Petite encyclopédie de la science ouverte

RÉSUMÉ

Données de recherche ouvertes

La non-ouverture des données scientifiques pèse sur les budgets. On estime qu'une part importante des connaissances scientifiques **disparaît** chaque année. Selon une étude réalisée en 2014, **moins de 50 % des jeux de données biologiques datant des années 1990 ont été récupérés**, au prix d'un investissement élevé en temps et en efforts. À titre de comparaison, 98 % des jeux de données publiés dans PLOS avec des identifiants uniques (DOI) sont toujours disponibles pour les futures recherches. Les données scientifiques ouvertes sont des ressources fondamentales pour un grand nombre d'activités de recherche, comme la méta-analyse, la reproduction des résultats ou l'accessibilité des sources primaires. De plus, leur valeur économique et sociale est importante, car les données scientifiques sont fréquemment exploitées par des professionnels non universitaires, des agences publiques et des organisations à but non lucratif.

Pourtant, les données scientifiques ouvertes impliquent certains coûts. Rendre les données à la fois téléchargeables et exploitables nécessite d'importants investissements en matière de documentation, de nettoyage des données, de licences et d'indexation. Il n'est pas possible de partager toutes les données scientifiques, et des contrôles sont souvent nécessaires pour vérifier l'absence d'informations personnelles ou de contenu soumis à droit d'auteur.

Pour être efficace, le partage des données doit être anticipé tout au long du cycle de vie de la recherche. Les nouveaux principes de gestion des données scientifiques visent à formaliser les cultures de données observées dans les communautés scientifiques et à appliquer des normes communes. Initialement publiés en 2016, les principes FAIR (Facilité à trouver, Accessibilité, Interopérabilité et Réutilisabilité) constituent un cadre de référence pour l'ouverture des données scientifiques.

Les politiques en faveur du partage des données sont passées d'un discours général d'encouragement au développement concret de services dédiés. Les premières initiatives remontent aux infrastructures informatiques pionnières : en 1957, le système WDC (World Data Center) visait à rendre facilement accessibles

un large panel de données scientifiques. Les programmes de données ouvertes étaient pourtant sévèrement limités par les lacunes techniques et les incompatibilités lors des transferts. Après 1991, le Web a fourni un cadre universel pour l'échange de données et entraîné une expansion massive des bases de données scientifiques. Et pourtant, bien des projets ont rencontré des problèmes critiques de durabilité à long terme.

Les infrastructures de science ouverte sont récemment devenues des vecteurs clés de la diffusion et de la gestion des données scientifiques ouvertes. Les référentiels assurent la conservation et la découvrabilité des ressources scientifiques. Les données qu'ils hébergent sont plus fréquemment utilisées et citées que celles publiées dans les fichiers supplémentaires.

Cet article est publié sur ce site web et simultanément sous la forme d'un article wikipédia mis à jour de manière indépendante.

Données de recherche ouvertes

Langlais, Pierre-Carl publié le 1 juin 2024

ARTICLE

Les **données scientifiques ouvertes** ou **données de recherche ouvertes** s'inscrivent dans la publication d'observations et de résultats d'activités scientifiques analysables et réutilisables par tous. L'un des principaux objectifs du mouvement des données ouvertes consiste à faciliter la vérification des théories scientifiques, en permettant l'examen de la reproductibilité des résultats par des tiers [1] et l'intégration de données provenant de nombreuses sources pour transmettre de nouvelles connaissances. [2]

Le concept moderne de données scientifiques est apparu dans la seconde moitié du XXe siècle, avec le développement de grandes infrastructures de la connaissance pour traiter les observations et les informations produites par la science. Rapidement, le partage et la distribution des données ont été identifiés comme des enjeux importants, mais ils étaient entravés par les limites techniques des infrastructures et l'absence de normes communes pour la communication des données. Dès l'origine, le World Wide Web fut conçu comme un protocole universel destiné au partage des données scientifiques, notamment dans le domaine de la physique des particules.

Définition

Données scientifiques

Le concept de données scientifiques ouvertes s'est développé en parallèle de celui de données scientifiques.

Les données scientifiques n'ont été formellement définies qu'à la fin du XXe siècle. Avant la généralisation de l'analyse informatique, les données étaient surtout une notion informelle assez souvent interchangeable avec celles de connaissance et d'information. [3] Les discours institutionnels et épistémologiques ont favorisé d'autres concepts et points de vue concernant les activités scientifiques : « Même les histoires des sciences et les commentaires épistémologiques ne mentionnent les données qu'en passant. D'autres ouvrages fondamentaux sur la création de sens dans le domaine de la science traitent des faits, des représentations, des inscriptions et des publications, avec peu d'intérêt pour les données en tant que telles. ». [4]

La première définition politique des données scientifiques à faire autorité remonte à 1999, quand les Académies nationales des sciences les ont présentées comme « des faits, des lettres, des nombres ou des symboles qui décrivent un objet, une condition, une situation ou d'autres facteurs ». [5] La terminologie a continué à évoluer : en 2011, les Académies nationales ont actualisé la définition pour inclure une grande variété d'objets datafiés tels que « les données de spectrographie, de séquençage génomique, de microscopie électronique et observationnelles, telles que les données de télédétection, géospatiales et socio-économiques, ainsi que d'autres formes de données générées ou compilées par des humains ou des machines », auxquelles s'ajoute « la représentation numérique de la littérature ». [6]

Alors que les formes de données restent vastes et incertaines, dernièrement les définitions et politiques standard avaient tendance à

circonscrire les données scientifiques aux secteurs informatiques et numériques. 7 Le projet pilote d'Horizon 2020 sur le libre accès aux données s'est volontairement limité à la recherche numérique : « Les « données de recherche numériques » sont constituées d'informations sous forme numérique (en particulier des faits ou des chiffres), collectées pour être examinées et utilisées comme base de raisonnement, de discussion ou de calcul. Elles incluent les statistiques, les résultats d'expériences et d'enquêtes, les mesures, les observations de terrain, les enregistrements d'entretiens et les images. ». 8

Dans l'ensemble, le statut des données scientifiques reste un point de discussion irrésolu entre les chercheurs, les communautés et les décideurs politiques : « Plus globalement, toutes les « données » qui présentent un intérêt pour les chercheurs devraient être traitées comme des « données de recherche ». ». 9 Des rapports politiques majeurs, comme la synthèse collective des Académies nationales des sciences sur la citation des données publiée en 2012, ont délibérément choisi une définition relative et nominaliste des données : « Nous consacrerons peu de temps aux questions de définition (par exemple, que sont les données ?), si ce n'est pour reconnaître que les données existent souvent dans les yeux de celui qui les regarde. ». 10 Pour Christine Borgman, la question principale n'est pas de définir les données scientifiques (« ce que sont les données ») mais de contextualiser le moment où les données sont devenues un point central de discussion au sein d'une discipline, d'une institution ou d'un programme de recherche national (« quand ce sont des données »). 11 Dans les années 2010, le développement des sources de données disponibles et la sophistication des méthodes d'analyse des données élargirent l'éventail des disciplines majoritairement concernées par la gestion des données aux « sciences sociales computationnelles, aux humanités numériques, aux données des médias sociaux, aux projets

de recherche en science citoyenne et aux sciences politiques ». [12]

Données scientifiques ouvertes

Les débats sur la gestion des données scientifiques ont consacré l'ouverture et le partage comme des thèmes majeurs, mais aussi fait des données une question pertinente au sein des institutions, des disciplines et des cadres politiques.

Selon Paul Edwards, le cheminement des données faut-il les partager, jusqu'à quel point et avec qui ? fut une cause majeure de la friction des données qui révéla les infrastructures jusqu'alors cachées de la science : « La métaphore de Paul Edwards sur la friction des données décrit les événements observés aux interfaces entre les « surfaces » de données : les points où les données circulent entre les personnes, les substrats, les organisations ou les machines (…) Chaque mouvement de données au niveau d'une interface possède un coût en temps, en énergie et en attention humaine. Chaque interface entre les groupes et les organisations ou entre les machines représente un point de résistance et un risque de brouillage, de mauvaise interprétation ou de perte de données. Dans les systèmes sociaux, la friction des données consomme de l'énergie et produit des turbulences et de la chaleur, c'est-à-dire des conflits, des désaccords et des traitements imprécis et désordonnés. ». [13] L'ouverture des données scientifiques est à la fois une friction des données en soi et un moyen de gérer collectivement ces frictions en aplanissant les questions complexes de propriété. Les cultures scientifiques ou épistémiques ont été reconnues comme des facteurs essentiels dans l'adoption de politiques de données ouvertes : « On pourrait s'attendre à ce que les pratiques de partage des données soient liées aux communautés et amplement déterminées par la culture épistémique. ». [14]

Dans les années 2010, les chercheurs et les décideurs politiques ont adopté de nouveaux concepts pour définir plus précisément les

données scientifiques ouvertes. Depuis leur introduction en 2016, les données FAIR sont devenues un axe majeur des politiques de recherche ouverte. L'acronyme FAIR décrit un type idéal de données Facilement trouvables, Accessibles, Interopérables et Réutilisables (en anglais « Findable, Accessible, Interoperable et Reusable »). Les données scientifiques ouvertes ont été classées en tant que communs ou biens publics dont l'administration, l'enrichissement et la conservation sont en premier lieu des actions collectives plutôt qu'individuelles : « Ce qui rend l'action collective utile pour comprendre le partage des données scientifiques, c'est qu'elle se concentre sur le rôle de l'ajustement des coûts et des bénéfices résultant des contributions à une ressource commune dans l'appropriation des gains individuels. ». 15

Historique

Développement des infrastructures de la connaissance (1945-1960)

Photographie du stockage des cartes perforées au US National Weather Records Center à Asheville (début des années 1960). Selon Paul Edwards, il y avait tellement de cartes perforées à l'époque qu'elles étaient stockées jusqu'à l'entrée de la pièce
Domaine public américain
Image reproduite dans Paul Edwards, A Vast Machine : Computer Models, Climate Data, and the Politics of Global Warming, p. 102

L'émergence des données scientifiques est associée à un glissement sémantique dans la perception courante des concepts clés que sont les données, l'information et la connaissance. [16] Suite au développement des technologies informatiques, les données et les informations sont de plus en plus souvent assimilées à des « choses » : [17] « Comme le calcul, les données ont toujours un aspect matériel. Les données sont des choses. Elles ne sont pas uniquement des chiffres, mais aussi des nombres, avec une dimension, un poids et une texture. ». [18]

Après la Seconde Guerre mondiale, les grands projets scientifiques se sont progressivement appuyés sur une infrastructure de la connaissance pour collecter, traiter et analyser d'importantes quantités de données. Le système à cartes perforées fut initialement employé à titre expérimental pour traiter les données climatiques dans les années 1920, avant une application à grande échelle au cours de la décennie suivante : « Durant l'un des premiers projets de création d'emplois du gouvernement [des États-Unis] pendant la Grande Dépression, les travailleurs de la Civil Works Administration poinçonnèrent quelque 2 millions d'observations de livres de bord consignées entre 1880 et 1933. ». [19] En 1960, les collections de données météorologiques du NWRC (National Weather Records Center) des États-Unis comptaient 400 millions de cartes et avaient une portée mondiale. Le caractère physique des données scientifiques était alors pleinement évident, au point de menacer la stabilité de bâtiments tout entiers : « En 1966, les cartes occupaient tellement d'espace que [le NWRC] commença à installer des armoires de stockage dans son hall d'entrée (voir

illustration). Les autorités craignaient sérieusement que le bâtiment ne s'effondre sous leur poids. ». [20]

À la fin des années 1960, beaucoup de disciplines et de communautés se mirent à adopter les infrastructures de la connaissance. L'ERIC (Educational Resources Information Center) innova en 1966 avec la création du premier référentiel bibliographique des données en libre accès sous forme électronique. La même année vit la création de MEDLINE, une base de données de citations bibliographiques provenant de revues biomédicales, accessible en ligne gratuitement et coadministrée par la National Library of Medicine et le National Institute of Health (États-Unis). Aujourd'hui associée à l'interface PubMed, MEDLINE compte actuellement plus de 14 millions d'articles complets. [21] Des infrastructures de la connaissance ont également vu le jour dans les domaines de l'ingénierie spatiale (NASA/RECON), de la recherche au sein des bibliothèques (OCLC Worldcat) ou des sciences sociales : « Durant les années 1960 et 1970, plus d'une douzaine de services et d'associations professionnelles sont apparus pour coordonner la collecte de données quantitatives. ». [22]

Ouverture et partage des données : premières tentatives (1960-1990)

Des discours et cadres politiques relatifs aux données scientifiques ouvertes ont émergé aussitôt après la création de la première grande infrastructure de la connaissance. Le système WDC (World Data Center, aujourd'hui World Data System) avait pour but de faciliter l'accès aux données d'observation en vue de l'Année géophysique internationale de 1957-1958. [23] Le Conseil international des unions scientifiques (aujourd'hui Conseil international pour la science) a créé plusieurs centres mondiaux de données afin de limiter le risque de pertes et de maximiser l'accessibilité des données. En 1955, il recommandait que les données soient mises à disposition dans un

format lisible par machine. [24] En 1966, le Conseil international pour la science lança CODATA, une initiative visant à « promouvoir la coopération en matière de gestion et d'utilisation des données ». [25]

Ces formes primitives de données scientifiques ouvertes n'ont guère connu de développement. Les frictions de données et les résistances techniques à l'intégration des données externes étaient trop nombreuses pour mettre en place un écosystème durable de partage. Ces infrastructures étaient le plus souvent invisibles pour les chercheurs, car la majorité des recherches étaient effectuées par des bibliothécaires professionnels. Non seulement les systèmes d'exploitation de recherche étaient compliqués à utiliser, mais ils imposaient un impératif d'efficacité en raison du coût prohibitif des télécommunications à longue distance. [26] Les concepteurs avaient prévu que ces systèmes seraient directement utilisés par les chercheurs, mais les obstacles techniques et économiques ont rendu la démarche impossible :

> Les créateurs des premiers systèmes en ligne présumaient que la recherche serait effectuée par les utilisateurs finaux, ce qui a guidé jusqu'à leur conception. MEDLINE et NASA/RECON furent conçus à l'intention des chercheurs en médecine et cliniciens d'une part, et des ingénieurs et scientifiques de l'aérospatiale d'autre part. Pour de nombreuses raisons, cependant, dans les années soixante-dix la plupart des utilisateurs étaient des bibliothécaires et des intermédiaires qualifiés qui agissaient au nom des utilisateurs finaux. En fait, certains chercheurs professionnels voyaient d'un mauvais œil que des utilisateurs finaux enthousiastes puissent accéder aux terminaux. [27]

Christine Borgman ne se souvient pas de débats politiques marquants sur la signification, la production et la circulation des données

scientifiques, sauf dans quelques domaines précis (comme la climatologie) après 1966. [28] Avant l'avènement du Web, les différentes infrastructures scientifiques étaient quasiment impossibles à interconnecter. [29] Les projets et les communautés s'appuyaient sur leurs propres réseaux non connectés au niveau national ou institutionnel : « Internet était presque invisible en Europe, parce que chacun travaillait sur un ensemble distinct de protocoles réseau. ». [30] La communication entre les infrastructures scientifiques n'était pas seulement un défi dans l'espace, mais aussi dans le temps. Lorsqu'un protocole de communication était abandonné, les données et les connaissances qu'il diffusait risquaient aussi de disparaître : « La relation entre la recherche historique et l'informatique a été durablement affectée par des projets avortés, des pertes de données et des formats irrécupérables. ». [31]

Partage des données scientifiques sur le Web (1990-1995)

À l'origine, le World Wide Web fut conçu comme une infrastructure de science ouverte. Le partage des données et de leur documentation était l'un des principaux objectifs de la présentation initiale du World Wide Web lorsque le projet fut dévoilé en août 1991 : « Le projet WWW fut lancé pour permettre aux physiciens des particules de partager des données, des informations et de la documentation. L'extension du Web à d'autres domaines et la mise en place de serveurs passerelles pour d'autres données nous intéressent fortement. ». [32]

Le projet est né d'une infrastructure de la connaissance parente : ENQUIRE. Ce logiciel de gestion de l'information avait été commandé à Tim Berners-Lee par le CERN pour les besoins particuliers de la physique des particules. La structure d'ENQUIRE était plus proche d'un réseau interne de données : elle connectait des « nœuds » qui « pouvaient se référer à une personne, un module logiciel, etc. et pouvaient être raccordés entre eux par diverses relations telles que

« fait », « inclut », « décrit », etc. ». [33] . ENQUIRE avait beau « permettre l'établissement de liens aléatoires entre les informations », il n'était pas en mesure de « faciliter la collaboration souhaitée par la communauté internationale des chercheurs en physique des particules ». [34] Comme toutes les infrastructures informatiques scientifiques d'importance antérieures aux années 1990, le développement d'ENQUIRE finit par être entravé par le manque d'interopérabilité et la gestion complexe des communications en réseau : « Même si ENQUIRE permet de relier des documents et des bases de données, et si l'hypertexte constitue un format d'affichage commun, il restait le problème de faire communiquer entre eux des ordinateurs équipés de systèmes d'exploitation différents. ». [35]

Le Web a rapidement supplanté les infrastructures de données scientifiques fermées, même lorsqu'elles étaient plus avancées sur le plan informatique. De 1991 à 1994, les utilisateurs du Worm Community System, une importante base de données biologiques sur les vers, ont basculé vers le Web et Gopher. Le Web possédait peu de fonctions avancées pour la recherche de données et la collaboration, mais il était facilement accessible. À l'inverse, le Worm Community System ne pouvait être consulté que sur des terminaux spécialisés présents dans certaines institutions scientifiques : « L'adoption du WCS, un puissant système sur mesure doté d'une interface pratique, entraîne des inconvénients à l'intersection des habitudes de travail, de l'usage des ordinateurs et des ressources de laboratoire (…) Le World Wide Web, quant à lui, est accessible depuis une grande variété de terminaux et de connexions, et l'assistance informatique pour Internet est aisément disponible dans la plupart des établissements universitaires à travers des services commerciaux relativement abordables. ». [36]

La diffusion sur le Web a complètement transformé l'économie de la publication des données. Dans le secteur de l'impression, « le coût de

reproduction de grands volumes de données est prohibitif », mais avec ce nouveau paradigme les frais de stockage de la plupart des jeux de données sont faibles. [37] Dans ce nouvel environnement éditorial, les principaux facteurs de limitation du partage de données ne sont plus techniques ou économiques, mais sociaux et culturels.

Définition des données scientifiques ouvertes (1995-2010)

Le développement et la généralisation du World Wide Web ont levé de nombreuses barrières techniques et frictions qui gênaient la libre circulation des données. Néanmoins, il restait encore à définir les données scientifiques et à mettre en œuvre une nouvelle politique de recherche pour concrétiser la vision initiale du réseau de données pressentie par Tim Berners-Lee. À ce stade, les données scientifiques étaient largement définies à travers un processus d'ouverture, l'application de politiques ouvertes ayant relancé l'intérêt de lignes directrices, de principes et d'une terminologie exploitables.

La recherche climatique a servi de champ d'expérimentation pour définir le concept de données scientifiques ouvertes, comme ce fut le cas pour construire la première grande infrastructure de la connaissance dans les années 1950 et 1960. En 1995, le GCDIS a formulé un engagement clair sur l'échange complet et ouvert de données scientifiques : « Les programmes internationaux de surveillance environnementale et de recherche sur le changement planétaire dépendent énormément du principe d'un échange de données complet et ouvert (les données et les informations doivent être mises à disposition sans restriction ni discrimination, pour un coût ne dépassant pas celui de la reproduction et de la distribution). [38] L'élargissement du cadre et de la gestion des infrastructures de la connaissance a aussi poussé au partage des données. En effet, l'« attribution de la propriété des données » entre un grand nombre de particuliers et d'institutions n'a cessé de se complexifier. [39] Les

données ouvertes créent un cadre simplifié qui permet à tous les contributeurs et utilisateurs d'y accéder. [40]

Les données ouvertes ont été rapidement identifiées comme un objectif clé du mouvement émergent de la science ouverte. À l'origine focalisées sur les publications et les articles scientifiques, les initiatives internationales en faveur du libre accès ont élargi leur périmètre à toutes les productions scientifiques majeures. [41] En 2003, la Déclaration de Berlin soutenait la diffusion des « résultats originaux de recherches scientifiques, de données brutes et de métadonnées, de documents sources, de représentations numériques de documents picturaux et graphiques [et] de documents scientifiques multimédia ».

Après 2000, des entités internationales comme l'OCDE (Organisation de coopération et de développement économiques) ont joué un rôle déterminant dans l'élaboration de définitions génériques et transdisciplinaires des données scientifiques, car les politiques relatives aux données ouvertes doivent s'appliquer au-delà des limites spécifiques d'une discipline ou d'un pays. [42] L'une des premières définitions des données scientifiques à faire autorité date de 1999. [43] Elle est issue d'un rapport des Académies nationales des sciences : « Les données sont des faits, des lettres, des nombres ou des symboles qui décrivent un objet, une condition, une situation ou d'autres facteurs. ». [44] En 2004, les ministres de la science de tous les pays de l'OCDE ont signé une déclaration qui stipule notamment que toutes les données d'archives à financement public doivent être mises à la disposition de tous. [45] En 2007, l'OCDE a « codifié les principes d'accès aux données de la recherche issues de financements publics » [46] à travers des Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics, qui définissent les données scientifiques comme « des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la

communauté scientifique comme nécessaires pour valider des résultats de recherche ». [47] Les Principes faisaient office de recommandation non contraignante et affirmaient que « l'accès aux données de la recherche accroît le retour sur l'investissement public dans ce domaine, il renforce la liberté de l'investigation scientifique, il encourage la diversité des études et opinions, il favorise de nouveaux domaines d'activité et permet l'exploration de thèmes qui n'avaient pas été envisagés par les chercheurs d'origine ». [48]

Mise en œuvre des politiques (2010-…)

Après 2010, les institutions nationales et supranationales ont adopté une position plus interventionniste. Elles ont mis en œuvre de nouvelles politiques pour garantir et encourager l'ouverture des données scientifiques, le plus souvent dans le prolongement des programmes de données ouvertes existants. En Europe, le « Commissaire européen à la Recherche, à l'Innovation et à la Science, Carlos Moedas, a fait des données de recherche ouvertes l'une des priorités de l'UE en 2015 ». [49]

Initialement publiés en 2016, les Principes directeurs FAIR [50] sont devenus un cadre de référence pour l'ouverture des données scientifiques. [51] Ils avaient été élaborés deux ans plus tôt durant l'atelier de politique et de recherche Jointly Designing a Data FAIRport (Conception conjointe d'un FAIRport de données) organisé au Lorentz Center de Leyde (Pays-Bas). [52] Au cours des délibérations de l'atelier, « l'idée s'est imposée que, grâce à la définition d'un ensemble minimal de pratiques et de principes directeurs acceptés par les communautés et bénéficiant d'un ample soutien, toutes les parties prenantes pourraient plus facilement et plus efficacement découvrir, consulter, intégrer, réutiliser et citer les immenses volumes d'informations générés par la science contemporaine si féconde en données ». [53]

Les Principes ne cherchent pas à définir les données scientifiques, qui restent un concept relativement plastique, mais ils s'efforcent de décrire « ce qui constitue une « bonne gestion des données » ». [54] Ils recouvrent quatre aspects fondamentaux « qui servent de guide aux producteurs de données » : Facilité à trouver, Accessibilité, Interopérabilité et Réutilisabilité. [55] En outre, ils visent à promouvoir l'exploitabilité informatique en explicitant la sémantique sous-jacente des données. [56] Reconnaissant pleinement la complexité de la gestion des données, les Principes ne prétendent pas introduire un ensemble de recommandations rigides mais plutôt des « degrés FAIR » ajustables en fonction des coûts organisationnels et des restrictions externes en matière de droits d'auteur ou de protection de la vie privée. [57]

Les principes FAIR ont été immédiatement validés par les principales organisations internationales : « Les principes FAIR ont connu un développement rapide, obtenant la reconnaissance de l'Union européenne, du G7, du G20 et de l'initiative « Big Data to Knowledge » (BD2K) aux États-Unis. ». [58] En août 2016, la Commission européenne a mis en place un groupe d'experts pour la « concrétisation des données FAIR ». [59] En 2020, les principes FAIR restaient « les normes techniques les plus avancées à ce jour pour les données scientifiques ouvertes ». [60]

À la fin des années 2010, les politiques de données ouvertes étaient bien soutenues par les communautés scientifiques. Deux grandes enquêtes commandées par la Commission européenne en 2016 et 2018 soulignent un avantage perçu par la quasi-totalité des répondants : « 74 % des chercheurs déclarent que l'accès à d'autres données leur serait bénéfique. ». [61] Toutefois, selon des observations plus qualitatives recueillies lors de ces mêmes enquêtes, « les déclarations idéalisées des scientifiques, en contraste avec leurs pratiques réelles,

révèlent une situation plus ambiguë ». [62]

Diffusion des données scientifiques

Publication et édition

Jusque dans les années 2010, l'édition de données scientifiques concernait principalement « la publication de jeux de données associés à un article de revue ». [63] Cette démarche était documentée au moyen d'une déclaration d'accessibilité des données (DAS, de l'anglais « Data Accessibility Statement »). Plusieurs typologies de DAS ont été proposées. [64] [65] En 2021, Colavizza et al. ont identifié trois catégories ou niveaux d'accès :

- DAS 1 : « Données disponibles sur demande ou similaires » [66]
- DAS 2 : « Données disponibles avec l'article et ses fichiers supplémentaires » [67]
- DAS 3 : « Données disponibles dans un référentiel » [68]

Les fichiers de données supplémentaires sont apparus dès la première phase de transition vers l'édition scientifique numérique. Bien que le format des publications ait largement conservé les contraintes de l'impression, les « informations complémentaires » pouvaient inclure des éléments additionnels. [69] En tant que publication, le fichier de données supplémentaire possède un statut ambigu. En théorie, il s'agit d'un document brut qui donne accès au contexte de la recherche. En pratique, il doit souvent faire l'objet d'une conservation spéciale en vue de la publication. Il se concentre généralement sur les sources de données primaires, et non sur l'ensemble des observations ou des mesures effectuées dans le cadre de la recherche : « Il est souvent difficile d'identifier les « données » associée à un article, un exposé de conférence, un livre ou une autre publication [car] les investigateurs collectent des données en continu. ». [70] Par ailleurs, la sélection des données est aussi influencée par l'éditeur. La politique éditoriale de la

revue détermine en grande partie « ce qui fait partie du texte principal et ce qui relève des informations complémentaires », et les éditeurs hésitent tout particulièrement à inclure de grands jeux de données potentiellement difficiles à administrer sur le long terme. [71]

Les jeux de données scientifiques sont de plus en plus reconnus comme des publications scientifiques autonomes. L'assimilation des données à des articles universitaires visait à accroître le prestige et la reconnaissance des jeux de données publiés : « L'argument implicite est que la familiarité encouragera la publication des données. ». [72] Cette approche fut privilégiée par plusieurs éditeurs et référentiels car elle a facilité l'intégration des données dans les infrastructures d'édition existantes et une ample réutilisation des concepts éditoriaux initialement centrés sur les articles. [73] Les articles de données ont été explicitement introduits en tant que « mécanisme d'incitation à la publication des données en science de la biodiversité ». [74]

Citation et indexation

Les premières bases de données numériques des années 1950 et 1960 ont immédiatement soulevé des questions de citabilité et de descriptions bibliographiques. [75] La mutabilité de la mémoire informatique était une source particulière de difficulté : contrairement à leur version imprimée, les données numériques ne présentent aucune garantie de stabilité à long terme. En 1965, Ralph Bisco souligna que cette incertitude affectait tous les documents associés, comme les livres de codage, dont l'obsolescence peut être très rapide. La gestion des données doit trouver un juste milieu entre l'amélioration continue et une certaine forme de stabilité générique : « Le concept d'archives de données fluides, modifiables et continuellement améliorées signifie que les différents traitements des études, y compris le nettoyage, ne doivent pas affecter significativement les analyses antérieures. ». [76]

Dans le domaine des bases de données, le sujet des métadonnées

bibliographiques structurées est débattu depuis les années 1960. [77] En 1977, l'American Standard for Bibliographic Reference a adopté une définition de l'expression « fichier de données » qui insiste fortement sur la matérialité et la mutabilité du jeu de données : ni les dates ni les auteurs n'étaient mentionnés, contrairement au support ou à la « méthode d'encapsulation ». [78] Deux ans plus tard, Sue Dodd introduisit une convention alternative, qui rapprochait la citation des données de la norme des références appliquée aux autres publications scientifiques : [79] elle recommandait d'utiliser les titres, l'auteur, les éditions et la date, ainsi que des mentions alternatives pour les sous-documentations telles que le livre de codage. [80]

L'indexation des jeux de données fut radicalement transformée par le développement du Web, qui aplanit considérablement les obstacles au partage d'informations. [81] Dans ce processus, l'archivage, la durabilité et la persistance des données sont devenus des aspects fondamentaux. Du fait de l'évolution structurelle continue des sites Web, les identifiants d'objets numériques (ou DOI, de l'anglais « Digital Object Identifier ») permanents furent introduits pour les articles scientifiques afin d'éviter les phénomènes de liens rompus. Au début des années 2000, des programmes pilotes commencèrent à attribuer des DOI y compris aux jeux de données. [82] Tout en résolvant les problèmes concrets de durabilité des liens, la création de DOI et de normes de citation des données enrichit le processus de légitimation, qui assimile les jeux de données à des publications scientifiques standard et peut s'appuyer sur des sources de motivation similaires (comme les indices bibliométriques). [83]

La facilité à trouver les jeux de données et leur accessibilité sont de précieux avantages pour le traitement des citations. D'après une étude de 2021 portant sur 531 889 articles publiés dans PLOS, lorsqu'un article de revue comporte « un lien vers des données archivées dans un référentiel public », le « gain relatif [est] de 25,36 % pour le nombre

un référentiel public », le « gain relatif [est] de 25,56 % pour le nombre de citations en général ». [84] La diffusion des données en tant qu'éléments supplémentaires n'entraîne pas de bénéfice significatif en termes de citations, ce qui suggère que « l'avantage citationnel des DAS est moins lié à leur simple présence qu'à leur contenu ». [85]
En 2022, la reconnaissance des données scientifiques ouvertes n'était pas totalement achevée. Le logiciel de référence Zotero n'avait pas encore d'entrée spécifique pour le jeu de données.

Réutilisation et impact économique

Dans la recherche universitaire, le stockage et la redondance des données scientifiques ouvertes ont montré tout leur intérêt. En revanche, la conservation des données scientifiques non ouvertes est peu efficace. Si elles ne sont pas complètement perdues, elles ne peuvent « être récupérées qu'au prix d'efforts considérables de la part des auteurs ». [86]

L'analyse des usages des données scientifiques ouvertes s'est heurtée aux mêmes écueils que pour les autres contenus de même nature : si l'accès gratuit, universel et indiscriminé a manifestement élargi la portée, l'étendue et l'intensité de la réception, il l'a aussi rendue plus floue en raison de l'absence de transaction.

Ces difficultés sont encore accrues par la nouveauté des données en tant que publication scientifique : « Dans la pratique, il peut être difficile de contrôler la réutilisation des données, avant tout parce que les chercheurs citent rarement leur référentiel. ». [87]

En 2018, un rapport de la Commission européenne estimait que la non-ouverture des données scientifiques selon les principes FAIR avait un impact négatif annuel direct de 10,2 milliards EUR et indirect de 16 milliards EUR sur toute l'économie de l'innovation. [88] La mise en œuvre de données scientifiques ouvertes à l'échelle mondiale « aurait un impact considérable sur le temps passé à manipuler les données et

un impact considérable sur le temps passé à manipuler les données et sur leurs modalités de stockage ». [89]

Pratiques et culture des données

Le partage des données scientifiques est ancré dans les cultures scientifiques et les communautés de pratique. Avec la généralisation des outils numériques, les infrastructures, les pratiques et les représentations collectives des communautés de recherche reposent de plus en plus sur une vision partagée de la nature des données et de ce qu'elles permettent de faire. [90]

Les machines épistémiques existantes peuvent être plus ou moins prédisposées au partage des données. Les facteurs importants sont par exemple les valeurs partagées (individuelles ou collectives), l'attribution de la propriété des données et les collaborations fréquentes avec des acteurs externes potentiellement réticents au partage d'informations. [91]

Émergence d'une culture des données ouvertes

Le développement des données scientifiques ouvertes ne se limite pas à la recherche scientifique. Il implique un large éventail de parties prenantes : « Les tenants du partage des données sont nombreux : les organismes de financement publics et privés, les entités politiques telles que les académies nationales et les comités de financement, les éditeurs de revues, les éducateurs, le grand public et les chercheurs eux-mêmes. ». [92] En l'occurrence, le mouvement en faveur des données scientifiques ouvertes recoupe largement des initiatives plus globales en faveur des données ouvertes. [93] En partie, les normes définissant les données ouvertes utilisées par un large éventail d'acteurs publics et privés ont été élaborées par des chercheurs autour de questions scientifiques concrètes. [94] Le concept de transparence a

particulièrement contribué à créer des convergences entre la science ouverte, les données ouvertes et le gouvernement ouvert. En 2015, l'OCDE a décrit la transparence comme une « raison d'être de la science et des données ouvertes ». 95

Christine Borgman a identifié quatre grands arguments en faveur du partage des données qui sont largement repris lors des débats réglementaires et publics sur les données scientifiques ouvertes : 96

- Reproductibilité de la recherche : le manque de reproductibilité est souvent attribué à une recherche opaque et à une analyse défaillante des données. Par conséquent, « [la reproductibilité de la recherche] est un argument en faveur du partage des données de recherche à la fois puissant et problématique ». 97 La reproductibilité s'applique uniquement à « certains types de recherches », surtout en sciences expérimentales. 98
- Accessibilité publique : cet argument selon lequel « les produits à financement public devraient être accessibles au public » plaide « en faveur d'un gouvernement ouvert ». 99 Bien que directement inspiré d'arguments similaires favorables aux publications en libre accès, sa portée est plus limitée car les données scientifiques ouvertes « profitent directement à un nombre beaucoup plus restreint de personnes, avec des avantages variables selon les parties concernées ». 100
- Valorisation de la recherche : pour le secteur privé, les données scientifiques ouvertes peuvent avoir une forte valeur ajoutée. Cet argument est particulièrement utilisé pour soutenir « la nécessaire mise en place d'un plus grand nombre de référentiels capables de recevoir et de conserver les données de recherche, de meilleurs outils et services d'exploitation des données, et d'autres investissements dans les infrastructures de la connaissance ». 101
- Stimulation de la recherche et de l'innovation : les données scientifiques ouvertes peuvent améliorer considérablement la qualité de la recherche privée et publique. Cet argument appelle à « investir dans les infrastructures de la connaissance afin de soutenir les données de recherche [qui seront] conservées selon des standards élevés de pratiques professionnelles ». 102

Cependant, la collaboration entre les différents acteurs et parties prenantes du cycle de vie des données est incomplète. Même au sein des institutions universitaires, la coopération reste limitée : « La plupart des chercheurs effectuent des [recherches liées aux données] sans consulter un gestionnaire de données ou un bibliothécaire. » 103

consulter un gestionnaire de données ou un bibliothécaire. ». [103]

Le mouvement mondial en faveur des données ouvertes a partiellement perdu sa cohésion et son identité au cours des années 2010, car les débats sur la disponibilité des données et les licences ont cédé le pas aux questions spécifiques sur le domaine d'étude : « Lorsqu'il n'est plus question de militer pour l'accès ouvert mais de créer des infrastructures de données et de les exploiter, les objectifs divergents des membres historiques du mouvement des données ouvertes apparaissent clairement, et il peut s'avérer complexe de gérer les tensions engendrées. ». [104] La portée très générique de la définition des données ouvertes, qui vise à englober un très large ensemble de cultures de données, ne tient pas bien compte du niveau plus élevé d'accessibilité et de contextualisation requis par la recherche scientifique : « L'ouverture des données, c'est-à-dire leur libre réutilisation, est une condition nécessaire mais non suffisante pour la recherche. ». [105]

Idéal et mise en œuvre : le paradoxe du partage des données

Depuis les années 2000, les enquêtes menées auprès des communautés scientifiques soulignent un décalage constant entre la notion idéale du partage des données et son application pratique : « Lorsqu'on demande aux chercheurs d'aujourd'hui s'ils sont prêts à partager leurs données, la plupart répondent par l'affirmative. Et lorsqu'on leur demande s'ils les publient effectivement, ils reconnaissent généralement ne pas le faire. ». [106] La culture des données ouvertes n'est pas apparue ex nihilo. Elle doit s'accommoder des pratiques en place concernant les données scientifiques et d'une série de facteurs systémiques potentiellement nuisibles au partage : « Dans certains domaines, les chercheurs sont activement incités à ne pas réutiliser les données… Les carrières se font en cartographiant des territoires jusqu'alors inexplorés. ». [107]

En 2011, 67 % des 1 329 scientifiques interrogés voyaient le manque de partage de données comme « un obstacle majeur au progrès scientifique » [108] alors que, pourtant, « seul environ un tiers (36 %) reconnaissent que leurs données sont facilement accessibles à autrui ». [109] En 2016, une enquête menée auprès de chercheurs en sciences de l'environnement a révélé qu'ils soutiennent massivement un meilleur accès aux données ouvertes (99 % les considèrent comme au moins « assez importantes ») et les exigences institutionnelles en matière de données ouvertes (88 %). [110] Cependant, « même quand il y a volonté de partager les données, on observe des écarts vis-à-vis des pratiques courantes, par exemple l'envie de consacrer du temps et des ressources à la préparation et au chargement des données ». [111]

La prévalence des données accessibles et facilement trouvables est encore plus faible : « Malgré plusieurs décennies de mesures politiques en faveur d'un accès libre aux données, les quelques statistiques disponibles indiquent de faibles taux de diffusion ou de dépôt des données. ». [112] Dans un sondage réalisé en 2011 pour Science, seuls 7,6 % des chercheurs partageaient leurs données sur un référentiel communautaire, les sites Web locaux hébergés par des universités ou des laboratoires étant privilégiés. [113] Par conséquent, « beaucoup déploraient le manque d'archives et de métadonnées communes comme principal obstacle à l'utilisation et au stockage des données ». [114]

Selon Christine Borgman, le paradoxe du partage des données est en partie dû aux limites des politiques de données ouvertes, qui tendent à se concentrer sur « l'obligation ou l'incitation des chercheurs à publier leurs données » sans répondre à « la demande de données ou d'infrastructures requises pour soutenir la diffusion et la réutilisation ». [115]

Facilités et obstacles pour les données scientifiques ouvertes

En 2022, Pujol Priego, Wareham et Romasanta ont souligné que les incitations au partage des données scientifiques étaient avant tout collectives et qu'elles incluaient la reproductibilité, l'efficacité et la qualité scientifiques, ainsi que des rétributions plus individuelles telles que le crédit personnel. [116] Le surplus de visibilité fait partie des avantages individuels : les jeux de données ouverts offrent un avantage conséquent en termes de citations, mais uniquement s'ils ont été partagés sur un référentiel ouvert. [117]

Les principaux obstacles sont la nécessité de publier en premier, les contraintes juridiques et les inquiétudes concernant la perte de crédit ou de reconnaissance. [118] Pour un chercheur, les jeux de données peuvent constituer un atout majeur à troquer contre « un nouveau poste ou de nouvelles collaborations » [119] et leur publication peut être difficile à justifier sauf à « obtenir une valorisation en retour ». [120]

L'inexpérience du partage des données, plutôt qu'un rejet pur et simple des principes de la science ouverte, constitue aussi un obstacle majeur. Plusieurs enquêtes menées au début des années 2010 ont montré que les chercheurs « demandent rarement des données à d'autres investigateurs et (…) se voient tout aussi rarement demander leurs propres données ». [121] Ce phénomène crée une boucle de rétroaction négative, car les chercheurs font peu d'efforts pour assurer le partage des données, ce qui en décourage l'utilisation efficace, alors que « la plus forte demande de réutilisation des données est observée dans les domaines à dépendance mutuelle élevée ». [122] Dans la pratique, la réutilisation des données peut aussi être sous-estimée parce qu'elles ne sont pas considérées comme une publication prestigieuse et que les sources originales ne sont pas citées. [123]

D'après une étude empirique de 2021 portant sur 531 889 articles publiés dans PLOS, les incitations et encouragements non contraignants ont peu d'incidence sur le partage des données : « Les

politiques des revues qui encouragent la DAS au lieu de l'exiger ou de la rendre obligatoire n'ont que peu d'effet. ». [124]

Statut juridique

L'ouverture des données scientifiques a fait émerger des questions juridiques sur les droits de propriété, les droits d'auteur, la protection de la vie privée et la déontologie. S'il est communément admis que les chercheurs « sont propriétaires des données qu'ils collectent dans le cadre de leurs recherches », ce « point de vue est erroné » : [125] la création d'un jeu de données peut engager les droits de nombreux acteurs supplémentaires tels que les institutions (agences de recherche, bailleurs de fonds, organismes publics), les producteurs tiers et les particuliers (pour leurs données personnelles). [126] La situation juridique des données numériques a donc été décrite comme un « faisceau de droits » attendu que « la catégorie juridique de la « propriété » (…) n'est pas un modèle idoine pour traiter la complexité des problèmes de gouvernance des données ». [127]

Droit d'auteur

Le droit d'auteur était au cœur de la littérature juridique sur les données scientifiques ouvertes jusqu'aux années 2010. La légalité du partage des données fut très tôt identifiée comme un point crucial. Contrairement au partage des publications scientifiques, le principal obstacle n'est pas le droit d'auteur mais l'incertitude : « Le concept de « données » [était] un fait nouveau, forgé à l'ère de l'informatique, alors que le droit d'auteur est apparu à l'époque des publications imprimées. ». [128] En théorie, les dispositions relatives au droit d'auteur ne s'appliquent pas au simple recueil de faits et de chiffres. Dans la pratique, la notion de données est beaucoup plus large et peut inclure des contenus protégés ou des agencements créatifs de contenus non

protégés par le droit d'auteur.

Le statut des données dans les conventions internationales sur la propriété intellectuelle est ambigu. Selon l'article 2 de la Convention de Berne, « toutes les productions du domaine littéraire, scientifique et artistique » sont protégées. [129] Pourtant, les données de recherche ne sont souvent pas une création originale émanant en totalité d'un ou plusieurs auteurs, mais plutôt « un ensemble de faits, généralement rassemblés à l'aide d'instruments ou d'équipements scientifiques automatisés ou semi-automatisés ». [130] Il n'existe donc pas de convention universelle sur le droit d'auteur des données, et les débats sur « le périmètre d'application du droit d'auteur » sont toujours d'actualité, avec des résultats différents selon la juridiction ou les spécificités du jeu de données. [131] Ce manque d'harmonisation résulte logiquement de la nouveauté des « données de recherche » en tant que concept clé des études scientifiques : « Le concept de « données » est un fait nouveau, forgé à l'ère de l'informatique, alors que le droit d'auteur est apparu à l'époque des publications imprimées. ». [132]

Aux États-Unis, dans l'Union européenne et plusieurs autres juridictions, les lois sur le droit d'auteur ont établi une distinction entre les données elles-mêmes (qui peuvent être un « fait » non protégé) et la compilation des données (qui peut être un agencement créatif). [133] Ce principe est largement antérieur au débat politique contemporain sur les données scientifiques, puisque les premiers jugements en faveur des droits de compilation remontent au XIXe siècle.

Aux États-Unis, les droits de compilation ont été définis dans le Copyright Act de 1976, qui mentionne explicitement les jeux de données : la compilation est « une œuvre formée par la collecte et l'assemblage d'informations préexistantes ou de données » (§ 101). [134] Dans sa décision de 1991 lors de l'affaire Feist Publications, Inc. vs Rural Telephone Service Co, la Cour Suprême a clarifié le périmètre

et les limites du droit d'auteur sur les bases de données. En effet, selon la Cour l'« assemblage » des données doit être manifestement original et les « faits bruts » figurant dans la compilation restent dénués de protection. [135]
Même dans les juridictions où l'application du droit d'auteur aux productions de données reste incertaine et en partie théorique, le flou juridique demeure considérable. La frontière entre un ensemble de faits bruts et une compilation originale n'est pas clairement délimitée. [136]

Bien que les organisations scientifiques connaissent généralement bien la législation sur le droit d'auteur, la complexité des droits sur les données crée des difficultés sans précédent. [137] Après 2010, les juridictions nationales et supranationales ont partiellement modifié leur doctrine sur la protection des données de recherche au moyen de droits d'auteur. Le partage étant encouragé, les données scientifiques ont aussi été reconnues comme un bien public informel : « Les décideurs politiques, les bailleurs de fonds et les institutions universitaires s'efforcent de faire admettre que, si les publications et les connaissances provenant des données de recherche appartiennent à leurs auteurs, ces mêmes données de recherche sont à considérer comme un bien public afin que leur potentiel de valeur sociale et scientifique puisse se matérialiser. ». [138]

Droits sur les bases de données

L'Union européenne présente l'un des cadres de propriété intellectuelle les plus rigoureux pour les données, avec une double couche de droits : le droit d'auteur pour les compilations originales (comme aux États-Unis) et le droit sui generis pour les bases de données. [139] Les critères d'originalité des compilations ont été harmonisés entre les États membres par la Directive de 1996 concernant la protection juridique des bases de données et par plusieurs jurisprudences majeures de la Cour de justice de l'Union européenne, comme Infopaq International A/

S vs Danske Dagblades Forening c ou Football Dataco Ltd et al. vs Yahoo! UK Ltd. Globalement, il a été reconnu que des efforts importants dans la production du jeu de données ne suffisent pas pour revendiquer des droits de compilation, car la structure doit permettre à l'auteur « d'exprimer son esprit créateur de manière originale ». [140] La Directive concernant les bases de données a également introduit un cadre de protection original pour les jeux de données, le droit sui generis conféré à tout jeu de données qui nécessite un « investissement substantiel ». [141] Bien que d'une durée limitée à 15 ans, le droit sui generis peut en théorie prendre un caractère permanent puisqu'il est renouvelable à chaque mise à jour du jeu de données.

La portée du droit sui generis est si vaste en matière de durée et de protection que la jurisprudence européenne s'en est très peu emparée dans un premier temps, car elle plaçait la barre très haut pour son application. Cette inclination à la prudence s'inversa dans les années 2010, lorsque la décision de 2013 dans l'affaire Innoweb BV vs Wegener ICT Media BV et Wegener Mediaventions renforça la position des propriétaires de bases de données et condamna la réutilisation des données non protégées dans les moteurs de recherche en ligne. [142] La consolidation et l'expansion des droits relatifs aux bases de données demeurent controversées dans la réglementation européenne, car elles viennent en partie contredire l'engagement de l'Union en faveur de l'économie des données et de la science ouverte. [143] En dépit de quelques exceptions concernant les usages scientifiques et pédagogiques, la portée de ce droit est limitée (aucun droit de réutilisation ultérieure) et il n'a pas été transposé dans tous les États membres. [144]

Propriété

Les problèmes de droits d'auteur en lien avec les jeux de données

scientifiques ont encore été renforcés par les incertitudes entourant la propriété. La recherche est en grande partie une activité collaborative qui implique un large éventail de contributions. Des initiatives telles que CRediT (Contributor Roles Taxonomy, « Taxonomie des rôles des contributeurs ») ont identifié 14 rôles différents, dont 4 explicitement liés à la gestion des données (analyse formelle, investigation, conservation des données et visualisation). [145]

Aux États-Unis, la propriété des données de recherche est généralement « déterminée par l'employeur du chercheur », l'investigateur principal agissant en tant que gestionnaire plutôt que propriétaire des données. [146] Jusqu'au développement des données de recherche ouvertes, les institutions américaines étaient généralement plus réticentes à renoncer aux droits d'auteur sur les données plutôt que sur les publications, qui sont considérées comme des actifs stratégiques. [147] Dans l'Union européenne, aucun cadre sur la propriété des données ne fait l'unanimité. [148]

Les droits supplémentaires des acteurs externes ont également été évoqués, notamment dans la recherche médicale. Depuis les années 1970, les patients revendiquent une certaine forme de propriété sur les données générées lors des essais cliniques, notamment avec d'importantes controverses sur « les sujets de recherche et les patients qui seraient ou non propriétaires de leurs tissus et de leur ADN ». [149]

Protection de la vie privée

De nombreux projets scientifiques s'appuient sur la collecte de données concernant les personnes, notamment dans la recherche médicale et les sciences sociales. Ces situations exigent que les politiques de partage de données prennent en compte de manière équilibrée la conservation et la protection des données personnelles. [150]

Les chercheurs, et plus particulièrement les investigateurs principaux, sont soumis à des obligations de confidentialité dans plusieurs juridictions. [151] Depuis la fin du XXe siècle, les données de santé sont de plus en plus réglementées soit par la loi, soit par des accords sectoriels. En 2014, l'Agence européenne des médicaments a clairement modifié les règles de partage des données issues des essais cliniques pour éviter la divulgation des données personnelles et de toute information à valeur commerciale. Cette évolution de la réglementation européenne « est susceptible d'influencer la pratique mondiale consistant à partager les données des essais cliniques dans un format ouvert ». [152]

Les programmes et pratiques de gestion de la recherche doivent être par nature ouverts, transparents et respectueux de la vie privée.

Licences libres

Le meilleur cadre juridique pour lever les restrictions et les ambiguïtés de la définition juridique des données scientifiques est assurément la licence ouverte. En 2003, la Déclaration de Berlin a appelé à renoncer universellement aux droits de réutilisation des contributions scientifiques incluant explicitement « les données brutes et les métadonnées ». [153]

Contrairement aux licences ouvertes pour les publications, dont la création fut assez rapide, les licences pour les données scientifiques ouvertes suivent un développement compliqué. Des protections et principes spécifiques, comme le droit sui generis sur les bases de données dans l'Union européenne ou la distinction entre faits simples et compilation originale, n'ont pas été anticipés. Jusque dans les années 2010, paradoxalement les licences libres pouvaient ajouter des restrictions à la réutilisation des jeux de données, notamment en ce qui concerne les attributions (facultatives pour les objets non protégés par le droit d'auteur comme les faits bruts) : « Dans de tels cas, dès lors

le droit d'auteur comme les faits bruts). [154] « Dans de tels cas, dès lors qu'aucun droit n'est attaché aux données de recherche, il n'y a pas lieu de leur associer une licence. ».

Pour contourner le problème, plusieurs institutions comme le Harvard-MIT Data Center commencèrent à partager les données dans le domaine public. [155] Cette approche garantit qu'aucun droit ne s'applique aux éléments non protégés par le droit d'auteur. Pourtant, le domaine public et certains outils associés comme la marque correspondante ne constituent pas un contrat juridique en bonne et due forme, et ils varient fortement d'une juridiction à l'autre. [156] Introduite en 2009, la licence Creative Commons Zero (ou CC0) fut immédiatement envisagée pour s'appliquer aux données. [157] Depuis, elle est devenue « l'outil recommandé pour diffuser les données de recherche dans le domaine public ». [158] Conformément aux principes de la Déclaration de Berlin, il ne s'agit pas d'une licence mais d'une renonciation, avec l'engagement suivant de la part du producteur des données : « Le Déclarant affirme par la présente céder, abandonner, et renoncer ouvertement, pleinement, définitivement, irrévocablement et sans conditions à tous ses Droits d'Auteur et Droits Voisins. ».

D'autres approches prévoient la conception d'une nouvelle licence libre pour démêler l'écheveau des attributions propres à la protection des bases de données. En 2009, l'Open Knowledge Foundation a publié la licence Open Database qui fut adoptée par de grands projets en ligne comme OpenStreetMap. Depuis 2015, toutes les licences Creative Commons ont été mises à jour pour s'appliquer pleinement aux jeux de données, les droits relatifs aux bases de données ayant été explicitement prévus dans la version 4.0. [159]

Gestion des données scientifiques ouvertes

La gestion des données s'est récemment installée au centre des débats politiques et de la recherche sur les données scientifiques ouvertes. Les principes de référence FAIR sont volontairement centrés sur les caractéristiques clés de la « bonne gestion des données » dans le contexte scientifique. [160] S'agissant de la recherche, la gestion des données est fréquemment associée aux cycles de vie des données. Divers modèles de cycles de vie à différents stades ont été théorisés par les institutions, les infrastructures et les communautés scientifiques, même si « de tels cycles de vie constituent une simplification de la vie réelle, qui est beaucoup moins linéaire et plus itérative dans la pratique ». [161]

Intégration au flux de travaux de la recherche

Alors que les premières politiques en faveur des données scientifiques ouvertes incitaient au partage de données en général, de plus en plus la complexité, les coûts sous-jacents et les exigences de la gestion des données scientifiques sont pris en compte : « Le partage des données est difficilement réalisable et justifiable par le retour sur investissement. ». [162] L'ouverture des données n'est pas une simple tâche supplémentaire. Elle doit être envisagée tout au long du processus de recherche, car elle « exige des changements dans les méthodes et pratiques de recherche ». [163]

L'ouverture des données de recherche change les modalités de gestion des coûts et des avantages. Le partage public de données induit un nouveau cadre de communication très différent de l'échange privé de données avec des collaborateurs ou des partenaires de recherche. La collecte, l'usage prévu et la limitation des données doivent être expliqués, car il n'est pas possible de s'appuyer sur des connaissances informelles existantes : « La documentation et les représentations sont les seuls moyens de communication entre le créateur et l'utilisateur des données. » [164] L'absence de documentation appropriée signifie que la

données. ». [164] L'absence de documentation appropriée signifie que la charge de la recontextualisation incombe aux utilisateurs potentiels et peut, en définitive, rendre le jeu de données inutilisable. [165]

La publication nécessite une vérification approfondie de la propriété des données et des responsabilités juridiques en cas d'utilisation abusive. Cette phase de clarification est encore plus complexe dans les projets de recherche internationaux qui peuvent croiser les juridictions. [166] En outre, le partage des données et les principes de la science ouverte offrent des avantages significatifs à long terme qui ne sont pas forcément visibles dans l'immédiat. La documentation des jeux de données permet de clarifier leur chaîne de provenance. Elle garantit que les données originales n'ont pas été excessivement altérées ou, si c'est le cas, que toutes les opérations ont été intégralement documentées. [167] La publication sous licence libre permet aussi de déléguer certaines tâches à des acteurs externes, par exemple la conservation à long terme.

La fin des années 2010 a vu apparaître une nouvelle littérature spécialisée dans la gestion des données au sein de la recherche, qui permet de codifier les pratiques et les principes réglementaires existants. [168] [169] [170]

Stockage et conservation

La disponibilité des données scientifiques non ouvertes décline rapidement : en 2014, une étude rétrospective des jeux de données biologiques a montré que « les chances qu'un jeu de données soit considéré comme toujours existant diminuent de 17 % par an ». [171] En conséquence, la « part des jeux de données encore existants est passée de 100 % pour 2011 à 33 % pour 1991 ». [172] La perte de données a aussi été soulignée comme un problème important au sein de revues majeures telles que Nature ou Science. [173]

Systématiquement, les enquêtes sur les pratiques de recherche montrent que les normes de stockage, les infrastructures et les flux de travaux demeurent insatisfaisants dans la plupart des disciplines. Très tôt, le stockage et la conservation des données scientifiques ont été identifiés comme des aspects critiques, en particulier s'agissant des données d'observation dont la conservation est considérée comme essentielle, parce qu'elles sont les plus difficiles à reproduire. [174] Une enquête menée en 2017-2018 auprès de 1 372 chercheurs contactés par l'intermédiaire de l'Union américaine de géophysique montre que seuls « un quart et un cinquième des répondants » mentionnent les bonnes pratiques de stockage des données. [175] Le stockage à court terme et non durable reste très répandu, 61 % des personnes interrogées stockant majoritairement ou totalement leurs données sur leur ordinateur personnel. [176] En raison de leur facilité d'utilisation à l'échelle individuelle, les solutions de stockage non durables sont plébiscitées dans la plupart des disciplines : « Ce décalage entre les bonnes pratiques et la satisfaction peut montrer que le stockage des données leur semble moins important que la collecte et l'analyse. ». [177]

Initialement publié en 2012, le modèle de référence de l'Open Archival Information System stipule que les infrastructures scientifiques doivent tendre vers la conservation à long terme, c'est-à-dire une durée « suffisamment longue pour prendre en compte les effets du progrès technologique, notamment les nouveaux médias et formats de données ou l'évolution de la communauté d'utilisateurs ». [178] Par conséquent, les bonnes pratiques de gestion des données impliquent à la fois le stockage (pour sauvegarder matériellement les données) et, plus important encore, la conservation « afin de préserver les connaissances sur les données pour en faciliter la réutilisation ». [179]

Le partage des données sur un référentiel public a contribué à sécuriser la conservation du fait de l'engagement à long terme des

infrastructures de données et de la redondance potentielle des données ouvertes. Une étude de 2021 portant sur 50 000 DAS d'articles publiés dans PLOS One a montré que 80 % des jeux de données pouvaient être récupérés automatiquement et que 98 % de ceux comportant un DOI pouvaient être récupérés automatiquement ou manuellement. En outre, l'accessibilité des publications plus anciennes ne diminuait pas significativement : « Grâce aux URL et aux DOI, les données et le code associés aux articles sont davantage susceptibles d'être disponibles dans le temps. ». [180] Aucun avantage majeur n'a été constaté lorsque les données ouvertes n'étaient pas correctement liées ou documentées : « Exiger que les données scientifiques soient partagées sous une forme ou une autre n'est pas suffisant pour les rendre conformes aux principes FAIR, car les études ont souvent démontré qu'une forte proportion des jeux de données ostensiblement partagés n'est pas réellement accessible. ». [181]

Plan et gouvernance

La gestion des données de recherche peut être organisée dans un plan de gestion des données ou PGD.

La création des plans de gestion des données remonte à 1966, lorsqu'il a fallu répondre aux besoins spécifiques de la recherche en aéronautique et en ingénierie, déjà confrontée à des frictions de données de plus en plus complexes. [182] Les premiers cas portaient sur des questions matérielles liées à l'accès, au transfert et au stockage des données : « Jusqu'au début des années 2000, les PGD étaient utilisés comme suit : dans certains domaines, pour des projets d'une grande complexité technique et à des fins limitées de collecte et de traitement des données à mi-étude. ». [183]

Après 2000, la mise en œuvre de grandes infrastructures de recherche et le développement de la science ouverte ont transformé le périmètre et l'objectif des plans de gestion des données. Les décideurs

politiques, plutôt que les scientifiques, ont joué un rôle déterminant dans cette évolution : « C'est à partir de 2009 que furent publiées les premières instructions générales destinées aux chercheurs sur la création des PGD, à la suite des parutions du JISC et de l'OCDE (…) Nous en déduisons que l'utilisation des PGD fut imposée à la communauté des chercheurs depuis l'extérieur. ». 184

Des études empiriques concernant les pratiques liées aux données dans la recherche ont « mis en évidence la nécessité pour les organisations d'offrir aux scientifiques une formation et une assistance plus formelles sur la gestion des données ». 185 Dans une enquête internationale menée en 2017-2018 auprès de 1 372 scientifiques, la plupart des demandes d'aide et de formalisation étaient associées au plan de gestion des données : « création de plans de gestion des données (33,3 %) ; formation aux meilleures pratiques de gestion des données (31,3 %) ; aide à la création de métadonnées pour décrire les données ou jeux de données (27,6 %) ». 186 De plus en plus, l'expansion des processus de collecte et d'analyse est venue remettre en cause bien des pratiques informelles et non codifiées de traitement des données.

La participation d'intervenants externes aux projets de recherche peut créer d'importantes tensions avec les principes de partage des données ouvertes. Les contributions des acteurs commerciaux peuvent reposer notamment sur une certaine forme d'exclusivité et d'appropriation des résultats finaux de la recherche. En 2022, Pujol Priego, Wareham et Romasanta ont élaboré plusieurs stratégies d'adaptation pour surmonter ces problèmes, comme la modularité des données (partage limité à une partie des données) et le délai (embargo d'un an avant la publication finale des données). 187

Infrastructures de science ouverte

La recommandation de l'Unesco sur la science ouverte, approuvée en novembre 2021, les définit comme « des infrastructures de recherche partagées qui sont nécessaires pour soutenir la science ouverte et répondre aux besoins des différentes communautés ». [188] Les infrastructures de science ouverte ont été reconnues comme un facteur crucial de la mise en œuvre et du développement des politiques de partage des données. [189]

Les principales formes d'infrastructures de données scientifiques ouvertes comprennent les référentiels, les plateformes d'analyse, les index, ainsi que les bibliothèques et les archives numérisées. [190] [191] Grâce aux infrastructures, les frais de publication, d'administration et d'indexation des jeux de données ne sont pas intégralement à la charge des chercheurs et des institutions. En outre, elles jouent un rôle clé dans la définition et l'adoption des normes de données ouvertes, notamment pour l'octroi de licences ou la documentation.

À la fin des années 1990, la création d'une infrastructure informatique scientifique à caractère public est devenue un enjeu politique majeur : [192] « Le manque d'infrastructures soutenant la diffusion et la réutilisation a été constaté dans une partie des premiers rapports politiques sur le partage de données. ». [193] La première vague de projets scientifiques en ligne dans les années 1990 et au début des années 2000 a révélé des questions de durabilité cruciales. Avec un financement alloué pour une période précise, les outils en ligne, les plateformes de publication et les bases de données critiques étaient difficilement administrables, [194] et les gestionnaires de projets vivaient dans l'angoisse « entre l'octroi de subventions et le financement opérationnel au long cours ». [195] Après 2010, la consolidation et l'expansion des infrastructures scientifiques commerciales, par exemple à travers l'acquisition des référentiels ouverts Digital Commons et SSRN par Elsevier, ont suscité une nouvelle vague d'appels à la constitution d'« infrastructures contrôlées par la

communauté ». [196] En 2015, Cameron Neylon, Geoffrey Bilder et Jenifer Lin ont publié l'ouvrage de référence Principles for Open Scholarly Infrastructure (Principes pour les infrastructures savantes ouvertes) [197] dont les préceptes ont été approuvés par de grandes infrastructures comme Crossref, [198] OpenCitations [199] et Data Dryad [200] . Avant 2021, les services publics et les infrastructures de recherche auront largement adopté la science ouverte comme partie intégrante de leur activité et de leur identité : « La science ouverte est le discours dominant auquel se réfèrent les nouveaux services en ligne destinés à la recherche. ». [201] Selon la feuille de route 2021 du Forum stratégique européen sur les infrastructures de recherche (ESFRI), en Europe les principales infrastructures ont adopté les principes de la science ouverte. « La plupart des infrastructures de recherche mentionnées dans la feuille de route de l'ESFRI sont à l'avant-garde du mouvement de la science ouverte et contribuent grandement à la transformation numérique en remodelant tout le processus de recherche à l'aune de ce paradigme. ». [202]

Les infrastructures de science ouverte représentent un plus haut niveau d'engagement en matière de partage des données. Elles s'appuient sur des investissements importants et récurrents pour garantir une maintenance et une documentation efficaces des données ainsi que pour « ajouter de la valeur aux données grâce aux métadonnées, à la provenance, à la classification, aux normes des structures de données et à la migration ». [203] En outre, les infrastructures doivent intégrer les normes et les usages prévus par les communautés scientifiques dont elles servent les intérêts : « Les plus performantes deviennent des collections de référence qui attirent des financements à long terme et peuvent servir d'étalon à leurs communautés. ». [204] L'application de normes ouvertes est l'un des principaux défis des grandes infrastructures ouvertes européennes. En effet, elle implique parfois de trancher entre des normes concurrentes et de garantir qu'elles soient mises à jour et accessibles à travers des

et de garantir qu'elles seront mises à jour et accessibles à travers des API ou d'autres points terminaux. [205]

La définition conceptuelle de l'infrastructure de science ouverte a été largement influencée par l'analyse d'Elinor Ostrom sur les communs, et plus particulièrement sur les communs de la connaissance. Dans le même esprit qu'Elinor Ostrom, Cameron Neylon sous-entend que les infrastructures ouvertes ne sont pas seulement caractérisées par la gestion d'un ensemble de ressources communes, mais aussi par l'élaboration d'une gouvernance et de normes partagées. [206] La diffusion des données scientifiques ouvertes soulève également d'épineuses questions de gouvernance. Au moment de déterminer la propriété des données, de choisir une licence libre et d'appliquer des règles de protection de la vie privée, « des négociations permanentes sont nécessaires » et il convient d'impliquer de nombreuses parties prenantes. [207]

Au-delà de leur intégration au sein de communautés scientifiques particulières, les infrastructures de science ouverte ont des liens étroits avec les mouvements en faveur de l'open source et des données ouvertes. Parmi les infrastructures européennes interrogées par SPARC, 82 % déclarent avoir partiellement construit des logiciels open source et 53 % possèdent des infrastructures technologiques exclusivement open source. [208] Les infrastructures de science ouverte intègrent de préférence des normes provenant d'institutions homologues. Parmi les infrastructures européennes, « les systèmes les plus couramment cités et donc les infrastructures essentielles pour beaucoup sont ORCID, Crossref, DOAJ, BASE, OpenAIRE, Altmetric et DataCite, la plupart étant à but non lucratif ». [209] Les infrastructures de science ouverte figurent ainsi parmi les nouveaux « communs de la science ouverte véritablement interopérables » fondés sur le principe que « les outils de recherche centrés sur le chercheur, peu coûteux, innovants et interopérables sont supérieurs au système actuel,

largement fermé ». 210

Plan

Définition

- Données scientifiques
- Données scientifiques ouvertes

Historique

- Développement des infrastructures de la connaissance (1945-1960)
- Ouverture et partage des données : premières tentatives (1960-1990)
- Partage des données scientifiques sur le Web (1990-1995)
- Définition des données scientifiques ouvertes (1995-2010)
- Mise en œuvre des politiques (2010-…)

Diffusion des données scientifiques

- Publication et édition
- Citation et indexation
- Réutilisation et impact économique

Pratiques et culture des données

- Émergence d’une culture des données ouvertes
- Idéal et mise en œuvre : le paradoxe du partage des données
- Facilités et obstacles pour les données scientifiques ouvertes

Statut juridique

- Droit d’auteur
- Droits sur les bases de données
- Propriété
- Protection de la vie privée
- Licences libres

- Licences libres

Gestion des données scientifiques ouvertes

- Intégration au flux de travaux de la recherche
- Stockage et conservation
- Plan et gouvernance
- Infrastructures de science ouverte

Notes

1. Spiegelhalter, D. Open data and trust in the literature. The Scholarly Kitchen. Consulté le 7 septembre 2018.
2. Wilkinson et al. 2016.
3. Lipton 2020, p. 19.
4. Borgman 2015, p. 18.
5. Lipton 2020, p. 59.
6. Lipton 2020, p. 59.
7. Lipton 2020, p. 61.
8. ARTICLE 29 2; DISSEMINATION OF RESULTS 2; OPEN ACCESS 2; VISIBILITY OF EU FUNDING, Draft of the H2020 Model Grant Agreement.
9. Lipton 2020, p. 61.
10. National Academies 2012, p. 1.
11. Borgman 2015, pp. 41;5.
12. Pujol Priego, Wareham & Romasanta 2022, p. 220.
13. Edwards et al. 2011, p. 669.
14. Pujol Priego, Wareham & Romasanta 2022, p. 224.
15. Pujol Priego, Wareham & Romasanta 2022, p. 225

15. Pujol Priego, Wareham & Romasanta 2022, p. 220.

16. Rosenberg 2018, pp. 5571;558.

17. Buckland 1991.

18. Edwards 2010, p. 84.

19. Edwards 2010, p. 99.

20. Edwards 2010, p. 102.

21. Machado, Jorge. Open data and open science. In Albagli, Maciel, Abdo. Open Science, Open Questions, 2015.

22. Shankar, Eschenfelder & Downey 2016, p. 63.

23. Committee on Scientific Accomplishments of Earth Observations from Space, National Research Council (2008). Earth Observations from Space: The First 50 Years of Scientific Achievements. The National Academies Press. p. 6. ISBN 978-0-309-11095-2. Consultation le 24/11/2010.

24. World Data Center System (2009-09-18). About the World Data Center System. NOAA, National Geophysical Data Center. Consultation le 24/11/2010.

25. Borgman 2015, p. 7.

26. Regazzi 2015, p. 128.

27. Bourne & Hahn 2003, p. 397.

28. Borgman 2015, p. 7.

29. Campbell-Kelly Garcia-Swartz 2013.

30. Berners-Lee Fischetti 2008, p. 17.

31. Dacos 2013.

32. Tim Berners-Lee, « Qualifiers on Hypertext Links », courriel envoyé le 6 août 1991 au groupe alt.hypertext.

33. Hogan 2014, p. 20.

34. Bygrave & Bing 2009, p. 30.

35. Berners-Lee & Fischetti 2008, p. 17.

36. Star & Ruhleder 1996, p. 131.
37. Borgman 2015, p. 217.
38. National Research Council (1995). On the Full and Open Exchange of Scientific Data. Washington, DC: The National Academies Press. doi:10.17226/18769. ISBN 978-0-309-30427-6.
39. Pujol Priego, Wareham & Romasanta 2022, p. 223.
40. Pujol Priego, Wareham & Romasanta 2022, p. 223.
41. Lipton 2020, p. 16.
42. Lipton 2020, p. 59.
43. Lipton 2020, p. 59.
44. National Research Council 1999, p. 16.
45. OECD Declaration on Open Access to publicly funded data Archived 20 April 2010 at the Wayback Machine.
46. Lipton 2020, p. 17.
47. OECD 2007, p. 13.
48. OECD 2007, p. 4.
49. Pujol Priego, Wareham & Romasanta 2022, p. 220.
50. Wilkinson et al. 2016.
51. Pujol Priego, Wareham & Romasanta 2022, p. 220.
52. Wilkinson et al. 2016, p. 8.
53. Wilkinson et al. 2016, p. 3.
54. Wilkinson et al. 2016, p. 1.
55. Wilkinson et al. 2016, p. 1.
56. Wilkinson et al. 2016, p. 3.
57. Wilkinson et al. 2016, p. 4.

58. van Reisen et al. 2020.
59. Horizon 2020 Commission expert group on Turning FAIR data into reality (E03464).
60. Lipton 2020, p. 66.
61. Pujol Priego, Wareham & Romasanta 2022, p. 241.
62. Pujol Priego, Wareham & Romasanta 2022, p. 241.
63. Borgman 2015, p. 48.
64. Federer et al. 2018.
65. Colavizza et al. 2020.
66. Colavizza et al. 2020, p. 5.
67. Colavizza et al. 2020, p. 5.
68. Colavizza et al. 2020, p. 5.
69. Borgman 2015, p. 217.
70. Borgman 2015, p. 216.
71. Borgman 2015, p. 216.
72. Borgman 2015, p. 48.
73. Borgman 2015, p. 48.
74. Chavan & Penev 2011.
75. Crosas 2014, p. 63.
76. Bisco 1965, p. 148.
77. Crosas 2014, p. 63.
78. Dodd 1979, p. 78.
79. Crosas 2014, p. 63.
80. Dodd 1979.
81. Crosas 2014, p. 63.

82. Brase 2004.
83. Borgman 2015, p. 47.
84. Colavizza et al. 2020, p. 12.
85. Colavizza et al. 2020, p. 10.
86. Vines et al. 2014, p. 96.
87. Lipton 2020, p. 65.
88. European Commission 2018, p. 31.
89. European Commission 2018, p. 31.
90. Pujol Priego, Wareham & Romasanta 2022, p. 224.
91. Pujol Priego, Wareham & Romasanta 2022, p. 224-225.
92. Borgman 2015, p. 208.
93. Davies et al. 2019, p. 1.
94. Borgman 2015, p. 44.
95. Lyon, Jeng & Mattern 2017, p. 47.
96. Borgman 2015, p. 208.
97. Borgman 2015, p. 209.
98. Borgman 2015, p. 209.
99. Borgman 2015, p. 211.
100. Borgman 2015, p. 212.
101. Borgman 2015, p. 212.
102. Borgman 2015, p. 212.
103. Tenopir et al. 2020, p. 12.
104. Davies et al. 2019, p. 6.
105. Borgman 2015, p. 283.

106. Borgman 2015, p. 205.
107. Borgman 2015, p. 213.
108. Tenopir et al. 2011, p. 7.
109. Tenopir et al. 2011, p. 9.
110. Schmidt, Gemeinholzer & Treloar 2016.
111. Schmidt, Gemeinholzer & Treloar 2016.
112. Borgman 2015, p. 206.
113. Science 2011.
114. Science 2011.
115. Borgman 2015, p. 207.
116. Pujol Priego, Wareham & Romasanta 2022, p. 226.
117. Colavizza et al. 2020, p. 12.
118. Tenopir et al. 2020, p. 5.
119. Borgman 2015, p. 217.
120. Borgman 2015, p. 217.
121. Borgman 2015, p. 213.
122. Borgman 2015, p. 213.
123. Borgman 2015, p. 223.
124. Colavizza et al. 2020, p. 13.
125. Lipton 2020, p. 127.
126. Lipton 2020, p. 127.
127. Kerber 2021, p. 1.
128. Lipton 2020, p. 119.
129. Lipton 2020, p. 119.

130. Lipton 2020, p. 119.
131. Lipton 2020, p. 119.
132. Lipton 2020, p. 119.
133. Lipton 2020, p. 119.
134. Lipton 2020, p. 122.
135. Lipton 2020, p. 122.
136. Lipton 2020, p. 123.
137. Lipton 2020, p. 126.
138. Pujol Priego, Wareham & Romasanta 2022, p. 224.
139. Lipton 2020, p. 123.
140. Article 6, Directive 2006/116/EC.
141. Lipton 2020, p. 124.
142. Lipton 2020, p. 125.
143. Lipton 2020, p. 125.
144. Lipton 2020, p. 125.
145. Allen, O'Connell & Kiermer 2019, p. 73.
146. Lipton 2020, p. 129.
147. Lipton 2020, p. 130.
148. Lipton 2020, p. 131.
149. Lipton 2020, p. 130.
150. Lipton 2020, p. 138.
151. Lipton 2020, p. 138.
152. Lipton 2020, p. 139.
153. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities.

154. Lipton 2020, p. 133.
155. Lipton 2020, p. 134.
156. Lipton 2020, p. 134.
157. Schofield et al. 2009.
158. Lipton 2020, p. 132.
159. Lipton 2020, p. 133.
160. Wilkinson et al. 2016, p. 1.
161. Cox & Verbaan 2018, p. 26-27.
162. Borgman 2015, p. 214.
163. Borgman 2015, p. 214.
164. Borgman 2015, p. 220.
165. Borgman 2015, p. 222.
166. Borgman 2015, p. 218.
167. Borgman 2015, p. 221.
168. Briney 2015.
169. Cox & Verbaan 2018.
170. Tibor 2021.
171. Vines et al. 2014.
172. Vines et al. 2014, p. 96.
173. Tedersoo et al. 2021.
174. Pujol Priego, Wareham & Romasanta 2022, p. 223.
175. Tenopir et al. 2020, p. 11.
176. Tenopir et al. 2020, p. 11.
177. Tenopir et al. 2020, p. 11.

178. CCSDS 2012, p. 1.
179. Lipton 2020, p. 73.
180. Federer 2022, p. 9.
181. Federer 2022, p. 11.
182. Smale et al. 2020, p. 3.
183. Smale et al. 2020, p. 4.
184. Smale et al. 2020, p. 9.
185. Tenopir et al. 2020, p. 13.
186. Tenopir et al. 2020, p. 13.
187. Pujol Priego, Wareham & Romasanta 2022, p. 239-240.
188. UNESCO Recommendation on Open Science, 2021, CL/4363.
189. Borgman 2015, p. 224.
190. Ficarra et al. 2020, p. 16.
191. Borgman 2015, p. 225.
192. Borgman 2007, p. 21.
193. Borgman 2015, p. 224.
194. Dacos 2013.
195. Skinner 2019, p. 6.
196. Joseph 2018, p. 1.
197. Neylon et al. 2015.
198. Crossref’s Board votes to adopt the Principles of Open Scholarly Infrastructure.
199. OpenCitations’ compliance with the Principles of Open Scholarly Infrastructure.
200. Dryad’s Commitment to the Principles of Open Scholarly Infrastructure.
201. Fecher et al. 2021, p. 505.

202. ESFRI Roadmap 2021, p. 159.
203. Borgman 2015, p. 226.
204. Borgman 2015, p. 225.
205. Ficarra et al. 2020, p. 23.
206. Neylon 2017, p. 7.
207. Borgman 2015, p. 229.
208. Ficarra et al. 2020, p. 29.
209. Ficarra et al. 2020, p. 50.
210. Ross-Hellauer et al. 2020, p. 13.

Rapports

- National Research Council (1999). A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases (Report). National Academies Press. Retrieved 2022-05-18.
- OECD (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding (Report). Paris: Organisation for Economic Co-operation and Development. Retrieved 2022-05-18.
- CCSDS (2012). Reference Model for an Open Archival Information System (OAIS) (Report). p. 135.
- European Commission (2018). Cost-benefit analysis for FAIR research data: cost of not having FAIR research data (Report). LU: Office des publications de l'Union européenne. doi:10.2777/02999. Retrieved 2022-06-18.
- Astell, Mathias; Hrynaszkiewicz, Iain; Allin, Katie; Penny, Dan; Mithu Lucraft; Baynes, Grace; Springer Nature Admin (2018). Practical challenges for researchers in data sharing – Springer Nature survey data (anonymised) (Report). Springer Nature. Retrieved 2022-09-11.
- Skinner, Katherine (2019). Mapping the Scholarly Communication Landscape: 2019 Census (Report). Educopia Institute. S2CID 201314019. Retrieved 2021-12-12.
- European Commission (2019). Horizon 2020 Annotated Model Grant A greements (Report). European Commission.
- Ficarra, Victoria; Fosci, Mattia; Chiarelli, Andrea; Kramer, Bianca; Proudman, Vanessa

(2020-10-30). Scoping the Open Science Infrastructure Landscape in Europe (Report). Retrieved 2021-10-31.
- ESFRI (2021). ESFRI Roadmap (PDF) (Report). ESFRI.
- Ross-Hellauer, Tony; Fecher, Benedikt; Shearer, Kathleen; Rodrigues, Eloy (2019-09-03). Pubfair: a framework for sustainable, distributed, open science publishing services (Report). Retrieved 2021-12-12.

Articles scientifiques

- Bisco, Ralph L. (1965-09-01). « Social Science Data Archives Technical Considerations ». *Social Science Information*. **4** (3): 129–150. doi:10.1177/053901846500400311. ISSN 0539-0184. S2CID 144164959.
- Dodd, Sue A. (1979). « Bibliographic references for numeric social science data files: Suggested guidelines ». *Journal of the American Society for Information Science*. **30** (2): 77–82. doi:10.1002/asi.4630300203. ISSN 1097-4571. Retrieved 2022-05-15.
- Buckland, Michael K. (1991). « Information as thing ». *Journal of the American Society for Information Science*. **42** (5): 351–360. doi:10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3. ISSN 1097-4571. Retrieved 2022-03-22.
- Star, Susan Leigh; Ruhleder, Karen (1996-03-01). « Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces ». *Information Systems Research*. **7** (1): 111–134. doi:10.1287/isre.7.1.111. ISSN 1047-7047. Retrieved 2021-12-22.
- Brase, Jan (2004). « Using Digital Library Techniques – Registration of Scientific Primary Data ». In Rachel Heery; Liz Lyon (eds.). *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. pp. 488–494. doi:10.1007/978-3-540-30230-8_44. ISBN 978-3-540-30230-8.
- Barateiro, José; Antunes, Gonçalo; Cabral, Manuel; Borbinha, José; Rodrigues, Rodrigo (2008). « Digital Preservation of Scientific Data ». In Birte Christensen-Dalsgaard; Donatella Castelli; Bolette Ammitzbøll Jurik; Joan Lippincott (eds.). *Research and Advanced Technology for Digital Libraries*. Vol. 5173. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 388–391. doi:10.1007/978-3-540-87599-4_41. ISBN 978-3-540-87598-7. Retrieved 2022-06-21.
- Schofield, Paul N.; Bubela, Tania; Weaver, Thomas; Portilla, Lili; Brown, Stephen D.; Hancock, John M.; Einhorn, David; Tocchini-Valentini, Glauco; Hrabe de Angelis, Martin; Rosenthal, Nadia (2009-09-10). « Post-publication sharing of data and tools ». *Nature*. **461** (7261): 171–173. Bibcode:2009Natur.461..171.. doi:10.1038/461171a. ISSN 0028-0836. PMC 6711854. PMID 19741686.
- Korsmo, F. L. (2010). « The Origins and Principles of the World Data Center System ». *Data Science Journal*. **8**: –55–IGY65. doi:10.2481/dsj.SS_IGY-011.

- Edwards, Paul N.; Mayernik, Matthew S.; Batcheller, Archer L.; Bowker, Geoffrey C.; Borgman, Christine L. (2011-10-01). « Science friction: Data, metadata, and collaboration ». *Social Studies of Science*. **41** (5): 667–690. doi:10.1177/0306312711413314. ISSN 0306-3127. PMID 22164720. S2CID 33973392.
- Tenopir, Carol; Allard, Suzie; Douglass, Kimberly; Aydinoglu, Arsev Umur; Wu, Lei; Read, Eleanor; Manoff, Maribeth; Frame, Mike (2011). « Data Sharing by Scientists: Practices and Perceptions ». *PLOS ONE*. **6** (6): –21101. Bibcode:2011PLoSO…621101T. doi:10.1371/journal.pone.0021101. ISSN 1932-6203. PMC 3126798. PMID 21738610.
- Chavan, Vishwas; Penev, Lyubomir (2011-12-15). « The data paper: a mechanism to incentivize data publishing in biodiversity science ». *BMC Bioinformatics*. **12** (Suppl 15): –2. doi:10.1186/1471-2105-12-S15-S2. ISSN 1471-2105. PMC 3287445. PMID 22373175.
- Campbell-Kelly, Martin; Garcia-Swartz, Daniel D (2013). « The History of the Internet: The Missing Narratives ». *Journal of Information Technology*. **28** (1): 18–33. doi:10.1057/jit.2013.4. ISSN 0268-3962. S2CID 41013. Retrieved 2022-01-04.
- Dacos, Marin (2013). « Cyberclio : vers une cyberinfrastructure au cœur de la discipline historique ». In Frédéric Clavert, Serge Noiret (ed.). *L'histoire contemporaine à l'ère contemporain* (Peter Lang ed.). Berne. pp. 29–41.
- Wallis, Jillian C.; Rolando, Elizabeth; Borgman, Christine L. (2013). « If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology ». *PLOS ONE*. **8** (7): –67332. Bibcode:2013PLoSO…867332W. doi:10.1371/journal.pone.0067332. ISSN 1932-6203. PMC 3720779. PMID 23935830.
- Vines, Timothy H.; Albert, Arianne Y. K.; Andrew, Rose L.; Débarre, Florence; Bock, Dan G.; Franklin, Michelle T.; Gilbert, Kimberly J.; Moore, Jean-Sébastien; Renaut, Sébastien; Rennison, Diana J. (2014-01-06). « The Availability of Research Data Declines Rapidly with Article Age ». *Current Biology*. **24** (1): 94–97. doi:10.1016/j.cub.2013.11.014. ISSN 0960-9822. PMID 24361065. S2CID 7799662. Retrieved 2022-09-11.
- Crosas, Mercè (2014-05-26). « The Evolution of Data Citation: From Principles to Implementation ». *IASSIST Quarterly*. **37** (1–4): 62. doi:10.29173/iq504. ISSN 0739-1137. Retrieved 2022-05-15.
- Tenopir, Carol; Dalton, Elizabeth D.; Allard, Suzie; Frame, Mike; Pjesivac, Ivanka; Birch, Ben; Pollock, Danielle; Dorsett, Kristina (2015). « Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide ». *PLOS ONE*. **10** (8): –0134826. Bibcode:2015PLoSO..1034826T. doi:10.1371/journal.pone.0134826. ISSN 1932-6203. PMC 4550246. PMID 26308551.
- Shankar, Kalpana; Eschenfelder, Kristin R.; Downey, Greg (2016-05-13). « Studying the History of Social Science Data Archives as Knowledge Infrastructure ». *Science*

Technology Studies. **29** (2): 62–73. doi:10.23987/sts.55691. ISSN 2243-4690. Retrieved 2021-12-23.

- Neylon, Cameron; Chan, Leslie (2016-04-18). « Exploring the opportunities and challenges of implementing open research strategies within development institutions ». *Research Ideas and Outcomes*. **2**: –8880. doi:10.3897/rio.2.e8880. ISSN 2367-7163. Retrieved 2021-11-01.
- Schmidt, Birgit; Gemeinholzer, Birgit; Treloar, Andrew (2016-01-15). « Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey ». *PLOS ONE*. **11** (1): –0146695. Bibcode:2016PLoSO..1146695S. doi:10.1371/journal.pone.0146695. ISSN 1932-6203. PMC 4714918. PMID 26771577.
- Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; Santos, Luiz Bonino da Silva; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J. G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; Hoen, Peter A. C. 't; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; Schaik, Rene van; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; Lei, Johan van der; Mulligen, Erik van; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend (2016). « The FAIR Guiding Principles for scientific data management and stewardship ». *Scientific Data*. **3**: 160018. Bibcode:2016NatSD…360018W. doi:10.1038/sdata.2016.18. PMC 4792175. PMID 26978244.
- Lyon, Liz; Jeng, Wei; Mattern, Eleanor (2017-09-16). « Research Transparency: A Preliminary Study of Disciplinary Conceptualisation, Drivers, Tools and Support Services ». *International Journal of Digital Curation*. **12** (1): 46–64. doi:10.2218/ijdc.v12i1.530. ISSN 1746-8256. Retrieved 2022-06-10.
- Witkowski, Tomasz (2017). « A Scientist Pushes Psychology Journals toward Open Data ». *Skeptical Inquirer*. **41** (4): 6–7. Archived from the original on 2018-09-15. although some scientists now agree that doing so could help prevent future retractions of scientific manuscripts.
- Besançon, Lonni; Peiffer-Smadja, Nathan; Segalas, Corentin; Jiang, Haiting; Masuzzo, Paola; Smout, Cooper; Billy, Eric; Deforet, Maxime; Leyrat, Clémence (2020). « Open Science Saves Lives: Lessons from the COVID-19 Pandemic ». *BMC Medical Research Methodology*. **21** (1): 117. doi:10.1186/s12874-021-01304-y. PMC 8179078. PMID 34090351.
- Rosenberg, Daniel (2018-11-01). « Data as Word ». *Historical Studies in the Natural Sciences*. **48** (5): 557–567. doi:10.1525/hsns.2018.48.5.557. hdl:21.11116/0000-0002-C567-C. ISSN 1939-1811. S2CID 149765492. Retrieved 2022-03-21.

- Joseph, Heather (2018-09-05). « Securing community-controlled infrastructure: SPARC's plan of action ». *College Research Libraries News*. **79** (8): 426. doi:10.5860/crln.79.8.426. S2CID 116057034.
- Federer, Lisa M.; Belter, Christopher W.; Joubert, Douglas J.; Livinski, Alicia; Lu, Ya-Ling; Snyders, Lissa N.; Thompson, Holly (2018-05-02). « Data sharing in PLOS ONE: An analysis of Data Availability Statements ». *PLOS ONE*. **13** (5): –0194768. Bibcode:2018PLoSO..1394768F. doi:10.1371/journal.pone.0194768. ISSN 1932-6203. PMC 5931451. PMID 29719004.
- Ross-Hellauer, Tony; Schmidt, Birgit; Kramer, Bianca (2018). « Are funder Open Access platforms a good idea? ». *SAGE Open*. **8** (4): 2158244018816717. doi:10.1177/2158244018816717. S2CID 220987901.
- Neylon, Cameron (2017-12-27). « Sustaining Scholarly Infrastructures through Collective Action: The Lessons that Olson can Teach us ». *KULA: Knowledge Creation, Dissemination, and Preservation Studies*. **1**: 3. doi:10.5334/kula.7. ISSN 2398-4112. Retrieved 2022-01-09.
- Allen, Liz; O'Connell, Alison; Kiermer, Veronique (2019). « How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship ». *Learned Publishing*. **32** (1): 71–74. doi:10.1002/leap.1210. ISSN 1741-4857. S2CID 67868432. Retrieved 2022-05-14.
- Smale, Nicholas Andrew; Unsworth, Kathryn; Denyer, Gareth; Magatova, Elise; Barr, Daniel (2020-01-01). « A Review of the History, Advocacy and Efficacy of Data Management Plans ». *International Journal of Digital Curation*. **15** (1): 30. doi:10.2218/ijdc.v15i1.525. ISSN 1746-8256. Retrieved 2022-06-21.
- Tenopir, Carol; Rice, Natalie M.; Allard, Suzie; Baird, Lynn; Borycz, Josh; Christian, Lisa; Grant, Bruce; Olendorf, Robert; Sandusky, Robert J. (2020-03-11). « Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide ». *PLOS ONE*. **15** (3): –0229003. Bibcode:2020PLoSO..1529003T. doi:10.1371/journal.pone.0229003. ISSN 1932-6203. PMC 7065823. PMID 32160189.
- van Reisen, Mirjam; Stokmans, Mia; Basajja, Mariam; Ong'ayo, Antony Otieno; Kirkpatrick, Christine; Mons, Barend (2020-01-01). « Towards the Tipping Point for FAIR Implementation ». *Data Intelligence*. **2** (1–2): 264–275. doi:10.1162/dint_a_00049. ISSN 2641-435X. S2CID 207828428.
- Colavizza, Giovanni; Hrynaszkiewicz, Iain; Staden, Isla; Whitaker, Kirstie; McGillivray, Barbara (2020-04-22). « The citation advantage of linking publications to research data ». *PLOS ONE*. **15** (4): –0230416. arXiv:1907.02565. Bibcode:2020PLoSO..1530416C. doi:10.1371/journal.pone.0230416. ISSN 1932-6203. PMC 7176083. PMID 32320428.
- Kerber, Wolfgang (2021). « Specifying and Assigning « Bundles of Rights » on Data: An Economic Perspective ». *SSRN Electronic Journal*. doi:10.2139/ssrn.3847620.

hdl:10419/234876. ISSN 1556-5068. S2CID 235457824. Retrieved 2022-05-14.

- Tedersoo, Leho; Küngas, Rainer; Oras, Ester; Köster, Kajar; Eenmaa, Helen; Leijen, Äli; Pedaste, Margus; Raju, Marju; Astapova, Anastasiya; Lukner, Heli; Kogermann, Karin; Sepp, Tuul (2021-07-27). « Data sharing practices and data availability upon request differ across scientific disciplines ». *Scientific Data*. **8** (1): 192. Bibcode:2021NatSD…8..192T. doi:10.1038/s41597-021-00981-0. ISSN 2052-4463. PMC 8381906. PMID 34315906.
- Fecher, Benedikt; Kahn, Rebecca; Sokolovska, Nataliia; Völker, Teresa; Nebe, Philip (2021-08-01). « Making a Research Infrastructure: Conditions and Strategies to Transform a Service into an Infrastructure ». *Science and Public Policy*. **48** (4): 499–507. doi:10.1093/scipol/scab026. ISSN 0302-3427. Retrieved 2021-12-22.
- Pujol Priego, Laia; Wareham, Jonathan; Romasanta, Angelo Kenneth S. (2022-02-07). « The puzzle of sharing scientific data ». *Industry and Innovation*. **29** (2): 219–250. doi:10.1080/13662716.2022.2033178. ISSN 1366-2716. S2CID 246795400. Retrieved 2022-06-18.
- Federer, Lisa M. (2022-08-24). « Long-term availability of data associated with articles in PLOS ONE ». *PLOS ONE*. **17** (8): –0272845. doi:10.1371/journal.pone.0272845. ISSN 1932-6203. PMC 9401135. PMID 36001577.
- Science Staff (2011-02-11). « Challenges and Opportunities ». *Science*. **331** (6018): 692–693. Bibcode:2011Sci…331..692.. doi:10.1126/science.331.6018.692. PMID 21311002. S2CID 109422723.

Thèses et ouvrages

- Bourne, Charles P.; Hahn, Trudi Bellardo (2003-08-01). *A History of Online Information Services, 1963-1976*. MIT Press. ISBN 978-0-262-26175-3.
- Borgman, Christine L. (2007-10-12). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA, USA: MIT Press. ISBN 978-0-262-02619-2.
- Berners-Lee, Tim; Fischetti, Mark (2008). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Paw Prints. ISBN 978-1-4395-0036-1.
- Bygrave, Lee A.; Bing, Jon (2009-01-22). *Internet Governance: Infrastructure and Institutions*. OUP Oxford. ISBN 978-0-19-956113-1.
- Edwards, Paul N. (2010-03-12). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Infrastructures. Cambridge, MA, USA: MIT Press. ISBN 978-0-262-01392-5.
- National Research Council (2012). Paul E. Uhlir (ed.). *For Attribution: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, DC: The National Academies Press.

_____._____
ISBN 978-0-309-26728-1. Retrieved 2022-03-22.
- Gaillard, Rémi (2014). *De l'Open data à l'Open research data: quelle(s) politique(s) pour les données de recherche ?* (Thesis). ENSSIB.
- Hogan, A. (2014-04-09). *Reasoning Techniques for the Web of Data*. IOS Press. ISBN 978-1-61499-383-4.
- Borgman, Christine L. (2015-01-02). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA, USA: MIT Press. ISBN 978-0-262-02856-1.
- Briney, Kristin (2015-09-01). *Data Management for Researchers: Organize, maintain and share your data for research success*. Pelagic Publishing Ltd. ISBN 978-1-78427-013-1.
- Regazzi, John J. (2015-02-12). *Scholarly Communications: A History from Content as King to Content as Kingmaker*. Rowman Littlefield. ISBN 978-0-8108-9088-6.
- Cox, Andrew; Verbaan, Eddy (2018-05-11). *Exploring Research Data Management*. Facet Publishing. ISBN 978-1-78330-280-2.
- Tim Davies; Stephen B. Walker; Mor Rubinstein; Fernando Perini, eds. (2019). *The State of Open Data: Histories and Horizons*. African Minds. Retrieved 2022-09-11.
- Lipton, Vera (2020-01-22). *Open Scientific Data: Why Choosing and Reusing the RIGHT DATA Matters*. BoD – Books on Demand. ISBN 978-1-83880-984-3. [*unreliable source?*]
- Tibor, Koltay (2021-10-31). *Research Data Management and Data Literacies*. Chandos Publishing. ISBN 978-0-323-86002-4.

Autres sources

- Neylon, Cameron; Bilder, Geoffrey; Lin, Jennifer (2015). « Principles for Open Scholarly Infrastructures ». *Science in the open*. Retrieved 2021-11-01.

Liens externes

- Research Data Canada
- Open Data In Science article (P Murray-Rust)
- Open Data about monitoring of deforestation in the Brazilian Amazon Rainforest
- OpenWetWare
- Open ConnectomeProject
- LinkedScience.org
- Collective Mind Repository for computer engineering

Ressources

Photographie du stockage des cartes perforées au US National Weather Records Center à Asheville (début des années 1960). Selon Paul Edwards, il y avait tellement de cartes perforées à l'époque qu'elles étaient stockées jusqu'à l'entrée de la pièce
Domaine public américain
Image reproduite dans Paul Edwards, A Vast Machine : Computer Models, Climate Data, and the Politics of Global Warming, p. 102