La science ouverte : aspects généraux

Qu'entend-on par Science Ouverte ?

Le comité national pour la science ouverte (CoSO) définit la science ouverte comme « la diffusion sans entrave des résultats, des méthodes et des produits de la recherche scientifique ». Elle concerne non seulement les publications mais aussi les données et les codes et logiciels issus des travaux des travaux de recherche disciplinaire, et s’applique aussi bien aux sciences techniques et mécaniques (STM) qu’aux sciences humaines et sociales (SHS). Initiative internationale, elle est ordonnée en France depuis 2021 par le deuxième plan national pour la science ouverte (PNSO) qui structure selon plusieurs axes les actions des institutions et organisations dans le domaine, et permet d’obtenir un financement pour un projet de recherche par le biais des infrastructures traditionnelles ou directement par le fonds national pour la science ouverte (FNSO).

- https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte-pnso/ Pour découvrir le deuxième plan national pour la science ouverte en détails.

Les principes FAIR

« FAIR » est un acronyme pour *facile à trouver, accessible, interopérable* et *réutilisable*. Ce sont les quatre grandes lignes qui garantissent une insertion optimale de la donnée dans l'écosystème numérique de la recherche. En d'autres termes, il s'agit de faire en sorte que la données et ses métadonnées associées soient :

1. Faciles à trouver pour les humains comme pour les ordinateurs, avec un identifiant pérenne et des métadonnées descriptives, dans un entrepôt de données ;
2. Toujours disponibles et accessibles. Dans le cas où la donnée est protégée, ce sont ses métadonnées qui restent ouvertes. L'accès doit pouvoir se faire à travers un protocole de communication standard, libre et ouvert, avec un accès par authentification si besoin ;
3. Généralement intégrées à d'autres données et sémantiquement compréhensible, ce qui permet leur échange et leur réutilisation. Cela implique une description par un vocabulaire contrôlé, le respect des principes FAIR et des liens vers d'autres données ;
4. Suffisamment décrites et partagées avec les licences les moins restrictives, ce qui permet la réutilisation la plus large possible avec l'intégration la moins lourde avec d'autres sources de données. Le partage des données suit les standards (schémas) de la communauté scientifique.

Quelques liens :

- https://www.go-fair.org/fair-principles/ (en anglais) Ce site donne des définitions précises pour chaque principe et des outils pour les mettre en œuvre.
- https://www.ccsd.cnrs.fr/principes-fair/ Le CCSD synthétise ces informations et donne les protocoles et standards adéquats pour chaque principe.

Qu'est-ce qu'un entrepôt de données ?

Il s’agit d’un espace de stockage pérenne et sécurisé pour les données de la recherche, avec des spécificités en fonction des disciplines et exigences de l’organisme en charge de l’entrepôt, mais toujours dans le respect des principes FAIR. Ces entrepôts peuvent être institutionnels, nationaux ou internationaux, ou disciplinaires.

Quelques liens :

- https://recherche.data.gouv.fr/fr/aide-en-ligne *Recherche Data Gouv* est un entrepôt de données interdisciplinaire mis à disposition de la communauté scientifique par le ministère de l’Enseignement Supérieur et de la Recherche. Son aide en ligne est destinée à aider les chercheurs à s’orienter parmi les entrepôts existants et, le cas échéant, à déposer leurs données sur sa plateforme.
- https://coop-ist.cirad.fr/gerer-des-donnees/deposer-des-donnees-dans-un-entrepot/1-qu-est-ce-qu-un-entrepot-de-donnees-de-recherche La Délégation en information scientifique et technique du Cirad a mis en place ce site pour répondre aux questions des chercheurs et professionnels de la recherche en termes, notamment, de gestion de données. Cette série d’articles revient sur le travail préparatoire en amont du dépôt des données de la recherche.
- https://doranum.fr/depot-entrepots/depot-et-entrepots-fiche-synthetique_10_13143_a3d4-7553/ DoRANum est une plateforme d’apprentissage destinée à qui s’intéresse à la science ouverte, mise à disposition par l’INIST-CNRS et le GIS « Réseau URFIST ». Dans cette fiche synthétique, il est question des étapes préliminaires à un dépôt de données, du choix de l’entrepôt à la vérification.
- https://www.ouvrirlascience.fr/selectionner-un-entrepot-thematique-de-confiance-pour-la-diffusion-des-donnees-de-recherche-note-methodologique/ Le CoSO donne quelques pistes pour s’orienter dans le choix d’un entrepôt « de confiance » pour les données de la recherche.

Pourquoi faire un DMP ?

Un plan de gestion de données (PGD, ou *Data Management Plan* en anglais) est tout d'abord un outil qui aide à prévoir dès l'amorce d'un projet de recherche la gestion qui sera faite des données produites. C'est aussi, depuis quelques années, une obligation pour la plupart des grands organismes de financement de la recherche comme l'ANR.

Très concrètement, c'est un document, associant les grands acteurs du projet et qui peut être évolutif, recensant les étapes du traitement des données, les standards adoptés, jusqu'au dépôt pour sauvegarde et partage dans un entrepôt de données. Il permet aux chercheurs de maintenir à jour leurs outils et protocoles en vue de l'analyse et de l'archivage des données produites.

Quelques liens :

- https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm La commission européenne a mis en ligne un guide reprenant la définition et les objectifs d'un plan de gestion de données, avec un modèle adapté aux projets qu'elle finance.
- https://dmp.opidor.fr/ De manière plus générale, OPIDoR est un outil de référence en France, mis à disposition par le CNRS et l'INIST-CNRS, permettant d'une part d'accéder à des modèles et exemples de DMP pour aider à la rédaction, et proposant un grand nombre de ressources sur comment rédiger et mettre en œuvre un DMP pour son projet de recherche.

Les aspects législatifs

Introduction

L’ouverture des données peut parfois être entravée par d’autres législations qui viennent se superposer aux recommandations pour la science ouverte : c’est le cas notamment pour les données à caractère personnel ou concernant le statut du producteur de la donnée.

Quelques premiers liens pour s’orienter :

- https://doranum.fr/aspects-juridiques-ethiques/ Le site de formation DoRANum propose une fiche synthétique, une vidéo et des ressources supplémentaires pour s’initier aux questions de droit du partage et de réutilisation.
- https://ethiquedroit.hypotheses.org/ Pour les chercheurs en SHS, notamment celles et ceux s’occupant de matériaux modernes et contemporains, le carnet de recherche Hypotheses *Éthique et droit* fait un état des lieux sur plusieurs de ces questions, telles que le droit à l’image ou plus généralement l’accompagnement dont peuvent bénéficier les chercheurs.

RGPD et données sensibles

Depuis 2016, le *Règlement général sur la protection des données* encadre au niveau européen l'usage qui peut être fait des données à caractère personnel et sensibles. Le texte complet et traduit peut être consulté sur le site de la CNIL : https://www.cnil.fr/fr/reglement-europeen-protection-donnees.

- https://www.cnil.fr/fr/comprendre-le-rgpd Cette même commission propose une série d'articles et de ressources qui donnent les clefs pour comprendre les obligations qui sont celles du chercheur dans le cadre de ses activités. La CNIL peut également être directement consultée pour des questions plus précises.

Droits de propriété intellectuelle

En fonction du statut du producteur d'une donnée scientifique, les droits moraux et d'exploitation ne seront pas les mêmes : cela pose la question, d'ordre juridique, des droits à céder ou acquérir pour tout partage ou toute réutilisation lorsque la licence n'est pas explicitement spécifiée.

Quelques liens :

- https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0347 Parmi les instruments juridiques mis à disposition par l'OCDE, cette recommandation sur l'accès aux données de la recherche financée sur fonds publics, allant jusqu'aux codes et logiciels, promeut notamment l'usage des licences ouvertes pour la protection de l'auteur.
- https://www.cairn.info/revue-francaise-d-administration-publique-2018-3-page-479.htm Cet article d'Aurélien Camus revient sur les définitions de propriété intellectuelle, notamment dans le cas d'une donnée administrative ou d'une œuvre de l'esprit. Il s'agit notamment de la protection des données publiques vis-à-vis d'un usage commercial.
- https://hal.science/hal-02791224 Ce guide assez complet reprend au cas par cas les conditions de diffusion des données pouvant être ouvertes (ou pas). Sont également abordées les modalités de diffusion, avec des points d'attention sur les données à caractère personnel et statistiques.

Qu'est-ce qu'un référentiel ?

Un référentiel est un répertoire d’autorités permettant d’identifier sans ambiguïté une entité nommée (personne, objet, etc.), géographique ou chronologique. Cela passe par l’attribution d’un identifiant pérenne à l’entité (de type DOI, ARK, ou autre) ainsi selon les cas qu’une description, localisation, étendue.

Quelques liens :

- https://doranum.fr/identifiants-perennes-pid/ La fiche synthétique et les ressources de DoRANum sur la question.
- https://isidore.science/vocabularies Le portail développé par la TGIR HumaNum, Isidore, recense et décrit les référentiels qu’il interroge pour apporter des données de bonne qualité à ses utilisateurs.

Qu'est-ce qu'une donnée de qualité ?

Définition

Une donnée de qualité est avant tout une donnée documentée (à l’aide d’un schéma de métadonnées) et pérennisée (grâce aux outils de stockage). Ce sont ces critères qui en feront une donnée pertinente et réutilisable dans le web sémantique.

Qu'est-ce qu'une métadonnée ?

Une métadonnée est une information rattachée à la donnée qui permet de la décrire (littéralement *une donnée sur la donnée*) : elle renseigne le contenu intellectuel, le contexte de production, les caractéristiques techniques des fichiers et des données ainsi que les propriétés et droits d'usage. On en trouve différents types : les métadonnées de description, qui donnent une idée précise du contenu de la ressource ; les métadonnées de gestion qui permettent d'accéder à la ressource (métadonnées d'identification, de provenance et de contexte) ; et enfin les métadonnées de préservation qui garantissent un accès pérenne au document et sa compréhension dans le long terme (métadonnées techniques, de structure, de droit).

Quelques liens :

- https://www.opendatasoft.com/fr/blog/metadonnees-ce-quil-faut-savoir-avant-de-publier-vos-donnees/ Le blog de la société OpenDataSoft reprend de façon plus détaillée ces éléments, et aborde la question des schémas de métadonnées. Il s'agit d'aider les producteurs à assurer une description optimale de leurs données avant ouverture.
- https://doranum.fr/metadonnees-standards-formats/ La fiche synthétique et les ressources de DoRANum sur la question.

Optimiser la pérennité des stockages

La question posée, en plus de celle des infrastructures, est celle du format des données qui doit être compatible avec les technologies en place. Un format de données est une convention utilisée pour représenter des données, des informations représentant un texte, une page, une image, un son, un fichier exécutable, etc. Il peut être ouvert (sa spécification est alors publiquement accessible) ou fermé (lorsque sa spécification est secrète). Éventuellement, un format peut aussi être normalisé (par une institution publique ou internationale, à l'exemple de ISO ou du W3C) lorsqu'il n'est pas propriétaire (soit s'il a été élaboré par une entreprise dans un but essentiellement commercial). Exemples de formats de métadonnées : CSV, XML, JSON, RDF…

- https://doranum.fr/stockage-archivage/ La fiche synthétique et les ressources de DoRANum sur la question.

Qu'est-ce que l'identité numérique des programmes ?

La présence ou visibilité numérique d’un programme de recherche est ce qui permet de replacer les différentes productions dans le contexte du projet avec lequel elles sont en lien. Il s’agit de mettre en relation les données créées lors d’un même programme.

Données et identité numérique des chercheurs

L’ouverture des données participe à la présence ou visibilité numérique du chercheur qui en est à l’origine. Cela passe essentiellement par son identification formelle dans un référentiel de personne et qui permet d’agréger toutes les données qui sont associées à son activité : il est en effet possible de créer des liens entre les entrepôts de données, les publications et un profil ORCID par exemple, lorsque les métadonnées ont été correctement renseignées.

Quelques liens :

- https://doranum.fr/identifiants-perennes-pid/ La fiche synthétique et les ressources de DoRANum sur la question.
- https://hal.umontpellier.fr/hal-03127068 Ce support de webinaire revient sur les différents moyens d’identifier un chercheur et son activité en ligne, faisant également un point sur le concept d’*e-réputation*.

Qu'est-ce qu'un data paper ?

Un data paper est une publication scientifique qui décrit précisément un jeu de données ouvert, et informe la communauté scientifique de son existence, de ses modalités et de son potentiel de réutilisation. En ce sens, il contient une partie descriptive avec des éléments communs aux articles classiques et des éléments spécifiques liés aux données. Les données, quant à elles, peuvent être intégrées dans l'article ou être déposées dans un entrepôt (auquel cas c'est leur identifiant qui permet d'établir le lien du data paper vers les données).

Quelques liens :

- https://coop-ist.cirad.fr/gerer-des-donnees/publier-un-data-paper/1-qu-est-ce-qu-un-data-paper Les différents articles et ressources de la Dist au Cirad autour des data papers.
- https://doranum.fr/data-paper-data-journal/ La fiche synthétique et les ressources de DoRANum autour de la question.
- https://data.ird.fr/datapapers/ Définition et ressources de l'Institut de Recherche pour le Développement.

Pour aller plus loin : sélection de ressources et d'outils

Ressources

La bibliothèque du comité pour la science ouverte (« Ouvrir la science ! ») propose une série de ressources pédagogiques à destination des doctorants et chercheurs. Voici deux premières publications pouvant vous aiguiller, le reste des textes (parfois plus officiels, des rapports, etc.) est disponible dans l'onglet « Ressources ».

- https://www.ouvrirlascience.fr/science-ouverte-entrez-dans-le-debat/
- https://www.ouvrirlascience.fr/partager-les-donnees-liees-aux-publications-scientifiques-guide-pour-les-chercheurs/

Pour les chercheurs en archéologie, le consortium labellisé HumaNum IR* et consacré à cette discipline a ouvert plusieurs guides permettant de prendre en main les outils et pratiques de la science ouverte : les sujets des thésaurus et vocabulaires contrôlés, des métadonnées et normes de préservation sont abordés.

- https://masa.hypotheses.org/livre-blanc-guide-des-bonnes-pratiques-en-archeologie
- https://masa.hypotheses.org/openguide

Plus largement, des organismes de financement comme l'ANR ou consortiums d'éditeurs comme Couperin mettent en avant leur politique en matière de science ouverte sur des pages ou sites dédiés. On peut y retrouver les grandes lignes qui fondent leur engagement, des témoignages et retours d'expériences mais aussi des liens vers textes et outils.

- https://anr.fr/fr/lanr/engagements/la-science-ouverte/
- https://scienceouverte.couperin.org/

Déjà mentionnée, la plateforme DoRANum propose des listes assez riches en termes de lexique de la science ouverte ainsi que de bibliographie. Elles peuvent intéresser ponctuellement des chercheurs désireux d'approfondir un point en particulier ou de se familiariser avec les concepts employés.

- https://doranum.fr/glossaire-donnees-recherche/
- https://doranum.fr/bibliographie-webographie-donnees-recherche/

Avant de finir, une série de guides plus généralistes et parfois en anglais montrant la portée institutionnelle et internationale de la science ouverte. À destination des doctorants et des chercheurs comme du grand public, il s'agit de découvrir l'environnement législatif et juridique des données et publications ouvertes.

- https://formadoct.doctorat-bretagneloire.fr/donnees_recherche
- https://guides.data.gouv.fr/ (avec des textes législatifs : https://guides.data.gouv.fr/publier-des-donnees/guide-juridique/chronologie-de-lopen-data)
- https://mantra.ed.ac.uk/
- https://data.europa.eu/elearning/fr/#/id/co-01

Enfin, pour celles et ceux qui seraient désireux de suivre de manière plus approfondie un cours au sujet de la science ouverte, un MOOC en ligne de France Université Numérique propose une formation autour de la question.

- https://www.fun-mooc.fr/fr/cours/la-science-ouverte/

Outils

- https://skosmos.loterre.fr/TSO/fr/ « Thésaurus de la science ouverte » développé au CNRS proposant une liste de définitions avec sources. Organisé de manière multilingue et hiérarchisé, cet outil offre un panorama exhaustif des bonnes pratiques et des acteurs de la science ouverte.
- https://opidor.fr/ Outre l'outil de préparation de plan de gestion de données déjà mentionné, OPIDoR propose également des services en termes de repérage des soutiens à la recherche pour les données scientifiques (Cat) et en attribution d'identifiants pérennes (PID).
- https://entrepot.recherche.data.gouv.fr/datapartage-datapapers-web/ Depuis 2023, l'entrepôt de données interdisciplinaire du MESR propose un outil de génération semi-automatique de Data Paper à partir du DOI attribué aux données déposées.
- https://pactols.frantiq.fr/opentheso/ Développé à la Maison de l'Orient et de la Méditerranée avec le soutien de la TGIR HumaNum, ce thésaurus appliqué à l'archéologie couvre l'ensemble des thématiques du domaine. Des formations sont proposées pour prendre en main les questions de vocabulaires contrôlés, d'alignement et d'interopérabilité.