SCIENCES HUMAINES & SOCIALES

Vade-mecum

Les **données de recherche** en sciences humaines et sociales

Sommaire

Illustration de couverture :
Foule © Hans / Licence Creative Commons / Pixabay

POURQUOI UN *VADE-MECUM* ?

La prospective de CNRS Sciences humaines & sociales, publiée en juin 2025, s'est appuyée sur une réflexion approfondie autour des dispositifs et des outils de la recherche dans un contexte de science ouverte, mettant en lumière plusieurs enjeux centraux liés aux données. L'accès aux données, leur gestion, leur mise à disposition ou encore la création de bases ouvertes figurent parmi les besoins clairement exprimés. Pour proposer des réponses adaptées à ces enjeux et progresser collectivement vers une science fondée sur la transparence des résultats, la traçabilité des matériaux qui les sous-tendent, ainsi que leur gestion raisonnée, il est indispensable que les acteurs concernés partagent une compréhension commune de l'écosystème des données en SHS, et de la politique conduite par l'institut sur ces questions. Pour la mettre en œuvre, CNRS Sciences humaines & sociales s'appuie sur les infrastructures de recherche étoiles (IR*) : Huma-num et Progedo.

HUMA-NUM
Une infrastructure nationale sociotechnique pour les données pluridisciplinaires en sciences humaines et sociales.

PROGEDO
Une infrastructure nationale pour les données et les méthodes en sciences sociales quantitatives.

Pour en savoir plus sur les infrastructures, voir page 15.

C'est dans cette perspective que ce *vade-mecum* pour les données en sciences humaines et sociales (SHS) est adressé en priorité aux directions d'unités et aux responsables de programmes de recherche (Programmes et équipements prioritaires de recherche [PEPR], Programmes prioritaires de recherche [PPR], Suivis ouverts des sociétés et de leurs interactions [SOSI]) et de projets financés relevant de CNRS Sciences humaines & sociales.

Son objectif principal est d'éclairer la politique de l'institut en matière de gestion et de partage des données de la recherche, dans le prolongement de la Feuille de route Science ouverte et du Plan Données du CNRS, et de fournir des repères dans l'écosystème des données en SHS. Il ne constitue ni un guide pratique, ni un document promotionnel.

LES DONNÉES DE RECHERCHE EN SHS

En SHS, les données de recherche peuvent être définies au sens large comme l'ensemble des matériaux qui contribuent à étayer et à valider des résultats scientifiques. La pluralité des disciplines relevant des SHS implique une grande hétérogénéité dans la nature des données produites, collectées ou réutilisées par les équipes de recherche. Celles-ci peuvent inclure des enquêtes et des statistiques, des images, des carnets de terrain, des entretiens audio ou vidéo et leurs transcriptions, ainsi que des corpus (textuels, iconographiques, sonores, audiovisuels), des données spatiales, expérimentales, ou encore des modélisations 3D. À cela s'ajoutent, plus récemment, les données du web. Cette variété implique de s'approprier ce terme à l'aune de chaque champ scientifique, de ses pratiques et de ses objets de recherche. **Les données produites par la recherche publique ont vocation à être ouvertes autant que possible.**

VERS DES DONNÉES FAIR

Pour rappel, les principes FAIR visent à garantir que les données soient : faciles à trouver ; accessibles ; interopérables ; réutilisables. Définis d'abord au niveau européen, les principes FAIR se sont peu à peu imposés comme des standards partagés pour la bonne gestion et diffusion des données de recherche[1]. Ils structurent la politique de données des agences de financement (Agence nationale de la recherche [ANR], union européenne) ; ils sont au cœur du Plan national pour la science ouverte et la grande majorité des établissements et organismes de recherche en font un levier central de leur politique interne de données.

Plusieurs niveaux de « FAIR-isation » sont possibles pour les données de recherche, le respect de ces principes pouvant être plus ou moins poussé selon la nature des données et les contraintes spécifiques qui s'y attachent. **Néanmoins, il est attendu que les données produites dans les laboratoires et dans le cadre de projets de recherche présentent un niveau satisfaisant de compatibilité avec les principes FAIR.**

Les projets et dispositifs présentés ci-dessous illustrent, à titre d'exemple, la diversité des données rencontrées en sciences humaines et sociales et témoignent d'une attention particulière portée à l'alignement de leurs pratiques de recherche sur les principes FAIR.

- **Réseau thématique SILEX : caractérisation et provenance d'une géo-ressource durant la préhistoire**

Silex est une structure nationale française créée en 2019, qui a pour objet la caractérisation et la provenance des silicites — des roches siliceuses (dont le silex) — qui sont très importantes dans les études préhistoriques.

- **SOSI Mobiliscope : la ville à toute heure**

Le Mobiliscope est un outil de géovisualisation qui donne à voir les variations de la population présente dans les territoires au cours des 24 heures d'une journée typique de semaine.

- **Anthropen : les frontières des données anthropologiques**

Dictionnaire situé au cœur des débats contemporains de l'anthropologie et des sciences de la culture, il publie régulièrement des contenus inédits.

- **Consortium 3D**

Ce consortium repose sur un réseau interdisciplinaire comptant plus de 26 institutions dans les domaines des SHS (patrimoine, archéologie, linguistique, ...) de l'informatique graphique, de l'optique.

- **Consortium DISTAM (DIgital STudies Africa, Asia, Middle East)**

DISTAM est un consortium d'Huma-Num destiné à accompagner et à consolider la transition numérique des études moyen-orientales, africaines et asiatiques en France.

- **Cocoon (COllections de COrpus Oraux Numériques)**

CoCoON est une plateforme technique qui accompagne les producteurs de ressources orales, à créer, structurer et archiver leurs corpus.

- **Chronocarto**

Chronocarto est une plate-forme WEBSIG destinée à la valorisation de données géoréférencées dans les domaines de l'archéologie, de l'ethnologie et de l'histoire.

- **Projet du Panel de Caen**

Le Panel de Caen est une enquête qualitative longitudinale, auprès d'un panel de jeunes vivant à Caen en Normandie, sur 20 ans. Elle a étudié les dynamiques de parcours, les processus biographiques, les bifurcations...

1. Pour les projets mobilisant des données autochtones, les principes FAIR peuvent être complétés par les principes CARE (Collective Benefits, Authority to control, Responsibility, Ethics), élaborés par la Global Indigenous Data Alliance. Ces principes mettent l'accent sur l'intendance éthique des données et sur la reconnaissance des droits et de la souveraineté des peuples autochtones en matière de gouvernance des données. Ils illustrent la capacité d'une communauté de recherche à définir et à mettre en œuvre des standards de gouvernance des données adaptés à ses enjeux éthiques et aux caractéristiques des données sur lesquelles elle travaille.

GÉRER LES DONNÉES DE RECHERCHE

LE PGD PROJET

Le plan de gestion de données (PGD), aussi appelé *Data Management Plan* (DMP) en anglais, constitue un outil essentiel pour anticiper et planifier les règles encadrant la gestion des données de recherche. Sa rédaction, rendue obligatoire par des financeurs comme l'ANR ou l'*European Research Council* (ERC) l'union européenne, a contribué à diffuser cette pratique au sein des communautés en SHS.

L'institut considère que la mise en place d'un plan de gestion de données, qui explicite la manière dont les données sont collectées, documentées et analysées afin de valider les résultats de la recherche, est un élément essentiel de la démarche scientifique et un facteur de réussite du projet. Pour cette raison, la rédaction d'un PGD est attendue pour tout projet de recherche dans lequel les données occupent une place centrale, y compris en l'absence d'exigence imposée par un financeur.

La rédaction de ces plans pourra être facilitée par l'utilisation de l'outil DMP OPIDoR, développé par l'Institut de l'information scientifique et technique du CNRS (Inist-CNRS). Celui-ci met à disposition des scientifiques et de leurs partenaires des modèles de plans de gestion de données permettant d'intégrer les recommandations des financeurs (ANR, ERC), des universités et des organismes de recherche. Le CNRS met à disposition, d'une part, des recommandations illustrant sa politique en matière de données et, d'autre part, des mesures préconisées par la déléguée à la protection des données (DPD) du CNRS pour ce qui regarde le traitement des données à caractère personnel.

Les chercheuses et chercheurs en SHS peuvent trouver un appui à la rédaction et à la relecture de leurs plans de gestion de données au sein des Maisons des Sciences sociales & des Humanités (MSH) ou des ateliers de la donnée. Ils peuvent également se référer au répertoire des services opérationnels de soutien à la rédaction des plans de gestion de données (SOS-PGD).

LA POLITIQUE DE DONNÉES À L'ÉCHELLE DU LABORATOIRE

La « FAIR-isation » des données produites au sein des laboratoires représente souvent un travail important pour les équipes de recherche. Une manière de limiter cette charge consiste à créer un cadre favorable à la production de données nativement FAIR. Cela suppose que l'ensemble des données produites soient, par défaut, soumises à un ensemble de règles encadrant leur gestion, leur documentation, leur partage et leur conservation. Définies en amont, ces règles produisent des effets positifs à long terme : elles permettent d'améliorer la connaissance et la qualité des données, d'anticiper les risques liés à leur obsolescence, de systématiser des procédures (notamment dans le cadre de l'application du règlement général sur la protection des données) et de garantir leur intégrité, intelligibilité et accessibilité dans

le temps. Elles facilitent également l'évaluation des coûts liés à leur gestion et les besoins en ressources humaines et financières.

Une première solution concrète peut consister à mettre en place une politique de données à l'échelle du laboratoire (pouvant prendre la forme d'un PGD d'entité — laboratoire, plateforme). Pour les structures volontaires, en particulier celles pour lesquelles la production de données constitue un enjeu central, l'institut encourage la rédaction de ce document de référence. Les laboratoires souhaitant se doter d'une politique de données sont d'ailleurs invités à se rapprocher du pôle science ouverte de l'institut.

LE SOUTIEN OPÉRATIONNEL DES ÉQUIPES DE RECHERCHE

Les équipes de recherche nécessitant un soutien opérationnel dans la gestion et la diffusion des données de recherche en SHS peuvent faire appel à la MSH qui se trouve sur leur site. Pour accéder à l'offre de services des infrastructures nationales Huma-Num et Progedo, mais aussi pour bénéficier d'une expertise dans le champ des données de la recherche tout au long de leur cycle de vie. Pour les quelques cas où une MSH et un atelier de la donnée coexistent sur un même site, la MSH assure la représentation disciplinaire des sciences humaines et sociales au sein de cet atelier.

Les ateliers de la donnée, labellisés par le ministère de l'Enseignement supérieur, de la Recherche et de l'Espace (MESRE) et présents sur l'ensemble du territoire métropolitain, sont des guichets d'accompagnement pouvant être sollicités dans le cadre d'un accompagnement de proximité. Non disciplinaires, ils peuvent assurer un soutien de premier niveau sur la gestion et le partage des données, par exemple, lors de la réalisation de PGD, du choix d'un entrepôt ou encore l'application des principes FAIR.

OUVRIR ET PARTAGER LES DONNÉES DE LA RECHERCHE

OUVRIR PAR DÉFAUT ET RÉUTILISER LIBREMENT

En France, la loi pour une République numérique de 2016 a instauré un principe d'*Open Data*, prévoyant l'ouverture par défaut et la libre réutilisation des données, lorsqu'elles correspondent à des données publiques. Cela signifie que les données de recherche doivent être mises en ligne spontanément et rendues librement réutilisables sans entrave. Cette libre réutilisation est atteignable notamment en apposant sur les données une licence CCY-BY 2.0 ou Etalab[2] avec pour seule restriction le respect de leur intégrité et la mention de leur source. Pour les bases de données, les chercheuses et chercheurs peuvent également faire le choix de la licence ODbL (*Open Database License*). L'utilisation de licences est essentielle pour établir un cadre légal clair, qui protège les droits des producteurs de données et garantit la reconnaissance des auteurs et de leur travail.

Dans certains cas exceptionnels listés par la loi (droits de propriété intellectuelle appartenant à des tiers, secrets administratifs, de données personnelles, protection du patrimoine scientifique et technique de la nation, par exemple), le principe d'ouverture par défaut des données laisse la place à une obligation de protection. En sciences humaines et sociales, ces exceptions sont particulièrement fréquentes. Il est donc essentiel, lorsqu'elles entrent dans ce cadre, de mettre en place des mesures appropriées de protection des données. Pour aider les équipes de recherche à comprendre et à appliquer le règlement général sur la protection des données (RGPD) dans le cadre de la recherche en SHS, l'institut et le service de la protection des données du CNRS ont élaboré un Guide pour la recherche.

CNRS Sciences humaines & sociales dispose également d'un comité d'éthique opérationnel (CEO), chargé d'éclairer les équipes de recherche sur les questions éthiques liées à leurs projets. Un formulaire de demande d'avis ainsi que le calendrier du CEO sont disponibles sur le site de CNRS Sciences humaines & sociales.

CRÉER LES CONDITIONS D'OUVERTURE

Lorsque les données de recherche relèvent de l'un des régimes dérogatoires prévus par la loi[3], notamment en cas de présence de données personnelles ou sensibles, cela ne signifie pas qu'elles doivent pour autant être systématiquement exclues de toute forme de partage. Il est souvent possible d'envisager des conditions d'ouverture encadrées, conciliant partage et respect des obligations juridiques. La science ouverte est compatible avec les logiques de

2. Licence ouverte / Open licence : https://www.etalab.gouv.fr/licence-ouverte-open-licence/
3. Pour en savoir davantage sur les exceptions à l'ouverture et à la libre réutilisation des données, voir Arènes C., Maurel L. et Rennes S. 2022, *Guide d'application de la Loi pour une République numérique pour les données de la recherche*, Comité pour la science ouverte, p. 8. (hal-03968218)

protection des données lorsque des nécessités l'imposent, d'où l'adage « aussi ouvert que possible, aussi fermé que nécessaire ».

Avant toute collecte ou réutilisation de données personnelles ou sensibles, il est nécessaire, en amont du projet, de prendre les mesures techniques et organisationnelles visant la protection des données telles que prévues par le RGPD. Les mesures doivent être validées par la Déléguée à la protection des données du CNRS, si désignée par la direction du laboratoire, ou par son homologue dans un établissement partenaire. Parmi ces mesures, la directrice ou le directeur d'unité, en tant que responsable de traitement, a l'obligation de documenter les traitements de données personnelles et de tenir à jour un registre des traitements.

La présence de données personnelles, particulièrement fréquente en SHS, constitue l'un des principaux obstacles ou arguments d'opposition au partage des données. Or, dans certaines situations, des procédés tels que la pseudonymisation ou l'anonymisation permettent tout de même d'envisager des solutions de partage adaptées. Ces mesures doivent être définies au cas par cas, en fonction des finalités, des usages visés et des risques pour les personnes. Lorsque des traitements garantissant une anonymisation suffisamment robuste sont mis en œuvre et qu'ils sont accompagnés d'une analyse de risque et de méthodes d'anonymisation dûment documentées, les données peuvent, dans certains cas, sortir du champ d'application du RGPD et être réintégrées dans le périmètre de l'*Open Data*. Lorsque la pseudonymisation est privilégiée, il est possible de mettre en œuvre des procédures d'accès, de partage contrôlé à des fins de recherche, à l'image de celles proposées par Progedo.

Ainsi, plutôt que de considérer la présence de données personnelles ou sensibles comme un empêchement définitif à toute forme de diffusion, **il convient d'adopter une approche proportionnée**, fondée sur l'analyse au cas par cas et sur la mise en place de conditions d'ouverture adaptées à la nature des données concernées. Lorsque la mise en œuvre de telles conditions n'est pas envisageable ou réalisable, l'absence de partage devra être dûment documentée et justifiée par des motifs légitimes, afin d'éviter qu'elle ne relève d'une décision purement discrétionnaire. Ces motifs légitimes seront rendus visibles, par exemple dans un plan de gestion de données ou dans des métadonnées associées.

LES DONNÉES LIÉES AUX PUBLICATIONS SCIENTIFIQUES : LE RÔLE DES REVUES EN SHS

Afin d'étayer les résultats présentés dans une publication, les données qui les sous-tendent doivent pouvoir être rendues accessibles, autant que leur nature le permet. L'ouverture, ou le partage des données à l'origine de résultats publiés, est indispensable pour garantir la transparence du processus de recherche et, plus largement, la crédibilité de la recherche en SHS. Pour garantir leur identification et leur accessibilité sur le temps long, il est important qu'elles soient citées dans un format standardisé, déposées en priorité dans un entrepôt « de confiance » dédié plutôt qu'exclusivement dans des rubriques de type *Supplementary Data* ou *Supplementary Materials*. Afin de prévenir toute appropriation par des éditeurs privés, **l'institut déconseille fortement le dépôt des données dans des entrepôts commerciaux**. Si des chercheuses et chercheurs se voient imposer une modalité spécifique pour la diffusion de données liées à l'article par un éditeur privé, il est recommandé de prendre contact avec le pôle science ouverte de l'institut.

Les revues en SHS, en tant qu'actrices majeures du processus de publication, ont un rôle à jouer dans le partage des données associées aux articles qu'elles diffusent. Certaines revues ont déjà initié des actions dans ce sens[4], mais elles restent aujourd'hui minoritaires. Les revues souhaitant développer ce lien publication-données sont invitées à se rapprocher du pôle science ouverte.

LES ENTREPÔTS DE DONNÉES

Les entrepôts de données sont des services en ligne permettant le dépôt, la curation, la description, la conservation, le référencement et la diffusion de jeux de données. Ils se distinguent des dispositifs de stockage utilisés pendant la durée d'un projet, permettant l'exploitation opérationnelle des données à court terme, sans nécessairement garantir leur mise à disposition.

Déposer ses données suppose une réflexion sur la priorisation des entrepôts existants. Pour prévenir les risques d'appropriation des données par des tiers, garantir leur souveraineté, éviter la multiplication des solutions de dépôt et donc l'éparpillement des matériaux de la recherche, **il convient de privilégier les entrepôts de confiance**[5]. Un entrepôt est qualifié d'entrepôt de « confiance » lorsqu'il respecte un certain nombre de critères exigeants, incluant notamment l'attribution d'un identifiant pérenne, la garantie d'une conservation à long terme, ainsi que la modération des dépôts.

L'institut recommande, en premier lieu, le recours systématique aux entrepôts de confiance portés par les infrastructures* Huma-Num et Progedo. Selon les besoins et les thématiques, il est possible de recourir à des entrepôts disciplinaires (les Collections de corpus oraux numériques [CoCoON], le Conservatoire national des données 3D [CND3D], la Banque de données du centre de données socio-politiques (CDSP), par exemple) ou bien, en ultime solution, d'envisager de se tourner vers l'entrepôt pluridisciplinaire porté par Recherche Data Gouv.

Au moment du partage, les entrepôts de données endossent le rôle de diffuseurs. Le déposant qui procède à la mise en ligne de données à partir de ces services garantit qu'elles soient finalisées, suffisamment documentées et légalement diffusables.

UN DÉLAI RAISONNABLE DE MISE À DISPOSITION

Lorsque les jeux de données ou bases de données relèvent de la responsabilité d'un organisme de recherche, ce dernier exerce une responsabilité financière et/ou administrative, notamment sur leur usage et leur réutilisation. À moins qu'elles ne soient soumises à un régime de protection spécifique ou à des restrictions légales de diffusion, ces dernières ont vocation à être ouvertes.

4. À titre d'exemples, CyberGéo propose une rubrique « *Data Paper* » et *Gallia — Archéologie des Gaules* affiche les jeux de données liés aux articles préalablement déposés dans l'entrepôt Nakala.
5. Collège Données de la recherche, 2024, *Sélectionner un entrepôt thématique de confiance pour le dépôt de données : méthodologie et analyse de l'offre existante*, site Ouvrir la Science. https://www.ouvrirlascience.fr/selectionner-un-entrepot-thematique-de-confiance-pour-la-diffusion-des-donnees-de-recherche-note-methodologique/

Conformément à l'article L.311-2 du Code des relations entre le public et l'administration (CRPA), cette ouverture s'inscrit dans la notion de « document achevé ». Appliquée aux données de la recherche, cette notion appelle **une appréciation contextualisée**, du fait du caractère souvent évolutif et cumulatif des jeux et bases de données, ainsi que de la diversité des standards disciplinaires.

Lorsque des données sont mobilisées dans le cadre d'une publication scientifique évaluée par les pairs[6], celles-ci peuvent être considérées comme achevées, ou comme ayant atteint un degré suffisant d'achèvement. Dans ce cadre, **l'institut considère que la publication scientifique constitue un moment pertinent pour ouvrir lesdites données**.

Concernant le caractère achevé des bases de données, lorsque seule une partie non substantielle de la base a été utilisée pour obtenir des résultats présentés dans une publication, il est possible de mettre à disposition uniquement la section exploitée.

L'institut privilégie une approche fondée sur un délai raisonnable de mise à disposition des données, plutôt que sur des délais d'embargo fixes. Cette logique vise à rechercher un équilibre entre, d'une part, une période d'exploitation scientifique prioritaire, notamment le temps nécessaire pour vérifier et documenter les données, et, d'autre part, la garantie que les données qui sous-tendent les résultats présentés dans la publication puissent être accessibles au reste de la communauté scientifique, et au-delà.

Si, en raison de la nature des données, leur mise à disposition est impossible, un *Data Availability Statement* (déclaration de disponibilité des données) constitue une information minimale sur les données qui sous-tendent la publication. Cette déclaration permet d'indiquer aux lecteurs où se trouvent les données, comment y accéder et quelles sont les coordonnées ou les procédures pour en faire la demande.

6. La publication scientifique évaluée par les pairs peut inclure les articles publiés dans des revues scientifiques ainsi que les monographiques s'appuyant sur des données.

LA VALEUR COLLECTIVE DES DONNÉES

VERS UN ÉLARGISSEMENT DES PRATIQUES DE PARTAGE

Outre l'ouverture des données dans le cadre d'une publication scientifique, d'autres situations se prêtent également à leur partage.

Les données expérimentales en SHS constituent tout un pan de la recherche où le partage est encore trop peu courant. Mettre en place des modalités de partage adaptées à ce type de données permet de tester leur robustesse, de permettre leur réutilisation dans d'autres contextes et plus largement de tendre vers une recherche cumulative.

Dans ce contexte, il importe également de considérer le cas des chercheuses et chercheurs qui, au cours de leur carrière, n'auraient que peu, ou pas du tout, partagé les donnés qu'ils ont menées à bien. Il est primordial de rendre visible et d'assurer la transmission de cette production scientifique. Ces données constituent non seulement l'héritage scientifique d'une carrière, d'un parcours de recherche singulier, mais aussi un patrimoine collectif de la recherche en ce qu'elles témoignent des objets, des méthodes, des outils mobilisés dans un contexte scientifique et épistémologique donné. Cette démarche suppose un travail important impliquant une sélection des données, dans certains cas la numérisation des matériaux de recherche, l'organisation des différents produits de la recherche ainsi que leur documentation avant un versement dans un entrepôt de confiance. Les chercheuses et chercheurs rattachés à CNRS Sciences humaines & sociales souhaitant entamer cette démarche sont encouragés à prendre contact avec le pôle science ouverte de l'institut.

BASES DE DONNÉES OUVERTES

Les bases de données constituent un moyen privilégié pour organiser, documenter, interroger et exposer les données de recherche. Qu'elles soient bibliographiques, archivistiques ou produites dans le cadre d'un projet spécifique, leur conception devrait intégrer en amont une réflexion sur leur pérennité et sur les conditions de leur maintien dans le temps. Lorsque cela est possible, il est là aussi recommandé de s'appuyer sur les services offerts par les infrastructures de recherche, notamment en matière d'hébergement. Pour les bases conçues à l'issue d'un projet, les données doivent d'abord être déposées et décrites dans un entrepôt de confiance avant, le cas échéant, d'être exposées sur un site web tiers.

ASSURER LA QUALITÉ ET LA TRAÇABILITÉ

QUALITÉ, DOCUMENTATION ET STANDARDS DE MÉTADONNÉES

La qualité de la recherche fondamentale en SHS est étroitement liée à celle des données collectées, réutilisées, analysées et exploitées, sur lesquelles elle fonde ses résultats (voir Vers des données FAIR). Lorsque les données ne sont pas suffisamment décrites, c'est-à-dire accompagnées d'informations sur leur contexte de production, leurs méthodes de collecte et d'analyse, leur traçabilité est fragilisée, avec un impact direct sur la fiabilité des résultats. À l'inverse, une documentation suffisamment détaillée facilite leur compréhension, leur utilisation et leur réutilisation, d'abord au sein de la communauté scientifique d'origine, puis dans d'autres disciplines et par des acteurs de la société civile (associations, entreprises, etc.). Dans ce cadre, la mise à disposition des données doit s'accompagner d'un travail garantissant, par défaut, un niveau de documentation aussi exhaustif que nécessaire.

Une attention particulière devra également être portée au choix des standards de métadonnées (tels que Dublin Core, *Data Documentation Initiative* [DDI], *Text Encoding Initiative* [TEI]), **y compris lorsque les données elles-mêmes ne peuvent être rendues accessibles**. Les métadonnées constituent alors le principal vecteur permettant d'identifier le contexte de production, la nature des données, leurs objectifs ainsi que les conditions encadrant leur réutilisation.

LE RÔLE DES IR*

Face à cet enjeu de qualité documentaire, les deux IR* apportent, chacune dans leur champ respectif et selon leurs méthodes, des réponses concrètes.

L'IR* Huma-Num propose, *via* Nakala, un système de modération des données[7]. Une fois le dépôt effectué, il est possible de solliciter un modérateur pour évaluer les données sur la base de critères documentaires. Un accompagnement est alors proposé pour améliorer la complétude et la qualité des métadonnées associées. Les chercheuses et chercheurs déposant dans Nakala sont fortement encouragés à demander le statut « modéré » pour les jeux de données déjà déposés ou les futurs dépôts. L'IR* Progedo assure ce travail de documentation en amont du dépôt dans l'entrepôt Quetelet-Progedo. Avant d'être mises à disposition, les données sont finement documentées, jusqu'au niveau des variables et structurées au standard international DDI.

7. Sans auteur, 2024, « Évaluer la qualité documentaire des données : le projet de modération dans Nakala », *Le blog d'Huma-Num et des Consortiums-HN*. https://doi.org/10.58079/11v7o

VERS UNE RECHERCHE REPRODUCTIBLE ?

Le développement de la recherche reproductible en sciences humaines et sociales constitue un axe stratégique identifié par l'institut dans sa prospective, dans un contexte où les enjeux de reproductibilité diffèrent selon les disciplines. La recherche reproductible constitue un gage de transparence et de fiabilité des méthodes scientifiques et des résultats, si bien que dans certains champs de recherche, des revues l'ont établie comme une exigence pouvant conditionner la publication des articles.

L'application des principes FAIR dans les pratiques de recherche constitue une première étape, en ce qu'elle permet notamment d'organiser l'accès aux données, aux algorithmes, aux codes et aux chaînes d'opérations. Pour garantir la reproductibilité, elle gagne toutefois à être complétée par une documentation approfondie des données, des codes et des chaînes de traitement, inscrite dans un environnement computationnel maîtrisé et, autant que possible, standardisé.

À l'échelle nationale, des initiatives et dispositifs existent pour accompagner cette dynamique. Le Réseau français de la recherche reproductible sensibilise, forme et fédère des communautés pluridisciplinaires autour de ces enjeux. Dans le domaine des sciences sociales quantitatives, l'unité *Certification agency for scientific code and data* (CASCAD), soutenue par l'institut, agit comme tiers de confiance pour certifier la reproductibilité des résultats associés aux publications.

INFRASTRUCTURES ET DISPOSITIFS STRUCTURANTS

CNRS Sciences humaines & sociales pilote les infrastructures nationales de recherche étoiles Huma-Num et Progedo. Ces dernières jouent un rôle central et structurant dans la gestion et la diffusion des données de recherche en SHS à l'échelle nationale, notamment en lien avec les vingt-et-unes MSH. Toutes deux sont par ailleurs labellisées Centres de référence thématiques[8] par la plateforme nationale des données, Recherche Data Gouv. Leur rôle est également renforcé par leur implication au sein des grands programmes PPR, PEPR ainsi que l'Appel à manifestation d'intérêt (AMI) SHS[9].

LES IR*

L'IR* HUMA-NUM

Huma-Num est l'infrastructure sociotechnique de référence pour les données pluridisciplinaires en sciences humaines et sociales, fondée sur trois piliers complémentaires : les communautés, une offre de services numériques et une infrastructure informatique souveraine permettant l'accès à des ressources pour le traitement, le stockage et le calcul. Elle porte la participation de la France dans l'infrastructure européenne *Digital Research Infrastructure for the Arts and Humanities* (DARIAH). La mission principale d'Huma-Num consiste, d'une part, à soutenir et fédérer des communautés de recherche réunies sous la forme de consortiums labellisés. D'autre part, en co-construction avec ces communautés, et sous un pilotage scientifique, l'infrastructure propose une grille de services et d'outils pour les données de recherche en SHS, comprenant notamment des solutions de stockage sécurisé, de la puissance de calcul et des outils pour traitement, une solution de dépôt, d'exposition et de référencement *via* l'entrepôt Nakala (entrepôt thématique de confiance de Recherche Data Gouv), un service d'archivage et la plateforme Isidore dédiée à la découvrabilité des données et publications scientifiques. Huma-Num n'entend pas répondre à l'ensemble des besoins liés aux données, mais à ceux clairement identifiés par les communautés elles-mêmes et intégrés dans une démarche scientifique et éthique. Afin d'assurer la pérennité, la souveraineté et la maîtrise des outils et des données par les communautés, Huma-Num s'appuie sur une infrastructure informatique fiable et sécurisée au sein du centre de calcul de CNRS Nucléaire & Particules.

L'IR* PROGEDO

Progedo est l'infrastructure de référence pour les sciences sociales quantitatives, reposant sur des données de qualité. L'IR* coordonne la participation française à la production des grandes enquêtes européennes, telles que l'*European Social Survey* (ESS) et la *Survey of Health, Ageing and Retirement in Europe* (SHARE). Par son implication dans ces dispositifs de grande ampleur, Progedo fédère des compétences en méthodes statistiques et computationnelles. L'infrastructure assure également la mise à disposition sécurisée de ces données, ainsi que de celles issues de la statistique publique, selon les standards internationaux. Grâce à des conventions signées avec les producteurs de données, Progedo organise l'accès à des jeux de données finement documentés à des fins de recherche *via* le catalogue Quetelet-Progedo. Les chercheuses et chercheurs ainsi que les institutions peuvent également effectuer des dépôts de données dans l'entrepôt Quetelet-Progedo (entrepôt thématique de confiance de Recherche Data Gouv).

8. Les centres de référence thématiques ont une portée nationale et disciplinaire. Ils soutiennent l'action en matière de gestion et de diffusion des données d'un champ scientifique. https://recherche.data.gouv.fr/fr/page/centres-de-reference-thematiques-expertises-par-domaine-scientifique

9. La participation de CNRS Sciences humaines & sociales dans l'AMI SHS se fait notamment sur le champ « circuit de la donnée » : https://www.cnrs.fr/fr/actualite/nous-esperons-accompagner-lemergence-de-clusters-shs-sur-des-themes-communs

La prospective de l'institut conforte leur importance dans le champ des données et des humanités numériques. **Il est donc indispensable qu'elles soient clairement identifiées par les communautés, afin que celles-ci puissent pleinement bénéficier de leurs services et de leurs expertises aux différentes étapes du cycle de vie des données.**

L'IR RNMSH ET LES MAISONS DES SCIENCES SOCIALES & DES HUMANITÉS

Les vingt-et-unes MSH forment une infrastructure de recherche nationale distribuée. Elles offrent des services essentiels aux communautés de recherche en SHS présentes sur chaque site. Le Réseau national des maisons des sciences sociales & des humanités (RnMSH) en assure la coordination et l'animation à l'échelle nationale.

Les MSH sont les points d'entrée, pour l'ensemble des laboratoires SHS de chaque site, des services d'Huma-Num et de Progedo *via* les pôles humanités numériques et les plateformes universitaires de données (PUD), grâce aux correspondants Huma-Num et Progedo. Ces services représentent la brique SHS des ateliers de la donnée.

Par ailleurs, de nombreuses MSH ont développé des plateformes technologiques regroupant des instruments de pointe, portées par des ingénieures et ingénieurs, et associant des chercheuses et chercheurs issus de différents laboratoires. Ces plateformes s'articulent autour d'objets et de thématiques prioritaires pour les tutelles. Au niveau du RnMSH, elles sont classées selon une typologie des données produites : Cogito (données cognitives), Audio-Visio (données audiovisuelles), Spatio (données spatiales), Scripto (données textuelles) et Data (données statistiques et quantitatives).

Ces plateformes, labellisées et répertoriées dans une base de données du RnMSH, entretiennent un lien direct avec les correspondants Huma-Num et Progedo.

COLLECTER SUR LE TEMPS LONG : LE DISPOSITIF SOSI

Face au constat selon lequel les dispositifs de soutien à la recherche existants peinent à répondre à l'un des besoins fondamentaux de la recherche en SHS, à savoir : « le temps, plus précisément la durée et la sécurisation des conditions pour qu'une recherche de longue haleine puisse se déployer »[10], CNRS Sciences humaines & sociales a mis en place en 2021 le dispositif SOSI (Suivi ouvert des sociétés et de leurs interactions). L'institut soutient aujourd'hui 13 SOSI, structurés sous la forme d'observatoires pour les sciences humaines et sociales, permettant notamment la collecte de données sur le temps long. Dans ces dispositifs, une attention particulière est portée à la documentation des méthodes de collecte ainsi qu'à la mise à disposition des données, afin d'en faciliter la réutilisation dans le monde académique et extra-académique.

10. Fabrice Boudjaaba, « Soutenir des recherches au long cours : les Suivis ouverts des sociétés et de leurs interactions (SOSI) », *CNRS Sciences humaines & sociales — La Lettre*, octobre 2024, p. 31. https://www.inshs.cnrs.fr/sites/institut_inshs/files/download-file/lettre_infoINSHS_90.pdf

ACCÉDER AUX DONNÉES DANS LE CADRE DE LA RECHERCHE EN SHS

La recherche en sciences humaines et sociales s'appuie largement sur des données existantes produites par des tiers (par exemple, des administrations ou des entreprises). Lorsqu'elles sont sensibles, personnelles ou confidentielles, ces données sont soumises à des régimes d'accès spécifiques. L'accès aux données constitue un enjeu majeur pour le développement de la recherche dans de nombreux champs scientifiques. Pour y répondre, CNRS Sciences humaines & sociales est impliqué dans divers dispositifs facilitant l'accès à ces données.

LE RÔLE DE PROGEDO DANS L'ACCÈS AUX DONNÉES PSEUDONYMISÉES

L'IR* Progedo coordonne les accès à des données issues d'administrations publiques et de services statistiques ministériels. Dans ce cadre, elle permet notamment la mise à disposition de données pseudonymisées à des fins de recherche[11]. Ces données font l'objet d'un important travail de documentation afin de permettre leur réemploi.

Progedo assure également l'accès à des catalogues de données internationaux tels que le *Consortium of European Social Science Data (CESSDA) data catalogue*, point d'entrée pour l'ensemble des fournisseurs nationaux de données au niveau européen.

L'ACCÈS AUX DONNÉES DE SANTÉ : LE SNDS

La recherche en SHS est pleinement concernée par la thématique de la santé. CNRS Sciences humaines & sociales collabore dans ce cadre avec d'autres établissements du Campus Condorcet au sein de la plateforme SHS santé.

La réalisation de projets de recherche dans ce champ nécessite un accès facilité aux données de santé. Depuis 2021, le CNRS bénéficie d'un accès permanent au système national des données de santé (SNDS)[12] dont l'usage est étendu à tous les membres des unités CNRS. Le SNDS regroupe principalement les données issues des remboursements de l'Assurance maladie depuis 2006, mais aussi des informations hospitalières. L'accès permanent du CNRS au SNDS permet l'ouverture de comptes sans demande préalable auprès de la CNIL, simplifiant ainsi les démarches pour les chercheuses et chercheurs habilités. L'institut est impliqué dans l'arbitrage de ces accès.

Une adresse unique a été mise en place pour centraliser spécifiquement ces demandes : CNRS-SNDS@cnrs.fr

11. Pseudonymisées, les données versées et les données réutilisées nécessitent des mesures de protection telles que prévues par le RGPD.
12. Centre national de la recherche scientifique (CNRS), 2023, « Les données de santé ouvrent des perspectives de recherche ». https://www.cnrs.fr/fr/actualite/les-donnees-de-sante-ouvrent-des-perspectives-de-recherche

L'ACCÈS AUX DONNÉES CONFIDENTIELLES : LE CASD

Le Centre d'Accès Sécurisé aux Données (CASD) a pour mission principale d'organiser et de mettre en œuvre des services d'accès sécurisés aux données confidentielles, issues de nombreux producteurs (l'Institut national de la statistique et des études économiques [INSEE], la Direction de l'animation de la recherche, des études et des statistiques [DARES], Santé publique France, France Travail, etc.). À ce jour, plus de 500 sources de données sensibles sont accessibles *via* le CASD[13]. L'accès aux données se fait à distance grâce à un dispositif technique sécurisé : la SD-Box, un boîtier informatique qui permet de se connecter à l'environnement où sont hébergées les données confidentielles[14]. Certaines PUD hébergées au sein des MSH sont équipées de cette SD-Box, facilitant ainsi l'accès aux données pour les chercheuses et chercheurs. Les laboratoires ne pouvant financer ce dispositif en interne peuvent se rapprocher de leur PUD de secteur.

13. Réseau national des Maisons des Sciences de l'Homme (RnMSH), État des lieux du réseau des plateformes Data, janvier 2025, p. 37. https://www.msh-reseau.fr/media/pages/plateformes/b824040ddd-1739888219/rapport-plateformes_data_vf.pdf
14. Le traitement de ces données doit être encadré en lien avec les déléguées et délégués à la protection des données.

OUTILLER ET FORMER DURABLEMENT LES COMMUNAUTÉS

JEUNES CHERCHEUSES ET CHERCHEURS : SENSIBLISER ET FORMER

Le déploiement d'une politique des données repose également sur la sensibilisation et sur la formation des jeunes chercheuses et chercheurs et doctorantes et doctorants. Assurer cette formation le plus tôt possible permet d'instaurer un cadre favorable à l'adoption de réflexes et de pratiques de recherche compatibles avec les exigences liées à la gestion et au partage des données. Il existe un certain nombre de ressources à destination de ce public comme le passeport pour la science ouverte du Comité pour la science ouverte ou la plateforme d'autoformation DoRANum. Ils peuvent être notamment partagés auprès des nouveaux arrivants au sein des laboratoires, des programmes de recherche et des réseaux.

IDENTIFIER LES BESOINS EN RESSOURCES HUMAINES ET EN COMPÉTENCES

Au sein des laboratoires ou des projets de recherche, il est essentiel d'identifier en amont les besoins en matière de ressources humaines et de compétences relatives aux données. Les fiches de poste relevant des branches d'activité professionnelles (BAP) D, E ou F, lorsqu'elles concernent les données de recherche, devraient inclure, autant que faire se peut, du temps consacré à l'ingénierie de données dans une perspective de science ouverte, à la question de leur mise à disposition et de leur réemploi, au-delà du temps du projet.

ANNEXE – BIBLIOGRAPHIE, SITOGRAPHIE

Retrouvez l'intégralité de la sitographie *via* le QR Code ci-dessous :

REMERCIEMENTS

Ce *vade-mecum* a été rédigé par Mathilde BERNIER pour le pôle science ouverte de CNRS Sciences humaines & sociales.

Avec la collaboration de Lionel MAUREL, successivement directeur adjoint scientifique science ouverte, édition scientifique et données de recherche et des Maisons des sciences sociales et des humanités, puis chargé de mission « données de la recherche » à l'institut.

Avec nos remerciements pour leur relecture et leurs apports :

La direction de CNRS Sciences humaines & sociales — Fabrice BOUDJAABA, Caroline BODOLEC.
Les directeurs scientifiques adjoints de CNRS Sciences humaines & sociales — Sylvia NIETO-PELLETIER, Pascale GOETSCHEL, Patricia CABREDO HOFHERR, Ricardo ETXEPARE, Cédric PATERNOTTE, Sandrine MALJEAN-DUBOIS, Franck LECOCQ, Nicolas ADELL, Anne-Cécile HOYEZ, Emmanuel HENRY, William BERTHOMIERE, Brice TROUILLET, Nicolas THELY.
Les directeurs des IR* — Olivier BAUDE, Nicolas SAUGER.
La déléguée à la protection des données du CNRS — Gaëlle BUJAN.
Le directoire du RnMSH — Christophe CHARLIER, Gilles POLET, Emmanuelle POULAIN-GAUTRET, Myriam DANON-SZMYDT.
Thomas LEBARBE
Thomas THEVENIN
Elsa SUPIOT
La déléguée à la protection des données du CNRS — Gaëlle BUJAN.
La responsable du pôle science ouverte, édition scientifique et données de recherche à l'institut — Astrid ASCHEHOUG.
La chargée de communication de l'institut — Zoë CHERON.

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'ESPACE

Liberté
Égalité
Fraternité

SCIENCES HUMAINES
& SOCIALES

3, rue Michel-Ange
75794 Paris Cedex 16
www.inshs.fr