquoi ? (2 réponses attendues)

- ☐ Cela favorise l'interopérabilité
- ☐ Cela facilite la réutilisation des données
- ☐ Cela permet d'obtenir un financement complémentaire
- ☐ Parce que !

Correction :

Il est fortement recommandé de **renseigner les métadonnées facultatives en plus des métadonnées obligatoires** car :

- cela favorise une plus grande interopérabilité
- cela permet une meilleure compréhension des données
- cela facilite la réutilisation de ces dernières.

Eh non, il ne faut pas rêver... aucun financement supplémentaire n'est attribué pour cela, mais les réutilisateurs de vos données vous diront merci !

Question 17/22

Qu'est-ce qui rime avec métadonnées ? (4 réponses attendues)

- ☐ Sobriété pour efficacité
- ☐ Plus il y en a, mieux c'est
- ☐ À renseigner pour tout relier (publications, logiciels, autres jeux de données…)
- ☐ Documenter pour être moissonnées
- ☐ Traçabilité et éternité assurées

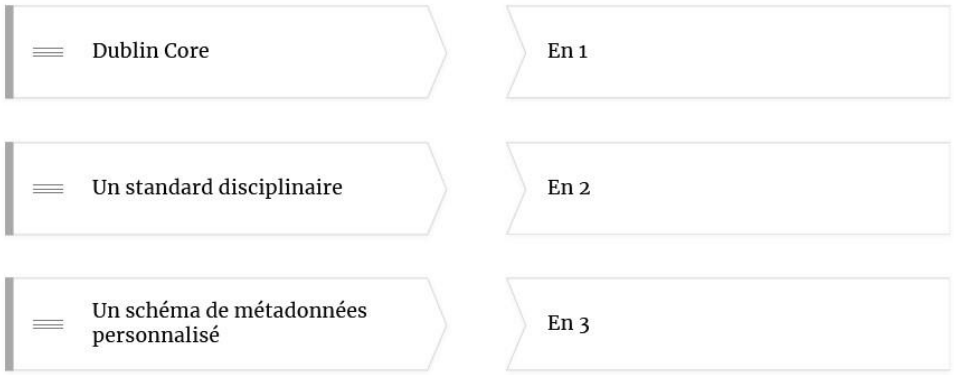
Correction :

La sobriété n'est pas de mise pour ce qui concerne les métadonnées. Au contraire, plus elles sont richement renseignées, mieux c'est !

Cela permet de correctement relier tous les travaux de recherche à son auteur (publications, logiciels, autres jeux de données…), de rendre les données faciles à trouver (le moissonnage étant basé sur les métadonnées) et de garder une trace, même si les données sont obsolètes et supprimées.

Question 18/22

Quel standard de métadonnées privilégier ? (Classer les propositions par ordre de priorité)



Correction :

Il est recommandé d'utiliser **prioritairement un standard de métadonnées disciplinaire** car il répondra mieux à vos besoins. S'il n'en existe pas dans votre domaine, vous pouvez utiliser le standard générique international Dublin Core. Il est beaucoup utilisé car il offre une grande interopérabilité, mais le nombre de champs proposé est relativement restreint.

Il est toujours possible de créer un schéma de métadonnées personnalisé, mais celui-ci rend vos données moins interopérables.

Pour connaître les éléments essentiels sur les métadonnées, vous pouvez consulter ce module interactif.

Question 19/22

Est-il possible de déposer des données en accès restreint ? (2 réponses attendues)

- ☐ Non, qui dit FAIR dit ouvert
- ☐ Oui, si l'entrepôt propose un protocole d'authentification et/ou une autorisation d'accès
- ☐ Oui, mais uniquement pendant la durée du projet
- ☐ Oui, avec une durée d'embargo

Correction :

Oui, il est possible de déposer des données en accès restreint lorsque c'est nécessaire (par exemple dans l'attente du dépôt d'un brevet ou d'une publication).

Il est bien sûr tout à fait possible de restreindre l'accès aux données durant le projet mais pas seulement. On peut définir un accès restreint au long cours pour des données accessibles uniquement à une communauté définie. Pour ce faire, il faut veiller à choisir un entrepôt qui permet de définir un embargo ou qui propose un protocole d'authentification et/ou une autorisation d'accès.

À noter que **les principes FAIR ne sont pas antinomiques avec un accès restrictif**. Les données doivent être partagées selon le principe « aussi ouvert que possible, aussi fermé que nécessaire ».

Pour en savoir plus sur les principes FAIR, vous pouvez consulter cette ressource DoRANum.

Question 20/22

Qui peut vous accompagner pour déposer ? (2 réponses attendues)

- ☐ Les ateliers de la donnée
- ☐ Le directeur de votre laboratoire
- ☐ L'administrateur ministériel des données
- ☐ Les documentalistes de votre université ou laboratoire

Correction :

Vous pouvez solliciter les documentalistes de votre université ou de votre laboratoire ou bien faire appel à un atelier de la donnée de l'écosystème *Recherche Data Gouv* s'il en existe un localement.

Le directeur de votre laboratoire n'est pas l'interlocuteur à privilégier, car il a un rôle de coordination à une échelle supérieure tout comme l'administrateur ministériel des données.

Question 21/22

Dans quels cas déposer ses données dans l'entrepôt *Recherche Data Gouv* ? (2 réponses attendues)

- ☐ Lorsqu'il n'existe pas d'entrepôt disciplinaire pour sa communauté
- ☐ Pour tous les projets de recherche français
- ☐ Lorsqu'il n'existe pas d'entrepôt institutionnel
- ☐ Pour tous les projets internationaux impliquant au moins un partenaire français

Correction :

Il n'y a pas d'obligation de déposer dans l'entrepôt national *Recherche Data Gouv* pour les projets français ou les projets internationaux impliquant au moins un partenaire français.

Il a été mis en place pour pallier à un manque **quand il n'existe pas d'entrepôt disciplinaire ou institutionnel**.

Question 22/22

À quel endroit déposer dans *Recherche Data Gouv* quand les co-auteurs sont affiliés à plusieurs institutions ? (1 réponse attendue)

- ☐ En priorité dans l'espace générique
- ☐ Dans chaque espace de rattachement des co-auteurs
- ☐ Dans un seul espace institutionnel choisi par l'ensemble des co-auteurs

Correction :

Les co-auteurs doivent se mettre d'accord sur le choix d'un espace institutionnel.

Il ne faut surtout pas multiplier les dépôts.

Par contre, il est utile de **signaler** un dépôt dans d'autres espaces.

À noter qu'il est possible de faire le dépôt dans l'espace générique s'il n'existe aucun espace institutionnel rattaché aux co-auteurs.