4ème journée pour la science ouverte au CNRS

LA SCIENCE OUVERTE ET LES DONNÉES DE LA RECHERCHE

30 novembre 2022

Recueil des interventions

Impression
CNRS/IFSeM/Secteur de l'imprimé

Conception
Sarah Landel, DircCom

Mise en page
Isabelle Debano, CNRS DDOR
WL, IFSeM/Secteur de l'imprimé

Octobre 2023

Photo de couverture : bibliothèque de stockage à bandes magnétiques du Centre de calcul de l'Institut de physique nucléaire et de physique des particules (CC-IN2P3). Cette infrastructure de recherche conçoit et opère un système de stockage de masse et de traitement de grandes quantités de données. © Cyril Fresillon / CC IN2P3 / CNRS Photothèque

SOMMAIRE

LA SCIENCE OUVERTE ET LES DONNÉES DE LA RECHERCHE

Alain Schuhl,
Directeur général délégué à la science, CNRS

Cette journée est consacrée aux données de la recherche. L'année dernière, elle se concentrait sur l'évaluation individuelle des chercheurs. Nous avons beaucoup de travail à accomplir pour permettre à toutes les communautés scientifiques d'avoir accès à des stockages de données qui permettent de retrouver, d'utiliser et d'interconnecter facilement les données. Une partie de la journée sera ainsi consacrée aux entrepôts de données, puis une autre aux données sensibles de la recherche. Nous terminerons la journée par une table ronde autour de la législation sur le partage des données de la recherche. Isabelle Blanc nous présentera l'action menée par le MESR sur l'entrepôt national de données, et les actions menées au niveau européen sur l'*European Open Science Cloud* (EOSC), qui est la déclinaison européenne de ce que nous faisons au niveau national. Nous voyons bien que le CNRS, qui est un acteur essentiel de l'enseignement supérieur et de la recherche, travaille aussi bien avec le ministère qu'au niveau européen.

Mon objectif est le suivant : 100 % d'accès ouvert pour les publications scientifiques.

Nous avons réalisé de nombreux recensements. Sur le plan du périmètre des unités mixtes de recherche (UMR), nous sommes passés de 50 % à 77 % de publications en accès ouvert depuis 2020, sans compter l'action que nous avons menée sur les comptes rendus annuels d'activité des chercheurs (CRAC). Cette action ne concerne malheureusement que les chercheurs du CNRS. En imposant le fait que les publications dans le CRAC soient extraites de HAL, nous avons augmenté considérablement le nombre de dépôts dans HAL, donc en accès ouvert. Pour les chercheurs du CNRS, sur l'année 2021, le taux d'ouverture des publications s'est porté à 94 %. Sans avoir mené d'action spécifique en la matière, le nombre d'identifiants HAL s'est établi à 75 %. Une action volontaire et forte a donc livré des résultats probants. Nous avons confié au CCSD des moyens financiers significatifs, à la hauteur des nouveaux enjeux. Un nombre notoire de publications sont dans HAL et non dans le *Web of Science*. Nous avons notamment travaillé sur l'exemple des sciences de l'information. Il apparaît que 4 000 publications sont dans le *Web of Science* et 8 000 dans HAL. Nous travaillons avec le CCSD pour le doter d'outils qui nous permettront de nous passer du *Web of Science* et d'avoir accès à toutes les publications.

Nous avons avec nos camarades allemands un débat, depuis quelque temps : le modèle de la science ouverte allemand réside dans les *Article processing charges* (APC)[1]. Nous estimons que ce système n'est pas vertueux, en ce qu'il favorise les personnes et les pays qui ont les moyens de payer. Nous travaillons actuellement avec l'Amérique du Sud, qui ne peut publier, faute de pouvoir payer les APC. Les communautés qui paient le plus d'APC sont en l'occurrence la biologie et la chimie. Dans un modèle 100 % APC, le coût des publications du CNRS passerait de 14 millions à 40 millions d'euros par an. Ce modèle, en plus d'être inégalitaire, n'est donc pas viable financièrement. Nous devons absolument trouver des solutions alternatives. Tel est le travail que nous avons conduit pour soutenir la bibliodiversité, qui consiste à offrir aux communautés scientifiques des moyens de publier sans payer d'APC. La publication a bien entendu un coût, et nous sommes très attachés à l'évaluation par les pairs. Si on considère le budget du CNRS, soit 3,6 milliards d'euros, divisé par le nombre de publications par an, cela donne un coût de publication de 80 000 à 100 000 euros. Le processus de publication lui-même n'est pas un scandale ; le problème réside dans le fait d'être à la merci de grandes industries qui augmentent considérablement le coût des APC uniquement sur la réputation des publications, et non pas sur le travail effectué.

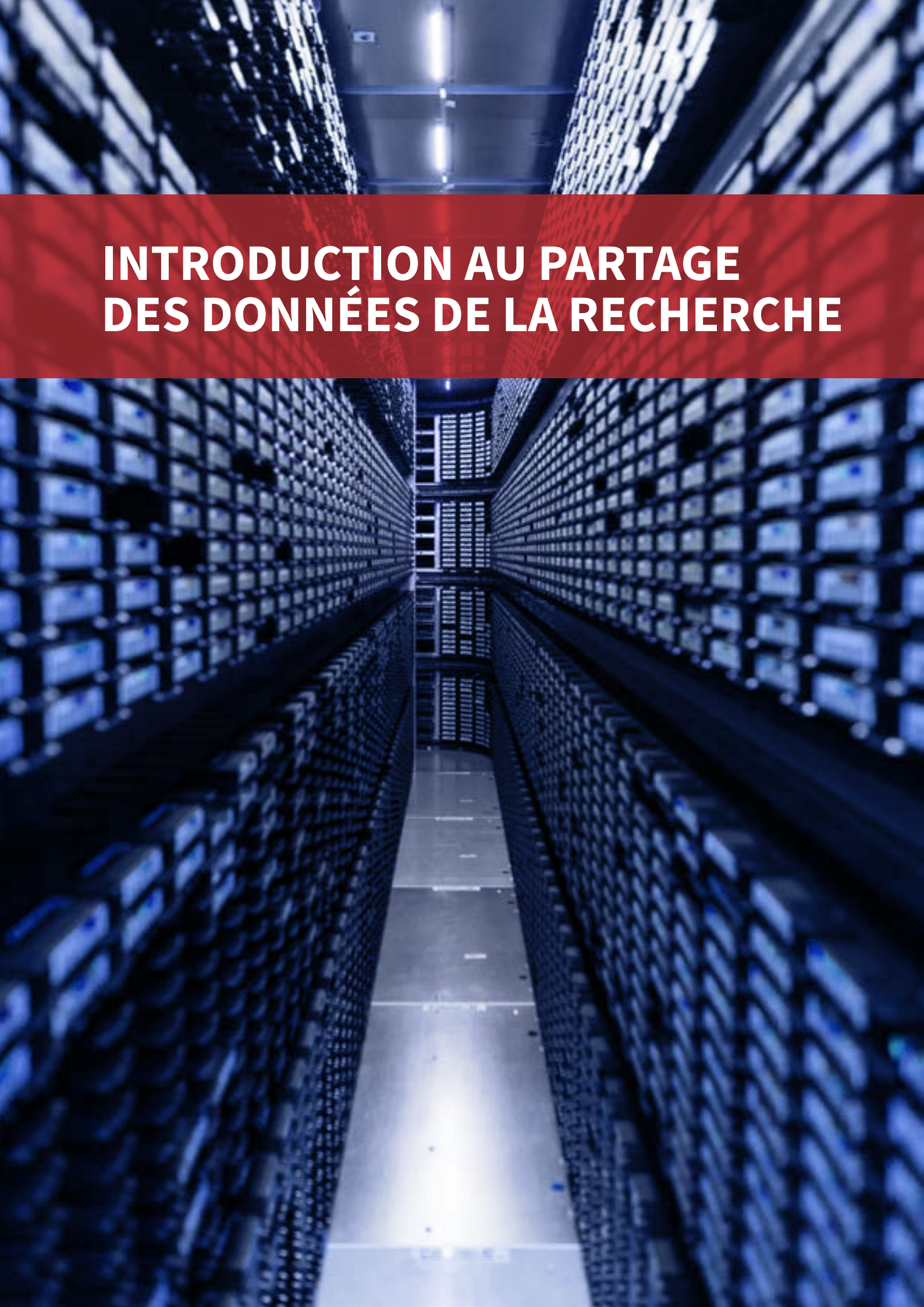
Le modèle que nous nous efforçons de développer est le modèle Diamant, qui consiste à ne pas payer d'abonnement ou d'APC, mais à construire des plateformes permettant de publier, avec pour seul coût celui supporté par les structures. Dans le cas des mathématiques, par exemple, le centre Mersenne[2], à Grenoble, a fêté ses cinq ans cette année. Durant cette période, la communauté s'est mobilisée pour faire en sorte que des plateformes permettent aux chercheurs de publier de très bons articles, avec une structure financée par le CNRS, l'université et les organismes, ce qui permet de contrôler le coût. Le modèle Diamant repose ainsi sur des plateformes supportées par des financements publics, qui permettent de publier en accès ouvert sans payer d'APC.

J'ai participé à une interview[3] accessible en ligne dans laquelle j'évoquais ce sujet. Une autre sera prochainement publiée. Nous ne pouvons pas dire aux chercheurs de ne pas payer d'APC, parce que certaines communautés n'ont pas de solution alternative. Elles doivent donc être en mesure de s'organiser. De nombreuses solutions existent, comme les PCI (peer community in)[4]. Pour les mathématiques, le centre Mersenne, et pour les SHS, OpenEdition[5], réalisent aussi un travail formidable. Ces plateformes doivent être développées. En 2018, nous avons mis un terme aux abonnements à Springer, et décidé de placer ces économies dans l'aide aux plateformes. Entre 2019 et 2022, le CNRS a injecté deux millions d'euros pour soutenir ces plateformes. Nous sommes parvenus à transférer la publication des Comptes rendus de l'Académie des Sciences[6], publiés jusqu'à présent par Elsevier, au centre Mersenne. Sur les sept catégories, six ont été transférées, une catégorie reste publiée directement par l'Académie des Sciences. Il s'est agi d'un réel choc de culture. En matière de publications, nous pensons immédiatement aux grandes publications prestigieuses, mais nous nous battons contre des éditeurs comme Elsevier, qui sont très coûteux, et vis-à-vis desquels, bien souvent, il suffit de payer pour être publié. Il s'agit de travailler avec les communautés sur ce sujet, sur lequel le Comité de Pilotage de la DDOR (Direction des données ouvertes de la recherche) travaille pour offrir à chacune des communautés des solutions alternatives. Nous ne pourrons cependant rien faire si les communautés ne se mobilisent pas. Nous voyons en l'espèce que la biologie est en train de se mobiliser. Les coûts des APC augmentent tant que nous n'avons pas d'autre solution.

Nous entendons souvent que la publication d'un article doit s'accompagner d'une cession des droits à l'éditeur. Cela n'est pas vrai. Une politique de non-cession des droits est développée au niveau européen et au niveau français. Nous disposons de tous les outils juridiques pour faire en sorte de ne pas avoir besoin de céder les droits aux éditeurs. Un article dans CNRS Info[7] expliquera tous les détails sur ce sujet. Une stratégie européenne est actuellement défendue par les agences de financement européennes. Les chercheurs doivent absolument résister sur le sujet. Si nous ne le faisons pas collectivement, nous ne parviendrons pas à résoudre ce problème. La stratégie permet à tous les auteurs de conserver leurs droits sur leur article, en appliquant eux-mêmes une licence spécifique sur les pages de leur manuscrit, qui signifie que les droits ne sont pas cédés à l'éditeur.

Nous avons également beaucoup travaillé sur l'évaluation individuelle des chercheurs. Celle-ci est capitale et permettra de gagner le combat de la science ouverte ou non. Nous devons avoir une évaluation qualitative, et non purement quantitative. Nous avons travaillé avec la CPCN (Conférence des Présidents des Sections du Comité national)[8] pour changer le dossier d'évaluation sur deux aspects : d'abord, le métier de chercheur recouvre la mission de produire des connaissances, mais également de les valoriser, participer à l'enseignement initial et tout au long de la vie, participer à l'expertise scientifique, etc. Depuis deux ans, nous avons proposé que les chercheurs, dans les dossiers d'évaluation, déclarent ce qu'ils ont réalisé dans la période précédente. Les chercheurs doivent pouvoir, dans une partie déclarative, expliquer en quoi leurs travaux ont fait avancer la science. Nous avons en outre limité à 10 le nombre de productions scientifiques, qui incluent des publications mais aussi des livres, des logiciels, etc. Il appartient donc aux chercheurs de décider parmi leurs productions scientifiques celles qu'ils estiment être les plus importantes et sur lesquelles ils souhaitent être évalués. Il s'agit de faire en sorte que le chercheur valorise ses réalisations, les explique et précise qu'il souhaite être évalué dessus. En limitant le nombre de publications à 10, l'objectif était de permettre aux membres des sections du Comité national de disposer du temps nécessaire pour apprécier le contenu de ces articles. S'agissant de la sélection des chercheurs, les sections font des propositions et la direction du CNRS prend la décision. Si des sections utilisent des critères qui ne sont pas compatibles avec la politique du CNRS, aucun chercheur ne sera recruté dans ces sections. Sur la question de la compatibilité des dossiers français avec le reste du monde, l'an dernier, dans le cadre de la PFUE et de l'appel de Paris sur l'évaluation[9], le CNRS a été un des premiers signataires de COARA[10], une alliance créée au niveau européen, dont le board sera prochainement constitué. L'une des candidates est la directrice de la DDOR, afin de faire entendre la voix du CNRS.

Sur le sujet des données de la recherche, nous avons lancé un plan des données de la recherche, fin 2020. Nous avons également créé la DDOR, afin de faire en sorte que les données soient le cœur de la recherche ouverte, en incluant les publications scientifiques, les données issues des recherches et les infrastructures de recherche. J'ai eu le plaisir d'assister au COPIL de la DDOR, qui manque aujourd'hui de temps pour étudier pleinement tous les problèmes à traiter. S'agissant d'EOSC, Suzanne Dumouchel représente à la fois le CNRS et l'ESR français au sein du board. Nous avons pu le faire grâce à nos réseaux internationaux et au G6, qui correspond au groupement de six grands organismes de recherche pluridisciplinaires européens. Je laisse la parole à Sylvie Rousset pour présenter ce sujet des données de la recherche. Antoine Petit et moi-même sommes très attentifs à cette politique de science ouverte, et consacrerons tous les moyens nécessaires pour avancer dans cette direction.



INTRODUCTION AU PARTAGE DES DONNÉES DE LA RECHERCHE

Sylvie Rousset,
Directrice, CNRS Direction des Données Ouvertes de la Recherche (DDOR)

Faut-il ouvrir toutes les données de la recherche ? Une donnée doit être ouverte ou protégée. L'ouverture des données s'entend selon l'expression « ouvert autant que possible, fermé autant que nécessaire ». Toutes les données de la recherche n'ont pas vocation à être ouvertes ou divulguées. Il existe des exceptions évidentes telles que les données spécifiques à caractère confidentiel, que cela soit du fait de leur caractère personnel, pour des raisons de concurrence industrielle ou pour des intérêts fondamentaux ou réglementaires des États. La décision d'ouverture ou de protection des données de la recherche doit être prise avec les services compétents du CNRS :

- les Services Partenariat Valorisation pour la propriété intellectuelle,
- la Délégation à la protection des données pour les données à caractère personnel,
- la Direction de la sûreté pour les questions relatives à la souveraineté.

Les données de la recherche correspondent à l'information utilisée pour faire une démonstration scientifique, avec tous les formats possibles et un écosystème incluant les infrastructures numériques, les *big data*, les services et les principes. Le cycle de vie de la donnée est fondamental ; les données sont produites, analysées, conservées, mais ont aussi vocation à être réutilisées et réinjectées dans le système, afin de produire de nouvelles connaissances. Ce cycle de vie de la donnée nous amène ainsi à vouloir bien traiter nos données et à les partager. Nos motivations sont d'abord scientifiques. Par ailleurs, un cadre légal d'ouverture des données doit être satisfait. Enfin, nous recherchons la rationalisation et la mutualisation de nos infrastructures et de nos moyens humains.

Lorsque nous avons rédigé le plan des données de la recherche du CNRS[11] (novembre 2020), des discussions portaient sur le fait de savoir si le fait d'entreposer des données revenait simplement à les stocker. Telle n'est pas notre définition de l'*entrepôt des données* de la recherche, qui est un service en ligne permettant le dépôt, la description, la conservation, la recherche et la diffusion des jeux de données de la recherche. Nous utilisons également les notions de *stockage* des données de la recherche, de *data centre* (qui correspond au bâtiment qui héberge les moyens de stockage) et de *mésocentre* (qui correspond aux moyens de calcul).

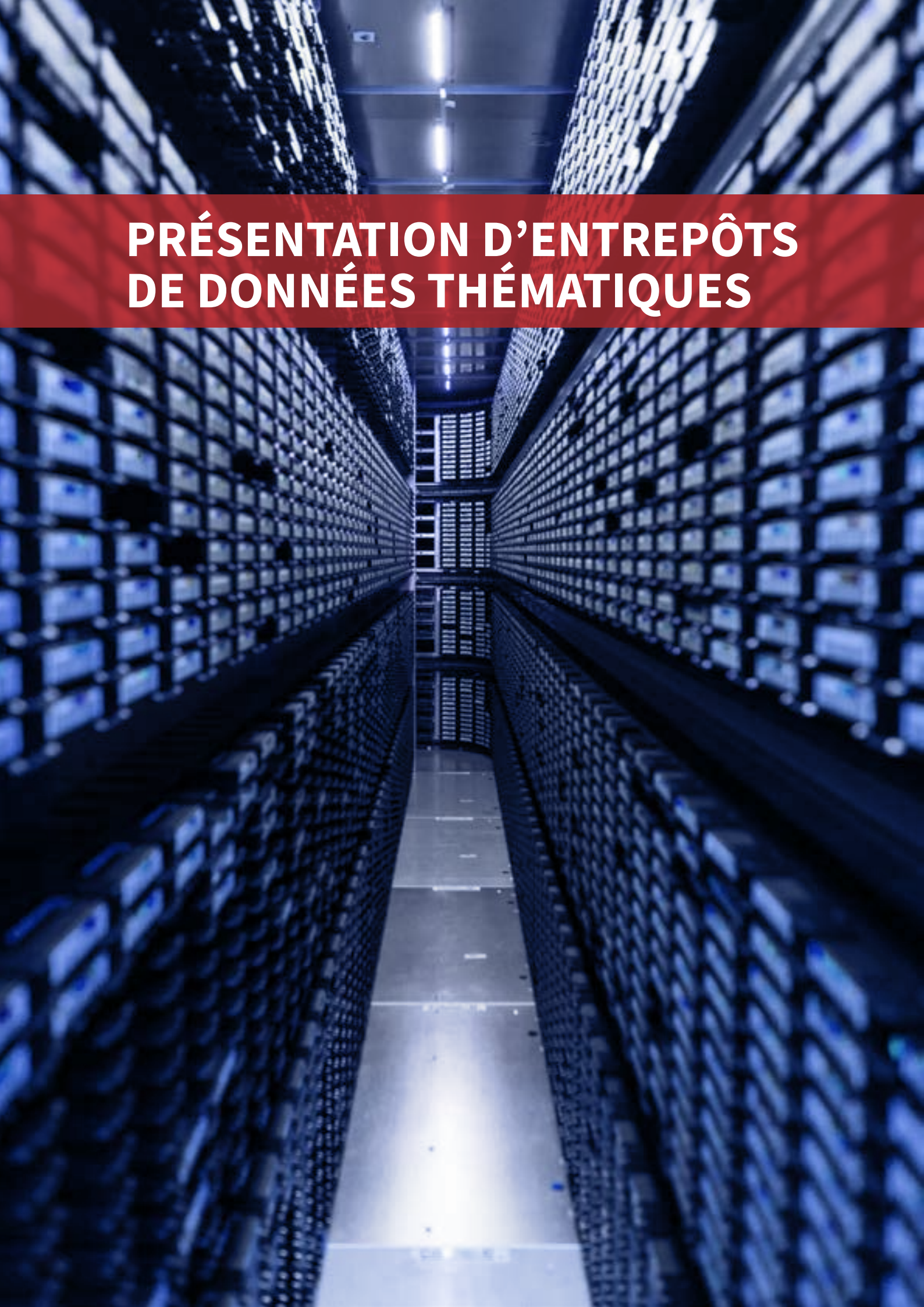
Nous avons lancé avec tous les instituts des cas d'usage, qui nous ont permis de discuter avec des communautés scientifiques restreintes. La DDOR et des spécialistes de l'INIST se sont ainsi réunis avec cinq équipes de cinq instituts différents, sur une année. Cela nous a permis de disposer d'un panorama de toutes les questions à aborder. Les questions politiques, en particulier, ont été très importantes. Les questions de la volumétrie et du stockage des données étaient elles aussi récurrentes.

En termes de réalisations, les services de l'INIST permettent aux chercheurs de traiter leurs données. Un groupe de travail piloté par Françoise Genova (CDS) et Paolo Lai (INIST) a permis de proposer un annuaire des services et entrepôts de données dont le CNRS est partie prenante : CNRS Données de la Rercherche[12]. Il permet également de naviguer sur une mappemonde. Nous réfléchissons déjà à la réalisation de ce type d'annuaire au niveau national.
S'agissant des outils et services proposés par l'INIST, trois sont particulièrement importants :

- l'aide pour les plans de gestion des données, via le service OPIDoR[13],
- le service en ligne de formation DoRANum[14] (avec neuf thématiques autour du cycle de vie de la donnée et des ressources pédagogiques librement accessibles),
- l'attribution de DOI pour les jeux de données via un lien avec DataCite[15]. La première assemblée générale de ce consortium français, qui rassemble plus de 175 utilisateurs s'est réunie en novembre 2022.

Enfin, concernant les grands chantiers en cours, nous souhaitons pouvoir créer un espace institutionnel CNRS[16] sur la plateforme Recherche Data Gouv[17], accompagner les scientifiques au dépôt de leurs jeux de données, créer des référentiels thématiques de métadonnées pour les communautés qui n'en disposent pas encore, traiter la problématique du stockage, ou encore répondre aux questions juridiques, qui ressortaient fortement des cas d'usage que nous avons étudiés.

Je vous propose de lancer la première session de cette journée. Nous savons que certaines communautés sont depuis longtemps très bien organisées pour gérer et partager leurs données. Nous souhaitons exposer les succès de l'organisation de ces communautés, à travers quatre exemples de structures labellisées comme centres thématiques de référence sur la plateforme du ministère et qui sont très ancrées au sein du CNRS.



PRÉSENTATION D'ENTREPÔTS DE DONNÉES THÉMATIQUES

1. Centre de Données astronomiques de Strasbourg (CDS)

Françoise Genova et Mark Allen,
CNRS - Observatoire de Strasbourg

Mark Allen est le directeur du Centre de Donnes astronomiques de Strasbourg[18] depuis septembre 2015, il a succédé à Françoise Genova qui l'avait dirigé de 1995 à 2015. L'objet de notre exposé est de réfléchir sur ce qu'il a fallu faire et qu'il faut continuer à faire pour développer un service permettant aux chercheurs d'accéder à des données sur la longue durée. Nous avons effectivement fêté les 50 ans du CDS en janvier 2022. Le Centre de Données astronomiques de Strasbourg a été créé en 1972 par l'Institut National d'Astronomie et de Géophysique (INAG), qui est maintenant l'INSU du CNRS, avec l'Université Louis Pasteur, qui est l'Université de Strasbourg. La mission du CDS n'a pas changé depuis 1972 : collecter les données utiles sur les objets astronomiques sous forme électronique, les améliorer en les évaluant de façon critique et en les combinant, distribuer les résultats à la communauté internationale et conduire des recherches en utilisant les données. Par rapport aux autres centres thématiques, le CDS a depuis l'origine un rôle international. Les mots clés en 1972 guident toujours l'action du CDS : données sous forme électronique, curation et qualité des données, et valeur ajoutée sur ces données. L'objectif de servir la recherche est également essentiel : il ne s'agit pas seulement d'accumuler des données de recherche, il faut qu'elles soient utiles aux chercheurs. La gouvernance est identifiée et effective. L'instance de pilotage stratégique et scientifique du CDS est son Conseil Scientifique. Il est intéressant de constater que depuis sa création au démarrage du Centre de Données, le Conseil Scientifique comprend six membres français et six membres étrangers. Les grandes organisations qui fournissent des moyens à la discipline, l'ESA, l'ESO, le CNES et la NASA, y sont représentées.

Le CDS est un des piliers de la pratique internationale de la science ouverte en astronomie. Il est inscrit sur la feuille de route nationale des infrastructures de recherche depuis la première version de celle-ci en 2008. Le CDS est une infrastructure de services à la donnée, qui est représentée dans le groupe de travail du CoSIN (Comité Services et Infrastructures Numérique) qui regroupe ces infrastructures. C'est également un centre de référence thématique de Recherche Data Gouv.

Le CDS est un des pionniers parmi les centres de données scientifiques, toutes disciplines confondues. Il a été pionnier du partage des données en astronomie, avec la base de données du satellite IUE (*International Ultraviolet Explorer,* qui a été opérationnel de 1978 à 1996). L'astronomie est aussi une des disciplines pionnières de la science ouverte. Nous avons identifié très tôt que le partage des données est essentiel à notre pratique scientifique parce que nous étudions un grand nombre d'objets, nous réalisons des statistiques sur les objets et nous utilisons de façon conjointe des données obtenues par des instruments différents pour comprendre les phénomènes physiques à l'œuvre dans l'Univers. Les objets sont souvent variables à différentes échelles, ce qui exige de pouvoir conserver et réutiliser les données sur le long terme.

Les services que nous avons inventés et que nous maintenons sur la durée sont utilisés tous les jours par la communauté dans son travail de recherche – on compte deux millions de requêtes par jour sur ces services. Nous avons également joué un rôle significatif dans la définition des standards d'échange et d'interopérabilité des données.

Nous réalisons ces tâches avec une équipe intégrée, qui regroupe les différents profils des métiers de la donnée : chercheurs, documentalistes (*data stewards*), ingénieurs informaticiens. Nous sommes en première ligne de l'évolution nécessaire de l'évaluation des chercheurs et des profils des personnels techniques dans le cadre de la science ouverte. Les astronomes assurent de la recherche, de l'enseignement et des tâches de services pour assurer la pertinence scientifique des services. Ils contribuent à la fois au CDS et à la communauté nationale et internationale. La polyvalence des documentalistes peut être soulignée, ainsi qu'une chaîne documentaire complexe, qui ajoute de la valeur aux données.

L'impératif catégorique étant de servir la recherche, le point de départ est constitué par les besoins scientifiques, qui structurent la réflexion, avec la qualité et la pertinence des services et des données. Nous devons aussi être au niveau technologique attendu par la communauté, afin de répondre à ses

►►►

besoins, qu'il faut savoir anticiper. Pour information, la première connexion française à internet a été mise en place pour une démonstration d'une base de données du CDS aux États-Unis, en 1998. Il est difficile mais indispensable de penser sur la longue durée, dans un contexte qui évolue constamment. Au-delà de la science et de la technologie, le contexte politique évolue lui aussi. Nous devons trouver notre place dans ce système en évolution.

Nous devons donc continuer à progresser sur la longue durée. La stratégie doit être explicite, évolutive et prendre en compte les différentes facettes du contexte. Il importe de créer, réunir et conserver les compétences, mais aussi de les faire évoluer. Nous devons par ailleurs continuer à construire et à tenir notre place dans le concert des nations. Nous fournissons des services essentiels à la communauté internationale. Un réseau de collaboration avec les acteurs majeurs (agences, journaux, base bibliographique de la discipline) s'est noué, avec une participation significative à l'*International Virtual Observatory Alliance*[19] qui définit les standards disciplinaires pour le partage des données. Enfin, il est indispensable d'obtenir le soutien sur le long terme de la gouvernance.

Pour le futur, le CDS doit continuer à renforcer le cœur de sa mission, la fourniture de services de données de référence à la communauté astronomique internationale et le *stewardship* (la curation) de ces données, et à participer au développement des standards internationaux qui permettent aux données et aux services d'être FAIR. Il lui faut aussi continuer à innover pour la prochaine génération de services de référence. Pour faire face à ces défis, et à l'augmentation du volume des données, il s'agit de recruter et former les data stewards et les informaticiens, d'améliorer continuellement les pipelines d'ingestion des données et les services, pour les rendre les plus efficaces et pertinents possible. Il faut passer au Peta/Exabyte, notamment pour le projet SKA[20] (*Square Kilometer Array*), qui sera structurant pour le futur, et s'inscrire pleinement dans la nouvelle ère de l'astronomie multi-messagers. Nous devons donc continuer d'innover. Une des évolutions en cours est l'intégration poussée des services du CDS dans les infrastructures de l'astronomie. Nous devons également maintenir notre capacité fondamentale à créer de nouveaux services, et trouver la bonne place pour le CDS dans l'EOSC.

Sur la durée, nous devons faire face à une charge de travail très lourde - qui continue d'augmenter - maintenir la confiance des utilisateurs via la pertinence des services par rapport à leurs besoins et un souci constant de la qualité du contenu et des services, assurer la veille technologie et la R&D comme partie intégrante du travail du CDS, entretenir le réseau de collaboration internationale avec tous les acteurs importants, et continuer à construire sur la collaboration et l'interopérabilité, malgré le risque de perdre en visibilité. Il est essentiel de continuer à nous assurer le soutien du CNRS/INSU et de l'Université, et de répondre au risque que tout le monde pense que tout va bien au CDS alors qu'il relève constamment des défis complexes, et enfin de préserver l'équipe du CDS sur la durée, en assurant la relève des générations successives.

2. Data Terra repository

Emmanuel Chaljub,
Data Terra - Directeur du pôle ForM@Ter

Je vous présente aujourd'hui l'infrastructure de recherche Form@Ter[21] puis le nouveau service d'entrepôt de données Data Terra[22] associé à Recherche Data Gouv.

En termes d'abord de contexte et d'enjeux, l'IR Data Terra s'inscrit dans un objectif d'appui à la recherche. L'objet qu'est la terre est un système complexe dynamique, avec de nombreux processus géophysiques et environnementaux à l'œuvre, qui opèrent à différentes échelles spatiales et temporelles, avec des interactions permanentes entre les différents compartiments de la terre solide, des surfaces continentales, de l'océan, de l'atmosphère et de l'anthroposphère. Pour comprendre ces processus, la recherche a besoin d'accéder à des données, de les croiser et de les analyser. Ces données sont complexes et hétérogènes, de par leur nature, leur niveau de transformation, leur fréquence d'acquisition, leur volume, de par les instituts et les organismes qui les financent. Elles sont d'origines multiples (satellites, observations de la Terre, *in situ*, campagnes temporaires d'observation, résultats d'expérimentation, sorties de modèles physiques et numériques). Certaines de ces données sont de plus en plus massives, avec par exemple des missions satellites de très haute résolution, de nouveaux capteurs, des réseaux très denses, des données de fibre optique ou encore des données venant de la science citoyenne. Toutes ces données sont issues de nombreux producteurs. L'objectif général de l'infrastructure de recherche est de développer un dispositif d'accès et de traitement de ces données multi-sources, ainsi que des services de haut niveau qui vont aider à comprendre et prévoir le fonctionnement et l'évolution du système Terre. L'objectif est donc de faciliter l'accès et l'utilisation des données et produits, de développer des services de visualisation et de traitement adaptés aux besoins, à l'accroissement de la volumétrie et aux avancées technologiques, de favoriser la mutualisation, l'interopérabilité, et de soutenir des approches multi et interdisciplinaires. L'IR Data Terra sert ainsi les communautés scientifiques et techniques, mais également les acteurs de l'action publique et de l'innovation. Sa mission est de mettre en œuvre une stratégie nationale, européenne et internationale.

Data Terra regroupe quatre pôles de données, correspondant aux quatre compartiments du système air : AERIS pour l'atmosphère, ForM@Ter pour la Terre solide, ODATIS pour l'océan, THEIA pour les surfaces continentales. Le pôle de données de biodiversité va rejoindre Data Terra d'ici 2025. Data Terra comptera 34 organismes et universités au lieu de 26 aujourd'hui. Ce périmètre inclut 30 centres de données et de services ou infrastructures de données, en particulier spatiales. Plus de 500 produits et services sont accessibles, pour plus de 15 000 utilisateurs.

S'agissant du paysage actuel des infrastructures, l'ensemble des infrastructures de recherche regroupe l'acquisition de données d'observation, d'expérimentation et de collection, ainsi que les infrastructures de recherche plus logistiques ou d'équipement transversal. L'ensemble de ces infrastructures contribue à alimenter Data Terra et le Pôle National de Données de Biodiversité[23] (PNDB), qui ont une nature d'e-infrastructure, depuis le stockage physique des données à l'entrepôt des données d'observation, qui a vocation à être certifié. La gestion des entrepôts de données s'opère au plus proche des services nationaux d'observation. Des centres de données et de services (CDS, CDOS) et des infrastructures de données et de services (IDS) fédèrent ces entrepôts et offrent des services de plus haut niveau. Ces centres de données sont eux-mêmes coordonnés par les pôles nationaux de données et de services.

Chacun des quatre pôles de données et de services de Data Terra alimente un catalogue de métadonnées qui expose l'ensemble des données des centres de données d'observation et de services, qui fédèrent eux-mêmes les entrepôts de données. Chaque pôle construit ainsi un catalogue qui alimente un catalogue fédéré de Data Terra, dont la colonne vertébrale terminologique est essentielle à la constitution du portail de la connaissance, développé par Data Terra. Les petits producteurs de données, quant à eux, ont des données de qualité, qui ont un potentiel de réutilisation, mais qui ne sont pas prises en compte dans les entrepôts de données ni les centres de données des pôles. Le pôle ODATIS, par exemple, propose un service d'entrepôts de données à destination de ces petits producteurs, SEANOE[24] (*SEA scieNtific Open data Edition*). Le service qui répond aux besoins d'entrepôt pour les données qualifiées d'orphelines est l'entrepôt de données Data Terra. L'entrepôt de données Data Terra est une des contributions à la science ouverte. Il a également, entre autres, la mission

►►►

de centre de référence thématique pour le Système Terre et Environnement. Data Terra est en outre une composante d'EOSC France, avec la volonté d'influer sur les services qui seront développés dans EOSC pour la communauté. Data Terra assure également une représentation française dans GO FAIR[25], une initiative internationale visant à mettre en œuvre les principes FAIR, avec le partage d'un secrétariat français avec la RDA.

L'entrepôt de données Data Terra est en construction. Il a commencé à être développé en 2021. Une équipe projet a été constituée avec des personnes de Data Terra, du BRGM[26] et d'une IR d'observation dans le domaine Terre solide. L'entrepôt est hébergé au BRGM et a été construit sur la base des besoins identifiés par les pôles : gestion des données produites dans les entités qui ne disposent pas de moyens pour les FAIRiser et les stocker, production de produits à valeur ajoutée issus de données d'observation mais qui ne sont pas restockées ni réutilisées. Les résultats d'une enquête conduite auprès de la communauté, après une réunion organisée par l'INSU avec l'ensemble des directions de l'Observatoire, sont apparus comme hétérogènes. La notion d'entrepôt, à l'époque, n'était pas claire pour tous. Il y a donc un besoin de stocker de façon simple. Pour réutiliser les données, une phase de qualification de la donnée est nécessaire, et elle suppose l'association de la communauté.

Il s'est par ailleurs agi d'analyser les solutions existantes et les retours d'expérience de la communauté. Ces éléments ont abouti au choix technique de la solution GeoNetwork[27]. En termes de fonctionnalités, l'entrepôt doit pouvoir accueillir des données géoréférencées, avec des métadonnées normalisées, qui permettent en particulier d'associer des vocabulaires, ce qui est fondamental pour la découverte et la réutilisation de ces données. En termes de volumétrie cible, les dépôts peuvent avoir une granularité différente, par fichier ou par jeu de données, avec des tailles allant de 5 à 100 Go. Un certain nombre d'améliorations sont encore prévues, notamment la gestion des versions des jeux de données ou la connexion à Software Heritage[28].

L'entrepôt était considéré, dès sa version 1, comme un entrepôt de confiance. La version 2 reste de confiance mais n'est pas encore certifiée. L'objectif est de faire en sorte que cet entrepôt soit certifié sans doute dans sa version 3.

En termes d'organisation, le déposant dépose son jeu de données, puis un modérateur s'assure que le type de données peut être candidat au dépôt dans Data Terra, c'est-à-dire qu'il a un potentiel suffisant de réutilisation et qu'il respecte la description standard des données et des métadonnées. Data Terra a besoin de scientifiques pouvant jouer le rôle de référent pour l'ensemble d'une communauté. Le référent fera le lien avec l'équipe de l'entrepôt et pourra identifier et former de façon initiale et continue les modérateurs de la communauté, et enfin accompagner les déposants dans l'amélioration des pratiques de dépôt. Un certain nombre de phases d'accompagnement et de communication ont ainsi été formalisées et peuvent être retrouvées sur les sites de Data Terra et de l'entrepôt : support, promotion, ateliers. Le pôle Form@Ter, qui ne disposait pas d'un entrepôt pour ses données, a bénéficié d'une expérience de quelques mois avec la communauté de sismologie du Résif-Epos (Réseau sismologique et géodésique français)[29]. Un test complet de dépôt a ainsi été réalisé de bout en bout sur des jeux de données identifiés comme éligibles à l'entrepôt. Les retours sont très instructifs, à la fois pour améliorer l'outil et faire prendre conscience aux chercheurs de l'intérêt de cette ingénierie de la donnée et de l'utilité de décrire au mieux les données déposées afin qu'elles soient réutilisables. Des problématiques liées à la définition des périmètres ont été identifiées. Les autres communautés ont été interrogées, pour le Pôle Océan ODATIS, il s'agit de pouvoir moissonner les données entreposées dans leur entrepôt de longue traîne. Pour les pôles surfaces continentales (THEIA) et atmosphère (AERIS), la discussion en est à un stade semblable à Form@Ter. Enfin, dans le cadre de GAIA Data[30], le service d'entrepôt de données de Data Terra permettra d'intégrer à ce projet la communauté CLIMERI-France[31] (Infrastructure nationale de modélisation du système climatique de la Terre, avec des modèles de simulation numérique), et le PNDB.

Le lancement est prévu en 2023. La solution technique est prête, avec une V2 opérationnelle, le facteur limitant est la mise en œuvre de la modération avec les communautés scientifiques. Des questions demeurent ouvertes, notamment la préservation à long terme (au-delà de 10 ans), les moyens humains pour assurer la continuité de ce service et les missions associées au rôle de centre de référence thématique. À titre d'exemple, SEANOE fonctionne avec 20 % d'un temps plein.

3. TGIR Huma-Num

Olivier Baude,
Huma-Num - directeur de la TGIR

Je suis très heureux de vous proposer cette présentation après les deux précédentes, car Huma-Num[32] doit beaucoup au CDS. Nous avons également une convergence assez forte avec d'autres infrastructures, notamment Data Terra. Huma-Num est une infrastructure de recherche qui s'inscrit dans une stratégie européenne comme nationale dans laquelle la dynamique de la science ouverte se concentre sur la question du cycle de vie des données. Elle a été créée il y a une dizaine d'années, à l'occasion de la fusion d'un équipement pour les données de la recherche en SHS et d'une infrastructure d'animation de communauté, afin d'apporter des services et un dynamisme au bon niveau et au bon moment pour les sciences humaines et sociales. Les données sur lesquelles nous travaillons proviennent de la tradition épistémologique des sciences humaines et sociales (ouvrages, revues, bibliographies, fonds d'archives, enquêtes, collections de documents). Une grande partie de la recherche, en matière de sciences humaines et sociales, ne se fait pas directement dans une dynamique de production de données mais est rattrapée par l'existence d'archives. Il y a deux types de publics des SHS : ceux qui sont nativement numériques, et ceux qui sont rattrapés par le numérique.

Une réflexion épistémologique classique porte en outre sur ce que sont les données. Nous bénéficions à ce niveau de l'approche des humanités numériques, c'est-à-dire sur des SHS cherchant à travailler sur de grands corpus d'œuvres, ce qui pose des questions de traitement pour la visualisation des données, leur structuration et leur organisation. Cette organisation est une organisation de la pensée elle-même et de la production scientifique. Les données sont conçues et repérées comme productions au sein d'un processus de pratique scientifique.

Huma-Num s'est construit autour d'un processus de recherche, en mettant au centre le projet de recherche et en l'environnant d'un certain nombre d'éléments accompagnant le processus de production scientifique au sein de ce projet de recherche. Nous disposons d'une infrastructure physique, de services numériques, d'une animation de communauté nationale et internationale, d'un pôle d'accompagnement à la qualité des données et d'un pôle d'innovation et de laboratoire scientifique, dans un écosystème regroupant d'autres infrastructures et d'autres partenaires, à la fois au niveau national et européen. S'agissant de l'infrastructure, nous avons choisi d'être hébergés par le Centre de calcul de l'IN2P3, dont je remercie le directeur, qui nous a permis de développer nos solutions dans un environnement compétent. Nous nous appuyons en outre sur le CINES[33] (Centre Informatique National de l'Enseignement Supérieur) pour l'archivage des données. Nous avons pour mission d'accompagner les projets de recherche jusqu'aux autres structures qui permettent de traiter les données, jusqu'au supercalculateur Jean Zay. Le besoin en traitement et puissance de calcul explose depuis quelques années en SHS. S'agissant des services que nous avons développés sur l'ensemble du cycle de vie, nous sommes aujourd'hui sous une forte pression d'utilisation. En un an, un tiers des acteurs de l'enseignement supérieur et de la recherche en SHS utilise des services Huma-Num, soit 15 000 utilisateurs. Un simple service comme un gestionnaire de fichiers et de partage, qui correspond souvent à la première demande dans les projets de recherche, donne lieu à une explosion des demandes et enregistre une très forte croissance. Nous hébergeons en outre plus de 800 sites et 1 000 bases de données. Nous voyons le développement de services qui sont orientés vers les pratiques des sciences humaines et sociales, avec la question des données brutes, dites tièdes, et des données chaudes. Nous proposons également trois services qui apportent une forte valeur ajoutée à la donnée : un entrepôt de données, Nakala, un moteur de recherche et assistant scientifique et un service de préservation à long terme.

L'entrepôt Nakala[34] est orienté vers la mise en place des principes FAIR et des valeurs de la science ouverte. Il a été développé en deux phases : une phase exploratoire depuis 2015, et depuis deux ans, une phase correspondant davantage aux pratiques actuelles. Les principes clés de cet entrepôt sont le dépôt de fichiers de tout type de format, l'ajout de métadonnées (Dublin Core, mais avec un vocabulaire extensible sur demande), l'organisation de ces dépôts en collection, la gestion des droits d'accès, l'attribution d'identifiants pérennes (Handle à l'origine, DOI depuis 2020), des visionneuses et un moteur de recherche sur les collections, données, métadonnées et contenus des fichiers. L'exposition des données se fait en

▶▶▶

OAI-PMH[35], en RDF[36] et en API REST, afin de tenir compte des différents types d'usages auxquels nous sommes confrontés. Plus récemment, nous avons développé un module de création de site web intégré à l'entrepôt, afin de faciliter l'exposition des données et de permettre à des projets de recherche d'adopter cette bonne pratique de gestion des données dans un entrepôt de données. En deux ans, nous avons vu une augmentation de 50 % de l'utilisation de Nakala, ainsi qu'une explosion du nombre de comptes et de fichiers gérés par l'entrepôt. Nous documentons et formons aujourd'hui à l'usage de cet entrepôt, en développant un certain nombre de documentations et d'actions de formation, ce qui demande beaucoup de ressources humaines. Nous avons en outre choisi d'avoir une relation directe avec les utilisateurs de l'entrepôt. Nous avons par ailleurs produit un POC (*Proof of Concept*) avec le CCSD dans le cadre d'un projet européen entre Nakala et HAL. Les données exposées dans Nakala contiennent ainsi un lien direct vers la fiche HAL, et inversement. Ce lien représente un enjeu fort dans nos perspectives. Il est au cœur de l'EquipEx+ COMMONS[37] (Consortium de moyens mutualisés pour des services et données ouvertes en SHS) que nous développons, à savoir un Equipex dédié au lien entre la continuité de la gestion des données et des publications par des services d'infrastructures.

D'autres services sont proposés : la gestion directe d'un moteur de recherche signalant l'ensemble des données des productions scientifiques en SHS et s'appuyant sur les technologies du web sémantique, afin d'approfondir la production d'un savoir à partir des données qui y sont liées. Cet entrepôt permet en outre de s'inscrire dans un processus d'archivage sur le long terme des données. Nous apportons ce service en lien avec le CINES, sous la responsabilité des Archives nationales.

Du fait de ce foisonnement, notre équipe est aujourd'hui très sollicitée par l'accompagnement des utilisateurs. Il s'agit d'un enjeu en termes de modération, mais aussi de montée en puissance de la qualité des données. L'ensemble de ces services est assuré avec des partenaires et des communautés scientifiques. Les partenaires sont les correspondants Huma-Num présents dans les Maisons des sciences de l'homme, dans les universités et dans les ateliers sur l'ensemble du territoire. Il s'agit également de coopérations avec d'autres partenaires et de la mise en place de consortiums qui intègrent l'ensemble des communautés en SHS, qui travaillent sur la production de bonnes pratiques, standards, méthodes et formations.

Huma-Num, enfin, a une activité européenne et internationale forte, qui nous permet de lier des activités de l'entrepôt et de membres d'Huma-Num aux projets européens et internationaux. Il s'inscrit au cœur de Recherche Data Gouv, en lien avec les ateliers de la donnée, en tant que centre de référence thématique.

4. Institut français de bioinformatique (IFB)

Jacques Van Helden,
Institut français de bioinformatique (IFB) - Co directeur

Sur le sujet des enjeux pour les données de la biologie, un changement de paradigme est intervenu au tournant du XXIe siècle. La biologie s'est ainsi convertie à une science de données. Tous les biologistes sont à présent amenés à analyser leurs données, ce que faisait plutôt la génomique, initialement. Ce phénomène s'est étendu à tous les domaines d'application et à toutes les thématiques de recherche fondamentale. Nous sommes en outre confrontés aux mêmes défis scientifiques et technologiques que nos collègues d'autres domaines (calcul intensif, intégration des données hétérogènes, apprentissage automatique, gestion des données, interopérabilité, FAIRisation). Les communautés de biologie-santé ne sont pas toujours formées pour analyser les données et pour l'usage informatique, d'où une forte demande de formation et d'accompagnement. Nos données ont en outre quelques spécificités ; nous gérons des objets très complexes, dont il est difficile d'obtenir des descriptions unifiées au sein des communautés. De plus, la biologie intégrative, consistant à intégrer différents types de données et à étudier les objets biologiques à plusieurs échelles, représente un réel enjeu. S'agissant de la protection des données de santé, nous sommes amenés, dans de nombreux projets, à rassembler les données de biologie moléculaire et de biologie cellulaire, avec des données de santé. Nous devons également faire face à des enjeux sociétaux de la bioinformatique : rationalisation des moyens, enjeu écologique, mutualisation des compétences. La biologie a des enjeux dans un ensemble de domaines d'application : recherche fondamentale, agriculture, santé, biotechnologie, environnement.

L'Institut français de Bioinformatique a été fondé en 2013 sur le PIA 1 (RENABI-IFB). Il a pour rôle de développer des services, des formations et de réaliser des développements innovants. Nous déployons une infrastructure numérique au niveau national, répartie sur différentes plateformes. Nous développons également des outils logiciels pour répondre aux nouvelles questions de la biologie. Nous développons des bases de connaissance, réalisons de nombreuses formations et avons ouvert dès 2022 sur notre feuille de route une mission « Science ouverte et interopérabilité », qui regroupe des initiatives que nous prenions dans d'autres missions. Nous avons également une mission « Innovation et prospective » et disposons d'un guichet central pour l'accompagnement des usagers, qui les redistribue vers les plateformes.

Nous fédérons 36 plateformes et équipes de recherche qui couvrent l'ensemble du territoire métropolitain. L'infrastructure a quatre tutelles : le CNRS, l'INSERM, l'INRAE et le CEA. Nos plateformes et équipes de recherche appartiennent à d'autres tutelles. Une unité coordonne l'utilisation des moyens mis à notre disposition pour nos missions nationales et internationales. Ces moyens sont relativement conséquents ; dans le cadre du PIA 1, nous avons disposé de 22,8 millions d'euros jusqu'à 2024, et dans le PIA 3 (MUDIS4LS), nous venons de disposer de 16,5 millions d'euros, pour mettre en place une infrastructure distribuée et assurer l'orchestration des flux de données depuis leur production et au travers du cycle de la donnée. Nous nous inscrivons en outre dans une articulation forte avec l'Europe, via ELIXIR[38].

Nous avons été nommés centre de référence thématique pour les données de Biologie-Santé. Les missions des CRT mettent l'accent sur le développement des bonnes pratiques et sur le rôle international. Par ailleurs, nous sommes amenés à manipuler des données de santé, mais nous n'avons pas la compétence pour l'ensemble des données de santé ni les données biologiques qui ne relèvent pas de la bioinformatique.

En tant que centre de recherche thématique, nous devons interagir avec les ateliers de la donnée, qui sont au contact des usagers et ont la mission de diffuser auprès d'eux les bonnes pratiques. Ils s'appuient pour cela sur ce que nous développons. Ces ateliers nous font remonter les besoins et les lacunes des communautés. Nous intervenons auprès de celles-ci au niveau de la production des données, pour leur structuration dans les bases de données, leur analyse, le stockage à chaud pendant la durée des projets, et la préservation à long terme en nous reposant sur des partenaires nationaux et pour la formation.

Pour jouer ce rôle de centre de référence thématique, nous nous appuyons sur des initiatives que nous avions déjà initiées dans le cadre de notre feuille de route précédente. Nous avions notamment entrepris un chantier avec l'ensemble des

▶▶▶

infrastructures qui produisent des données. Nous collaborons ainsi depuis 2017 pour la mutualisation des ressources numériques, mais aussi pour mettre en place des plans de gestion de données modulaires. Nous assurons également du *data procuring*, en accompagnant les chercheurs pour leur faciliter le dépôt dans les entrepôts internationaux et en veillant à améliorer la qualité des métadonnées. Nous avons en outre mis en place, grâce à un support de l'ANR et aux appels flash pour la science ouverte, deux initiatives. La première consiste à développer des plans de gestion de données interopérables du point de vue informatique, ce qui nous permet d'interconnecter l'allocation de nos ressources numériques avec un plan de données. Le PGD devient ainsi un outil d'accompagnement tout au long de la vie du projet. OpenLink vise quant à lui à lier électroniquement le cahier de bord électronique avec tous les endroits où la donnée sera traitée au fil de son parcours. Nous travaillons également sur l'interopérabilité, avec un outil FAIR Checker qui permet de mesurer le degré de FAIRisation d'une ressource internet. Nous avons organisé deux formations orientées vers la FAIRisation, l'une sur la donnée (Fair Data) et l'autre sur la FAIRisation du code que produisent les bioinformaticiens (FAIR Bioinfo). Nous avons adopté un principe de « *train the trainer* » : les premières versions de ces formations ont été orientées vers des bioinformaticiens des plateformes qui doivent se les approprier pour les redéployer à leur tour.

Le projet Espaces numériques mutualisés pour les sciences du vivant (MUDIS4LS), financé par l'équipe Equipex +, a un volet technologique, avec le développement d'une infrastructure numérique distribuée sur les différents datacentres régionaux et nationaux, et couvre également l'articulation avec les plans de gestion de données ainsi que des *implementation studies* visant à mettre en pratique cette infrastructure numérique, en la mettant dans les mains des différentes communautés ciblées. L'orchestration des flux de données repose sur une généralisation des plans de gestion de données interopérables. Initialement, ceux-ci visaient à interconnecter le plan de gestion de données avec l'allocation de ressources sur notre infrastructure numérique. Nous allons désormais les généraliser, en définissant avec toutes les infrastructures productrices de données, qui produira les données puis où elles doivent être déposées sur notre infrastructure. Nous jouons également le rôle de *data procuring*, en facilitant le dépôt des données dans les entrepôts internationaux et dans l'entrepôt national que nous souhaitons développer, Bio Dataverse, de façon automatisée. L'automatisation des plans de gestion de données permettra aux agences de financement d'évaluer le devenir des données.

En termes de politique générale de protection des données, nous travaillons avec la déléguée à la protection des données (DPO) du CNRS. Chacune de nos plateformes travaille avec sa propre DPO. Cette politique consiste à ouvrir autant que possible et fermer autant que nécessaire. Nous devons cependant traiter des métadonnées de santé associées aux données de patients. Nous sommes, par exemple, impliqués dans plusieurs projets de séquençage génomique. Dans ce cadre, la donnée n'est pas anonymisable. Une donnée qui a une valeur prédictive pour la santé verra sa valeur augmenter avec l'évolution des connaissances des génomes. De plus, elle n'est pas seulement personnelle mais familiale. Un certain nombre de projets sont affectés par cet important défi, notamment un projet international, *European Genomic Data Infrastructure*, dans le cadre duquel le PFMG et l'IFB sont partenaires.

Nous avons un très fort ancrage sur nos plateformes, réparties partout en France, mais nous bénéficions aussi de nos interactions avec nos partenaires européens, dans le cadre d'ELIXIR, dont un grand nombre d'initiatives sont liées à la donnée. Le *Service Deliver Plan* vise à labelliser toutes les initiatives nationales, notamment les bases de données, outils, formations, etc. Certains types de données sont labellisés *Core Data Resources*[39]. L'enjeu est de définir des données qui, si elles venaient à disparaître, mettraient en danger toute la recherche en biologie. Dans le cadre d'EDAM, nous développons en outre une ontologie des termes de la bioinformatique. Nous espérons un effet levier par la mobilisation des communautés. Le projet ELIXIR-CONVERGE, quant à lui, financé par l'Union européenne en 2019 à hauteur de 5 millions d'euros, vise à créer un réseau d'experts en gestion de la donnée et développer des outils de partage de la donnée. En 2020, l'Union européenne a décidé d'injecter 5 millions d'euros supplémentaires afin d'ajouter trois work packages dédiés au partage des données Covid-19.

En conclusion, la mise en œuvre des standards sera un défi conséquent. Il s'agira en outre de concilier l'ouverture et la protection des données sensibles, de toucher l'ensemble des communautés de biologie-santé ou encore d'assurer l'intégration des données hétérogènes de la biologie. En termes de points d'appui, nous sommes bien dotés, financièrement, par les PIA. Nous bénéficions également d'un très fort engagement de la communauté bioinformatique (plateformes, équipes de recherche, enseignants) ainsi que d'une collaboration active entre les infrastructures nationales de biologie-santé via le club des INBS et IBISA. Les réseaux nationaux mis en place par le MESRI sont quant à eux structurants. Enfin, nous bénéficions de la mutualisation de nos réalisations avec nos pairs européens.

RECHERCHE DATA GOUV : UN ÉCOSYSTEME AU SERVICE DU PARTAGE ET DE L'OUVERTURE DES DONNÉES DE RECHERCHE

Isabelle Blanc,
MESR, Administratrice ministérielle des données, des algorithmes et des codes sources

Mon intervention a pour objet la présentation de la plateforme Recherche Data Gouv (RDG). En premier lieu, je rappellerai qu'à l'origine la création de cette plateforme répond à l'ambition, posée par la loi pour une République numérique (2016), dans le premier PNSO (2018) et renouvelée dans le deuxième PNSO (2021) : l'ouverture de l'ensemble de nos produits et de nos méthodes de recherche, c'est-à-dire, les publications, les données et les codes sources. La première étape de l'open access avec les publications était ce qu'il y avait de plus simple à faire, par contre, l'ouverture des données et des codes n'est pas si évidente.

Les engagements pris au niveau du ministère en mettant des moyens et en étant facilitant, sont ceux présents dans le deuxième PNSO (2021) et dans la politique des données, des algorithmes et des codes sources probablement moins bien connue. Ces deux politiques sont complémentaires et superposables, et dans ce cadre, je pilote cette politique et la feuille de route qui comporte une cinquantaine d'actions et je pilote également directement les deux axes du PNSO : l'axe des données et l'axe des codes et logiciels. L'ambition est donc de « soutenir la structuration, la préservation, le partage, l'ouverture, la découverte des données de la recherche » afin de favoriser les pratiques de réutilisation de ces données de recherche. Et c'est dans ce contexte, que se situe la création de Recherche Data Gouv.

La politique des données amène aussi la réalisation d'actions et missions autour de la production de guides d'accompagnement, métiers et activités sur les données de la recherche, pour mieux recruter et reconnaitre et mieux accompagner l'ensemble des parcours. Dans les activités complémentaires, on note le développement d'une doctrine juridique sur la base des cadres juridiques existant afin de pouvoir vous accompagner grâce à la création d'un dispositif de ressources juridiques à destination de toutes les communautés qui en ont besoin.

En ce qui concerne Recherche Data Gouv, avant tout, le premier objectif est « Ne pas laisser d'équipes de recherche sans solution pour ouvrir ou partager leurs données » : angle principal et principe directeur de RDG. Ensuite, faire de l'accompagnement un élément central du dispositif : cela se traduit par la structuration depuis le ministère d'un maillage des offres d'accompagnement en proximité de la recherche. On constate qu'autour de 80 % des équipes de recherche sont plutôt démunies face aux questions liées aux données de la recherche. Il est donc nécessaire de les sensibiliser, former ou accompagner. Au-delà de ce qui existait dans certains domaines déjà très avancés sur la question (infrastructures, etc.), il y avait nécessité de compléter ce maillage avec un premier niveau d'accompagnement au plus près du terrain et en incitant les établissements en partenariat avec les équipes.

Vous l'avez tous souligné ; nous manquons de compétences sur ces métiers multiples, qui regroupent à la fois des compétences scientifiques de documentalistes ou de bibliothécaires, mais aussi d'informaticiens. Il s'agit de savoir comment ces compétences sont regroupées sur un même territoire pour éviter que chaque établissement et chaque unité ait besoin de les développer. L'objectif est donc cette mutualisation entre établissements de services et compétences pour accompagner les équipes au plus près de leurs préoccupations, en leur apportant un premier niveau de réponse. Ce développement va se poursuivre au fil des années.

Vous avez bénéficié du témoignage d'un certain nombre de centres de référence thématiques. Nous avons en effet souhaité les intégrer dans l'écosystème Recherche Data Gouv en raison de leur expérience de longue date. L'objectif était de s'appuyer sur cette expertise et de faire en sorte que toutes les données concernant les centres de référence thématiques et les besoins d'accompagnement soient redirigés vers eux. Leur rôle est encore en construction, au même rythme que la construction des ateliers de la donnée.

Les centres de ressources, quant à eux, interviennent en appui aux ateliers de la donnée, qui sont en proximité des équipes de recherche, qui ont besoin de mutualiser des ressources. Ces centres de ressources ont ainsi fédéré des initiatives existantes, notamment les initiatives CNRS/INIST. Un centre de ressources est dédié à l'entrepôt-catalogue, qui développe et maintient l'entrepôt et accompagne les utilisateurs. Un centre de ressources est en outre dédié aux compétences, afin d'assurer la montée en compétence des ateliers de la donnée. Par ailleurs, une étude est lancée sur les métiers de la donnée.

L'ensemble de l'écosystème (ateliers, centres de référence thématiques, centres de ressources, entrepôt) est accessible depuis le portail Recherche Data Gouv. Les modules entrepôt et catalogue de données sont une possibilité offerte aux équipes de recherche. Chaque établissement assure la modération de ces données. Cet entrepôt disciplinaire est à conseiller aux équipes de recherche lorsque des entrepôts thématiques de confiance n'existent pas. L'objectif du catalogue est de repérer et signaler les données déposées dans ces entrepôts thématiques de confiance. Un travail est ainsi en cours sur la définition des critères de confiance d'un entrepôt.

L'entrepôt et le catalogue sont développés par et pour la communauté de recherche, bien qu'il s'agisse d'une initiative

▶▶▶

financée et soutenue par le Ministère. Elle a été confiée à l'INRAE, qui a proposé d'adapter et de mutualiser cette solution. Sept établissements l'ont rejoint, dont le CNRS. Ils contribuent directement au développement de l'entrepôt, qui a ouvert en juillet dernier. Nous avons une gouvernance collective, incluant les partenaires du projet et les représentants de la communauté ESR. Les décisions collégiales conduisent ainsi aux différents choix que nous opérons.

Pour rejoindre l'écosystème, il est possible de créer un espace institutionnel et de devenir atelier de la donnée. Depuis le 8 juillet 2022, 16 espaces institutionnels sont actifs, quatre sont en cours d'instruction ou de création, 391 jeux de données (qui représentent plusieurs dizaines ou centaines de fichiers) ont été déposés, 80 000 fichiers ont été téléchargés. Des ateliers sont en cours de labellisation et un certain nombre de sessions de formation ont eu lieu, regroupant de nombreux participants. Vous pouvez retrouver toutes les actualités sur le site de Recherche Data Gouv. Merci à tous pour votre attention.

Ateliers de la donnée (Liste mise à jour, novembre 2023)

Quatorze ateliers labellisés

- **dat@UBFC,** Atelier de la donnée UBFC - service de gestion et d'ouverture des données de la recherche de l'Université de Bourgogne Franche-Comté (dat@UBFC), présenté par l'Université de Bourgogne Franche Comté
- **Open DATA & SCIENCE URCA**, Ouvrir et valoriser les données de la recherche à l'URCA en conformité avec les bonnes pratiques de la science ouverte, présenté par l'Université de Reims Champagne Ardenne
- **ADOC**, Accompagner aux DOnnées les Chercheurs et chercheuses en Lorraine. Parce que nous choisir, c'est la solution ad hoc !, présenté par l'Université de Lorraine
- **ADELE**, Atelier DonnéE aLsacE Helpdesk, présenté par l'Université de Strasbourg
- **CDGA**, Cellule Data Grenoble Alpes, présenté par l'Université Grenoble Alpes
- **Data Univ Eiffel**, Atelier data Université Gustave Eiffel
- **ADCVL**, Atelier de la donnée Centre-Val de Loire présenté par l'Université de Tours
- **LORD**, Lille Open Research Data présenté par l'Université de Lille
- **ECODOR**, Vers un écosystème des données de la recherche : pour un accompagnement mutualisé sur le site de Montpellier présenté par l'Université de Montpellier
- **ADN**, Accompagnement de la Donnée à et avec Nanterre, présenté par l'Université Paris Nanterre
- **ADOO**, Atelier de la Donnée d'Occitanie Ouest présenté par l'Université Fédérale de Toulouse Midi-Pyrénées
- **ADN**, Atelier de la donnée en Normandie, présenté par Normandie Université
- **DatASaclay**, Dat'Atelier de l'Université Paris-Saclay présenté par l'Université Paris-Saclay
- **Données Condorcet**, Atelier de la donnée du Campus Condorcet

Cinq ateliers « sur la trajectoire de labellisation »

- **GDsAM**, Guichet de la Donnée du site d'Aix-Marseille
- **ARDoISE**, Atelier Rennais de la Donnée : Information et Soutien aux Equipes de recherche, présenté par l'Université de Rennes
- **ADSSD**, Atelier de la donnée de la Seine-Saint-Denis, présenté par l'Université Paris 8
- **G.O.T.O.DO UCA**, Gérer Ouvrir Trouver Obtenir les Données de l'Université Clermont Auvergne
- **OREADE Nantes Université**, ORientation Et Accompagnement pour les Données de la RecherchE à Nantes Université

Source : recherche.data.gouv.fr

EUROPEAN OPEN SCIENCE CLOUD(EOSC)

Suzanne Dumouchel,
CNRS, Chargée de mission EOSC au CNRS
et membre du Board de l'Association EOSC

Je vais vous parler d'EOSC en abordant deux perspectives : d'une part, l'action que nous menons au sein du CNRS, d'autre part, l'association EOSC et ses principales actions. Volker Beckmann apportera quant à lui une vision ministérielle de la participation de la France à la construction d'EOSC.

S'agissant d'abord de l'activité du CNRS dans EOSC, nous avons mis en place deux groupes de travail : un atelier de travail, composé essentiellement de membres opérationnels du CNRS dans les projets EOSC ou dans les groupes de travail de l'association européenne, et un atelier stratégique et opérationnel, composé de membres du COPIL de la DDOR et de directeurs d'infrastructures du CNRS, pour avoir une vision opérationnelle mais aussi plus stratégique. Le CNRS participe en outre à différents projets estampillés EOSC : des projets de développement de service, de structuration de communautés scientifiques, des projets portant sur des problématiques liées à EOSC (comme le projet FAIR-IMPACT), des projets sur la mise en place du cœur d'EOSC (à l'image d'EOSC Future), et des projets sur les compétences FAIR (comme Skills4EOSC). Le CNRS est membre de l'association EOSC, il participe à diverses taskforces de l'association et plusieurs de ses services sont intégrés dans le portail EOSC. Nous comptons aujourd'hui une vingtaine de personnels CNRS impliqués dans les groupes de travail de l'association, sans parler de la participation aux différents projets ou initiatives européens qui gravitent autour de l'écosystème EOSC. Le CNRS est représenté dans le Conseil d'Administration de l'association EOSC et participe au Collège EOSC-France, afin de travailler sur la stratégie EOSC au niveau national.

Plusieurs chantiers ont été lancés dans les groupes de travail :

1. Répondre à l'enquête de l'Association ESOC sur les activités et services CNRS liés à EOSC (engagement des membres de l'association) ;
2. Analyse de faisabilité d'un catalogue français de services EOSC à partir du catalogue Cat Opidor ;
3. Animer un blog de publication des textes principaux relatifs à EOSC mais traduits en français.

L'association EOSC est constituée de représentants des organisations. Elle est pilotée par un board composé de huit directeurs, élus pour trois ans, et d'un président. L'assemblée générale de l'association EOSC s'est tenue hier et avant-hier. Elle a réélu Karel Luyben au poste de président pour trois ans, ainsi que Sarah Jones et Ignacio Blanquer dans leurs postes de directeurs. L'association EOSC s'organise autour de 13 *taskforces*, qui s'articulent autour de cinq *advisory groups*, ouverts aux membres et observateurs de l'association. Nous avons un personnel français dans chacune des *taskforces*. Hier, l'assemblée générale a validé la création d'un groupe de travail ouvert exclusivement aux infrastructures de recherche européennes qui sont sur la feuille de route ESFRI (*European Strategic Forum for Research Infrastructure*)[40] et qui sont membres de l'association européenne EOSC. Son rôle sera de renforcer la présence de la communauté scientifique dans EOSC, de contribuer à l'engagement des chercheurs dans EOSC et de contribuer techniquement aux enjeux de la création d'EOSC pour les infrastructures de recherche. Les *advisory groups* travaillent sur des thématiques techniques, et ce groupe de travail permettra de nourrir le dialogue avec ces groupes consultatifs. Les discussions sont en cours pour créer un second groupe, réunissant cette fois les universités voire d'autres acteurs. Enfin, la perspective nationale passe par l'engagement des organisations mandatées par chacun des ministères et les événements tripartites.

S'agissant de la gouvernance d'EOSC, l'association a été créée en juillet 2020. Elle dispose d'un secrétariat opérationnel depuis un an seulement. Un partenariat tripartite a été signé en juin 2021, structuré autour de la Commission européenne, des États membres et de l'association EOSC (qui représente la communauté scientifique). Le ministère français est bien entendu représenté. Volker Beckmann coordonne l'activité des États membres au sein de ce groupe. Les événements tripartites sont organisés dans chacun des pays et réunissent au minimum une personne de la Commission européenne, une personne du board de l'association et un représentant des États membres, généralement celui du pays concerné. Le premier événement tripartite a été organisé en France en avril dernier. Le SRIA (*Strategic Research and Innovation Agenda*) et le MAR (*Multi-Annual Roadmap*) sont deux documents très importants au sein de l'association ESOC. Ils doivent être régulièrement mis à jour. Le SRIA vise à identifier les actions à mettre en place dans les années à venir pour qu'EOSC soit fonctionnel et utilisable. Le MAR a quant à lui pour objet d'identifier des priorités en fonction d'un calendrier. Nous rentrons à présent dans une seconde phase du MAR, qui correspond à la production de valeur ajoutée. La première phase était consacrée à l'organisation et à la structuration d'EOSC. Des priorités ont ainsi été identifiées dans le MAR pour 2023 et 2024. Le programme de travail présenté à l'assemblée générale repose sur trois axes : le partenariat avec EOSC, le projet Focus EOSC et le soutien à l'activité de l'association EOSC.

Volker Beckmann,
MESR, Chargé de mission EOSC

Je vais revenir sur les réalisations au niveau national et européen. En tant qu'organismes de recherche, vous dépendez des politiques et stratégies nationales. En France, nous mettons en place des stratégies très fortes, via la feuille de route pour les infrastructures de recherche ou le plan national pour la science ouverte, mais cela n'est pas le cas dans tous les pays. Il est donc nécessaire d'avancer sur le plan de la science ouverte en Europe. Nous avons décidé d'accélérer notre action, considérant que nous ne progressions pas à une vitesse acceptable.

Au sein du *steering board*, nous discutons des politiques nécessaires pour faire avancer la science ouverte. Trois défis ont été identifiés en 2022, qui ont donné lieu à des publications et recommandations adressées aux pays pour leur permettre d'avancer sur un certain nombre d'aspects :

- *EOSC and data literacy*,
La science ouverte est nécessaire pour diffuser l'expertise au niveau des laboratoires et des universités, et pour augmenter la connaissance. Il ne s'agit pas seulement de créer de nouveaux postes, mais de mettre en œuvre des formations. Au niveau européen, l'objectif est de créer un cadre européen pour les compétences en matière de données FAIR, ainsi qu'un réseau européen de *data competence centers*.

- *EOSC and the commercial partners*,
Nous souhaitons inviter les partenaires commerciaux à participer à l'action d'EOSC, qui s'adresse également au secteur commercial. Un travail doit également être conduit sur l'analyse des risques et bénéfices.

- *EOSC sovereignty on FAIR Data*.
Ce thème pose le sujet de la qualité des données. Nous souhaitons assurer, au niveau d'EOSC, une qualité satisfaisante des services et données.

Nous avons réalisé cette année une quatrième publication. Au-delà de la production des solutions pour la science ouverte, nous devons assurer une surveillance au niveau national et au niveau européen, afin d'identifier les politiques en place, les bonnes pratiques à partager au sein des pays, etc. L'enjeu est de créer un monitoring entre les pays européens, notamment pour identifier les manques et communiquer, valider, développer et affiner les politiques. Nous souhaitons en outre qu'EOSC soit considéré comme un des *European Data Spaces*[41], qui sont axés sur certaines thématiques (santé, énergie, transport, etc.). Vous pouvez consulter les résultats de cette activité de monitoring sur une page web dédiée.

Au niveau national, le collège EOSC-France est intégré dans la comitologie des ministères. Il est composé des représentants des grands organismes français. Des groupes de travail sont chargés de fournir des avis et des recommandations. Nous avons des discussions au sein des collèges, et émettons des recommandations auprès du secrétariat. La dernière instance est le comité de pilotage. Nous assurons donc une connexion directe entre les experts et les décideurs, pour assurer une prise de décision top-down permettant la mise en place de la science ouverte en France.

S'agissant des prochaines étapes, il s'agit de s'engager dans la mise en œuvre des recommandations des quatre *opinion papers*. Nous souhaitons en outre disposer d'un catalogue des bonnes pratiques existantes. Notre message est le suivant : « *EOSC is here to stay; there's no turning back*. » Il est nécessaire de partager les données, infrastructures, services et ressources humaines pour faire avancer notre recherche au niveau national, européen et mondial.

DONNÉES DE SANTÉ ET DONNÉES SENSIBLES

1. Health DataHub et accès aux bases de données SNDS

Emmanuel Bacry,
CNRS - Directeur scientifique du Health Data Hub

Le Health Data Hub (HDH) est un groupement d'intérêt public visant à faciliter l'accès aux données de santé pour une utilisation secondaire de ces données, donc pour la recherche. Il a été créé à la suite du rapport Villani (2018) et s'inscrit dans la stratégie nationale sur l'intelligence artificielle. Cette mission est subdivisée en quatre grandes missions :

- la mise en place d'un guichet unique d'accès aux données de santé,
- la création d'une plateforme sécurisée et à l'état de l'art,
- la mise en place d'un catalogue de données de santé,
- l'animation de l'écosystème, qu'il soit national ou international.

Un an après la création du Hub, la plateforme technologique a été mise en production et emploie 84 personnes. Nous accompagnons 74 projets. Nous comptons plus d'une centaine de partenariats actifs et avons effectué plus de 300 interventions depuis début 2021.

Au sujet d'abord du guichet unique, la CNIL[42] a mis en place des procédures simplifiées pour accéder aux données. S'agissant des méthodologies de référence, un chercheur ou médecin dans un hôpital peut accéder aux données produites par l'hôpital pour faire de la recherche avec l'autorisation de la gouvernance locale de l'hôpital. Les pratiques des gouvernances locales sont cependant extrêmement hétérogènes en la matière. Aucune procédure n'a été standardisée sur ce sujet. Il est courant d'attendre trois à quatre ans avant d'avoir accès aux données de santé, voire de ne pas y avoir accès, parfois sans raison. En dehors de ces procédures d'accès rapide, il existe une procédure nationale où le Health Data Hub joue le rôle de guichet. Une demande est déposée avec son aide (voir le kit pédagogique). Il s'assure que le dossier est complet puis le transmet à un comité indépendant, qui évalue le projet scientifiquement et éthiquement, ainsi que l'intérêt public du projet. La réponse est communiquée en un mois. Si le projet est accepté, un autre dossier d'homologation doit être complété puis est adressé à la CNIL, qui vérifie par exemple que les citoyens ont bien été informés. La réponse de la CNIL est fournie sous deux mois. Une prolongation de deux mois peut être demandée. À l'issue de ce délai, un accord implicite d'accès est acquis.

Depuis trois ans, 300 projets sont soumis chaque année au CESREES (Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé). 540 ont été soumis à la CNIL. Plus de 60 % des projets requièrent l'accès au Système national des Données de Santé[43] (SNDS). 35 % de l'ensemble des projets opérés demandent un chaînage de ces données avec des données cliniques. Le SNDS joue donc un rôle extrêmement important au sein du Hub et dans la recherche sur les données de santé française. Le SNDS est une base de données médico-administratives, qui ne comprend donc quasiment pas de données cliniques, et qui a été constitué pour gérer les remboursements. Il s'agit d'une compilation de trois bases chaînées entre elles : le SNIIRAM[44] (qui correspond aux remboursements de la Carte vitale, et qui regroupe 67 millions de personnes, sans aucun biais de sélection), le PMSI (qui est l'équivalent du SNIIRAM à l'hôpital) et le CépiDc (causes médicales de décès issues de l'INSERM). Cette base est unique au monde. Ses applications sont considérables, notamment en pharmacovigilance. Si elle était remontée pendant la première vague Covid, par exemple, elle aurait permis de déterminer immédiatement que l'hydroxychloroquine n'avait aucun effet. Une étude EPI-PHARE unique au monde a été conduite sur l'impact du vaccin sur les personnes de plus de 50 ans, soit 25 millions de personnes. Les enjeux de santé publique et de surveillance sont donc considérables. Depuis peu, le CNRS bénéficie d'un accès permanent, pour tout chercheur d'unité mixte CNRS. Pour accéder à cette base SNDS, il est donc possible d'éviter la procédure nationale standard. L'accès peut se faire sur le portail de la CNAM, de façon limitée d'un point de vue de recherche méthodologique, ou sur la plateforme UE, qui donne accès à des puissances de calcul plus importantes.

Il est cependant important de disposer d'autres bases. Dans l'idéal, un réseau d'entrepôts de données interopérables permettrait des extractions à la demande. Cette interopérabilité n'existe pas aujourd'hui, et les accès aux données sont extrêmement longs. Depuis sa création, le Hub souhaite mettre en place un catalogue de bases de données important, fixé par décret mais est dans l'attente d'un accord CNIL.

Le HDH est aussi très actif en matière d'open source. Des travaux sont conduits sur l'algorithmie, notamment sur le SNDS,

►►►

à travers des nombreux appels à projets. Tous ces algorithmes sont publiés. Le Hub a organisé un data challenge qui a rencontré un grand succès, tandis que six autres seront mis en place cette année, ouverts à l'international. Un symposium a également été organisé avec le MIT. Des appels à projets sont lancés avec de grandes institutions. Un projet est en cours sur la santé environnementale avec le ministère de la Transition écologique. Un autre vient de se terminer avec le Québec.

Enfin, le HDH est extrêmement actif à l'international et a de nombreux partenariats. Des partenariats sont en cours de montage avec Israël et le Japon. Un projet dont le Hub est particulièrement fier : il est aujourd'hui le leader du consortium qui opérera le pilote de l'Espace européen des Données de Santé (EHDS). Ce consortium réunit huit plateformes nationales (France, Finlande, Danemark, Norvège, Belgique, Allemagne, Croatie, Hongrie), l'Agence européenne du Médicament, une autre agence européenne ainsi que des infrastructures de recherche européennes et internationales. Deux ans sont prévus pour réaliser ce pilote, dont le but est double. Le premier objectif est de construire un réseau entre les infrastructures existantes afin de disposer d'un portail centralisé permettant d'interroger toutes les métadonnées disponibles sur le réseau, et d'un formulaire d'accès unique permettant d'adresser toutes les données de l'ensemble des pays. Il existe aujourd'hui une grande hétérogénéité dans les modes d'accès aux données, suivant les pays. Il s'agit également de tester la capacité du réseau à transférer des données individuelles. Le second objectif vise à réaliser des cas d'usage à l'échelle européenne. Cinq cas ont été sélectionnés par la Commission, qui font intervenir plus ou moins de pays (six au maximum).

2. Protection des données sensibles sur le supercalculateur Jean Zay

Guillaume Harry,
IDRIS – CNRS, Responsable de la sécurité des systèmes d'information

Cet exposé présente un retour d'expérience, sur deux ans, sur le traitement de données sensibles sur le supercalculateur Jean Zay d'IDRIS. L'IDRIS est une unité du CNRS qui est centre national pour le calcul intensif. Il héberge et exploite le supercalculateur Jean Zay. L'IDRIS met à disposition toutes les ressources humaines et matérielles pour lancer des projets en intelligence artificielle (IA) ou en calcul intensif (High performance computing ou HPC[45]) ainsi qu'un support dédié aux utilisateurs, afin de les aider à avancer plus rapidement dans leur projet. Le supercalculateur Jean Zay n'est pas totalement financé par le CNRS, mais par le Grand Équipement national de Calcul intensif[46] (GENCI), qui finance également deux autres calculateurs d'envergure nationale, au CINES et au TGCC (très grand centre de calcul du CEA) et coordonne l'Equipex+ MesoNET[47] qui fédère des mésocentres. Jean Zay est aujourd'hui le calculateur le plus puissant, en production, pour la recherche académique. Il compte ainsi plus de 1 200 projets, plus de 2 000 utilisateurs. Depuis peu, certains utilisateurs travaillentsur des données sensibles.

Les données sensibles sont des données à caractère personnel, à savoir toute information permettant d'identifier une personne physique. Inversement, l'anonymisation permet de rendre toute identification de la personne impossible. Un traitement de données à caractère personnel correspond à toute opération portant sur des données personnelles. Pour caractériser une donnée à caractère personnel, un numéro d'identifiant concernant une personne physique suffit. Il s'agit donc de faire preuve de beaucoup de prudence vis-à-vis de ces données. Les photographies et les fichiers audio sont également concernés. Des projets d'intelligence artificielle visant à traiter des enregistrements audio, si la personne peut être reconnue, manipulent ainsi des données à caractère personnel. Certaines données à caractère personnel justifient plus de prudence encore : les données sensibles. Les données sensibles correspondent à des informations qui révèlent l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses, les données biométriques, les données de santé et les données concernant la vie sexuelle.

Le règlement général de protection des données (RGPD) prévoit un certain nombre d'obligations de sécurité, mais également une obligation de minimisation des données. Ainsi, lorsque des données sont déposées sur le supercalculateur, seules doivent être incluses les données nécessaires à l'objectif de recherche. Le RGPD pose également l'obligation d'informer les personnes concernées ainsi que le délégué à la protection des données. Ces obligations s'appliquent aux responsables de traitement, mais également aux sous-traitants, dont GENCI, qui fournit le supercalculateur. Pour traiter les données à caractère personnel sur le supercalculateur, un contrat doit donc être signé entre le responsable de traitement et GENCI. Il n'est pas possible de recueillir les données sensibles sans le consentement éclairé de la personne concernée. Il est également nécessaire de définir les acteurs qui interviendront sur ces données.

Parmi les données sensibles, les données de santé sont régies par le Code de santé publique. Selon l'article L1111-8, toute donnée recueillie à l'occasion d'activités de prévention, de diagnostic, de soins ou de suivi social et médicosocial doit être hébergée sur un datacenter certifié hébergeur de données de santé. Dans le cas contraire, par exemple si les données sont acquises à des fins de recherche ou pseudonymisées, il n'est pas obligatoire de recourir à un hébergeur certifié. Il s'agit donc de s'assurer auprès des potentiels utilisateurs que les données ont bien été recueillies dans ce cadre.

L'utilisation du supercalculateur Jean Zay suppose d'abord de savoir si le projet va traiter des données à caractère personnel, dont des données de santé. Si tel est le cas, GENCI sera le sous-traitant et un contrat devra être conclu. Un responsable de traitement est systématiquement désigné. Dans le formulaire, les données à caractère personnel traitées doivent être spécifiées. Il est précisé s'il s'agit de données sensibles de santé, si elles ont été acquises à des fins de recherche et si elles ont été pseudonymisées. Si les données ont bien été acquises à des fins de recherche, le formulaire reçoit un accord et la sous-traitance sera formalisée. Tout au long du projet, le RGPD doit être respecté. Ainsi, si une personne demande le retrait ou la modification de ses données, l'utilisateur sera tenu de le faire. En cas de problème de sécurité, GENCI doit avertir le responsable de traitement sous 72 heures. Enfin, le centre organise la gestion de crise le cas échéant.

Vos données ne sont pas exposées. Chaque utilisateur dispose d'un espace dédié, auquel les autres utilisateurs n'ont pas accès. Pour certains projets, les données peuvent être mises en commun sur un espace. À tout moment, vous savez donc qui a accès à vos données. Vous avez notamment la possibilité d'intervenir sur toute intervention du support de l'IDRIS sur ces données.

PANORAMA LÉGISLATIF POUR LE PARTAGE DES DONNÉES DE LA RECHERCHE ET CAS D'USAGE

Table ronde

Panorama législatif pour le partage des données de la recherche et cas d'usage

Sylvie Rousset, Directrice de la DDOR, introduit cette table ronde qui réunit les intervenants suivants :

- **Adrien Boussaguet** - Responsable du pôle industrie et valorisation, SPV, CNRS
- **Gaëlle Bujan** - Déléguée à la protection des données, CNRS
- **Lionel Maurel** - DAS Science Ouverte, Edition scientifique et Données de Recherche, InSHS - CNRS
- **Stéphanie Rennes** - Juriste, INRAE
- **Jacques Van Helden** - Institut Français de Bioinformatique (IFB)

Adrien Boussaguet,
Responsable du pôle industrie et valorisation, SPV , CNRS

Je souhaite vous proposer un retour d'expérience des questions de science ouverte qui ont été rencontrées dans des services partenariats et valorisation, qui sont impactés à plusieurs titres par cette thématique.

Science ouverte et montage de projets

Dans le cadre de l'accompagnement des chercheurs dans le montage de projets, nous avons intégré dans les règles de subventionnement des deux principaux financeurs (ANR et Commission européenne) depuis quelques années, des contraintes assez importantes. La rédaction d'un DMP a été rendue obligatoire en 2017 par la Commission européenne et en 2019 par l'ANR. Ce DMP est à remettre dans les 6 premiers mois et à mettre à jour au milieu et à la fin des projets ANR, et selon un planning défini par le chercheur pour les projets européens. Pour les deux financeurs, les publications doivent être publiées dès le début sous la licence Creative Commons (CC-BY ou équivalent), et les données de recherche doivent être mises à disposition pour pouvoir être réutilisées (sauf à justifier d'une thématique innovante ou sensible). L'intervention du Service Partenariat et Valorisation se limite essentiellement au rappel de ces obligations et au partage d'une trame s'agissant du DMP. Il est arrivé que les collègues se sentent démunis lorsque des chercheurs leur demande de valider leur DMP. Tout au plus est-il possible de le comparer à d'autres et de vérifier si le document est complet. Des chargés d'affaires aux projets européens ont admis s'être formés aux questions de science ouverte (HAL, open access DMT) grâce à des tutoriels YouTube. L'INIST du CNRS a récemment réalisé des actions de sensibilisations sur ce sujet auprès des équipes et cela contribuera certainement à augmenter l'implication de notre service. Il serait pertinent d'y consacrer un module récurrent au sein du cycle de formation SPV à destination des nouveaux entrants, au vu du turn-over assez important dans certains services.

Science ouverte et publications scientifiques

Nous sommes parfois sollicités par des chercheurs dans le cadre de la négociation du contrat d'édition qui les lie avec un éditeur pour la publication dans une revue ou un ouvrage scientifique. La titularité des droits d'auteurs n'appartenant pas à l'employeur, et n'étant nous même pas forcément formés sur ces thématiques, il est difficile de les conseiller. Tout au plus peut-on les informer de leurs droits et sur ce que prévoit la loi République Numérique, mais notre accompagnement ne peut aller plus loin.

Science ouverte et valorisation

À l'issue de la recherche, lorsqu'une valorisation est possible sous la forme d'un dépôt de brevet ou d'un savoir-faire, il est crucial que les résultats restent confidentiels. Nous avons donc plutôt tendance à inviter le chercheur à la prudence quant au contenu de sa publication. Il n'est pas rare que des brevets ne puissent être attribués en raison de publications existantes (voire même dans un cas d'espèce, par le résumé du projet sur le site de l'ANR !), qui avaient détruit l'antériorité de l'invention. Pour la même raison, le chercheur peut se priver d'un financement de prématuration, si l'étude du projet par le comité de sélection révèle que le chercheur en a trop dit dans ses publications. Ceci est d'autant plus préjudiciable lorsqu'une création de start-up ou un concours scientifique sont envisagés. Pour le chercheur, cela révèle la contradiction supposée entre la démarche de libre diffusion et celle de la valorisation, encore qu'une définition précise de ces termes soient nécessaire. Concrètement, dans les cas que nous avons rencontré, si le DMP avait été bien rempli dès le début, cette problématique aurait pu être anticipée. Alors que les laboratoires y voient essentiellement une contrainte, il s'agit là d'une solution.

Science ouverte et partenariats

Lorsqu'on met en place un partenariat avec une entreprise, qui va financer une étude et collaborer avec un laboratoire, la logique de libre diffusion peut être renversée. L'intérêt de l'industriel est bien souvent d'améliorer son savoir-faire ou de pouvoir breveter pour exploiter, en gardant un maximum

▶▶▶

d'informations confidentielles. Dans nos contrats de collaboration, nous défendons la politique du CNRS qui est de protéger au maximum la liberté de publication de nos chercheurs, tout en respectant les contraintes économiques de nos partenaires. Cela passe par l'exclusion des résultats attendus de la recherche de la confidentialité, et par des clauses qui encadrent la manière dont un partenaire peut valider les publications des autres. A titre d'exemple, lorsque SNCF Voyageurs, à la veille de la mise en concurrence du rail, a souhaité renforcer les règles de confidentialité sur les données publiées par le doctorant concernant les statistiques de de fréquentation des trains par ligne (ces informations étant devenues extrêmement précieuses pour la SNCF), nous avons urgemment négocié un avenant pour trouver un compromis afin que les données publiées dans la thèse ne soient fournies que par axe, sans porter atteinte à la qualité scientifique de la production.

Science ouverte et compliance

Enfin, le dernier impact de la science ouverte sur les activités du SPV a trait à la compliance vis-à-vis des droits des tiers et des règles éthiques, notamment les droits de propriété intellectuelle ou les données personnelles. Il convient de se demander si, sur un contenu que le chercheur déposera sur une plateforme et tentera de valoriser, le droit d'auteur s'applique, des données patient apparaissent, etc. Des enregistrements audio de langues rares, par exemple, font intervenir du droit d'auteur, des données personnelles et du droit à la voix. Aux chercheurs dont la mission est d'alimenter une base de donnée ouverte répertoriant toute la richesse linguistique des cinq continents, nous avons remis des formulaires d'autorisation, qu'ils font signer quand ils le peuvent et qui les couvrent. La question s'est posée de savoir si, pour tous les enregistrements recueillis depuis plusieurs décennies, il était nécessaire de régulariser la situation. Une évaluation des risques doit être réalisée, et au cas par cas, il est nécessaire de faire preuve de souplesse.

Gaëlle Bujan,

Déléguée à la protection des données, CNRS

Les données personnelles bénéficient d'un cadre réglementaire très spécifique et dense : règlement européen sur la protection des données personnelles, Loi Informatique et liberté, Loi pour une République numérique, Code de la recherche, Code pénal, Code de la santé publique… Ces éléments juridiques encadrent l'utilisation des données. Une donnée personnelle correspond à toute information ou groupe d'informations permettant d'identifier ou de réidentifier directement ou indirectement une personne. Ces données peuvent être utilisées pour la recherche dans un cadre sécurisé. La personne doit en effet rester maîtresse de l'utilisation et de la réutilisation de ses données.

La recherche applique les principes généraux de l'utilisation des données personnelles, qui portent sur le fondement légal, en général pour la recherche au CNRS l'intérêt public, lié à la mission de recherche publique (décrite dans le décret du 24 novembre 1982 portant fonctionnement et organisation du CNRS), la finalité « recherche » de l'utilisation des données, la pertinence des données pour la recherche, la conservation des données (qui ne sont généralement utilisées que pour le temps de la recherche), la transparence de l'information sur l'utilisation des données (qui est particulièrement scrutée par la CNIL, et qui suppose que chacun connaisse l'utilisation qui est faite des données le ou la concernant) et la sécurisation et la protection des données.

Le RGPD a consacré la possibilité pour la recherche de traiter des données à caractère personnel. Nous nous sommes donc saisis de son article 89. À titre d'exception, nous avons en effet la possibilité, à des fins de recherches scientifiques, de réutiliser des données. Nous devons trouver l'articulation adéquate entre notre volonté de faire progresser la connaissance et l'utilisation des données personnelles en les protégeant. Dans cette question de la réutilisation des données, l'anticipation est particulièrement importante, puisque nous devons informer les personnes que nous sommes susceptibles de réutiliser leurs données à d'autres fins scientifiques que celle prévue initialement et qu'elles peuvent s'y opposer.

De plus, la réutilisation est possible uniquement si le traitement antérieur était licite et donc respectait le RGPD. Parmi les données réutilisables, les données anonymes, qui ne permettent pas d'identifier une personne, échappent au RGPD. Des jeux de données sont également pseudonymisés et peuvent être réutilisés, il peut s'agir des informations qui résultent des enquêtes diffusées par l'infrastructure de recherche Progedo, ou encore les ressources linguistiques dans Ortolang. Certaines données sont indirectement identifiantes et peuvent être accessibles via des entrepôts sécurisés, comme le centre d'accès sécurisé aux données (CASD) ou les entrepôts de données de santé (tels que l'entrepôt de l'AP-HP). Si la première collecte a respecté la réglementation, lors de la réutilisation, ces démarches de protection des données personnelles doivent à nouveau être suivies pour limiter les risques pour les personnes d'une exploitation illégale de leurs données.

Une question a été posée quant à l'accès permanent du CNRS au système national de données de santé (SNDS). La demande de compte utilisateur pour les chercheurs dans des équipes CNRS qui traite des données à des fins de recherche en santé est à formuler auprès du service protection des données, qui dispose d'une adresse dédiée (CNRS-SNDS@cnrs.fr).

Lionel Maurel,
DAS Science Ouverte, Edition scientifique et Données de Recherche, InSHS - CNRS

Stéphanie Rennes vous a parlé du principe d'ouverture par défaut, Gaëlle Bujan des exceptions à ce principe liées à la protection des données personnelles, et Adrien Boussaguet a évoqué les droits des tiers partenaires industriels, qui constituent une autre exception. Je reviendrai au principe d'ouverture des données, avec un focus particulier sur le sujet des licences. Les chercheurs doivent en effet choisir une licence pour la diffusion de leurs données. Dans la loi, l'application d'une licence n'est pas strictement nécessaire, même quand les données doivent être diffusées en open data. Un décret d'application contient une série de licences parmi lesquelles les administrations peuvent choisir pour préciser les conditions de réutilisation des données et rendre la réutilisation plus sécurisée.

Une chercheuse s'est adressée à l'INSHS pour obtenir des conseils quant aux choix d'une licence. Son projet, le Mobiliscope, est porté par le laboratoire Géographie-Cités. Il propose de la géovisualisation autour des questions de mobilité urbaine, permettant de visualiser les mouvements pendulaires autour d'une cinquantaine de villes en France, au Québec et en Amérique du Sud, s'appuyant sur des données issues d'enquêtes ayant lieu tous les 10 ans. Le projet a été lauréat du prix Science ouverte organisé par le ministère de l'Enseignement supérieur et de la Recherche, dans la catégorie « Créer les conditions de la réutilisation des données de recherche. Le Mobiliscope est un objet complexe, qui associe des données, des logiciels (dont la partie principale est produite en laboratoire) et des contenus graphiques éditoriaux. Ces différents objets n'ont pas tous la même nature, au sens juridique, et renvoient à des licences particulières à utiliser. Le premier choix opéré par les chercheurs portait sur le logiciel, en l'occurrence un logiciel sous licence libre AGPL. Celle-ci permet en effet la libre réutilisation, avec une condition de réciprocité. Ainsi, toute personne qui récupère ce code pour le rediffuser devra le faire sur la même licence. En ce qui concerne les données, la licence ODbL est l'équivalent de la licence GPL sur les bases de données. Pour les visuels (cartes, graphiques), qui correspondent à des œuvres de l'esprit protégées par le droit d'auteur, et qui ne peuvent se voir appliquer de licence de logiciel libre, nous avons conseillé aux chercheurs de retenir des licences creative commons, en particulier la licence CC-BY-SA, qui pose elle aussi une condition de réciprocité.

La chercheuse, en collectant les données, s'est rendue auprès de certains établissements publics qui ont tenté de lui imposer des conditions de réutilisation qui auraient été contraires à l'ouverture, et qui ne respectaient pas la loi pour une République numérique. Nous lui avons conseillé de leur demander d'appliquer la loi, ce qu'elle a fait. Ces établissements lui ont ensuite remis les données dans une forme permettant la réutilisation et l'ouverture. La loi pour une République numérique est donc un outil qui permet de rendre les données librement réutilisables.

Pour chaque couche du projet, il convient donc d'identifier la nature de l'objet à ouvrir, en fonction de laquelle des licences spécifiques s'appliquent. Certaines ouvrent totalement la réutilisation, à condition de citer la source. D'autres posent une condition de réciprocité. S'agissant des œuvres de l'esprit, ce qui est couvert par le droit d'auteur n'est pas soumis au principe d'ouverture par défaut.

Stéphanie Rennes,
Juriste, INRAE

Je vous remercie de cette invitation, qui me permet de partager avec mes collègues du CNRS un sujet passionnant, que j'introduirai par une présentation sur le principe général d'ouverture des données et ses exceptions.

S'agissant d'abord du cadre juridique national, un certain nombre de corpus de lois organisent l'accès aux données, à la fois les données publiques et les données de la recherche. L'accès aux données administratives remonte à 1978, dans un contexte de promotion de la transparence de l'action de l'administration. Dans les années 2015 et 2016, cette volonté de faciliter l'accès et les modalités de réutilisation des informations du secteur public a été accentuée, avec la loi dite Walter de décembre 2015, qui concerne les modalités de réutilisation des informations du secteur public et qui instaure un principe de gratuité à leur égard, et la loi du 7 octobre 2016 dite loi pour une République numérique, qui organise la circulation des savoirs dans un environnement numérique, l'ouverture des données publiques et des codes sources et met en place un service public de la donnée. Cette loi est codifiée dans près de 30 codes, parmi lesquels le Code de la recherche, le Code de la propriété intellectuelle et le Code des relations entre le public et l'administration (CRPA). Sur un plan plus opérationnel, des décrets sont venus régir certains pans de l'activité de recherche : par exemple, un décret portant sur l'intégrité scientifique (donnant une place centrale au plan de gestion des données) en 2021, ou encore un décret concernant la « fouille de données » en juin 2022.

Le cadre politique national se met également en place, avec une circulaire du Premier ministre (avril 2021) qui instaure une politique publique de la donnée, des algorithmes et des codes sources, et le plan national pour la science ouverte de 2018 puis 2021. Sur le plan européen, nous pouvons citer le règlement européen sur la gouvernance des données, ainsi que des directives, par exemple la directive sur le droit d'auteur et les droits voisins dans le marché unique numérique, qui a des impacts sur l'activité de recherche, et la directive sur les données ouvertes et la réutilisation des informations du secteur public. En France, les deux plans nationaux pour la science ouverte ont donc consolidé la place de la recherche publique dans le dispositif juridique d'ouverture des données publiques. Comme évoqué précédemment, l'accès aux données publiques et les modalités de leur libre réutilisation ont été organisés par plusieurs lois, aujourd'hui principalement transposées dans le Code des relations entre le public et l'administration (CRPA). Ces dispositions marquent un tournant majeur, puisqu'elles indiquent que sauf exception, les données produites ou reçues par une administration dans le cadre de sa mission de service public sont ouvertes par défaut et réutilisables gratuitement. Conformément à la mission de service public qui leur est confiée, et suivant les cadres juridiques et politiques précités, les établissements de recherche organisent l'accès aux données qu'ils détiennent selon le principe directeur « aussi ouvert que possible, aussi fermé que nécessaire ». Ces dispositifs concernent principalement les données publiques. Les établissements sont visés par les réglementations et garants de la mise en œuvre de l'open data, qui est un mode d'accès et de réutilisation des données qui coexiste avec d'autres modes, selon la nature des données. En pratique, le plan de gestion des données et les licences sont des outils incontournables de bonne gestion juridique et scientifique des données.

Le principe directeur « aussi ouvert que possible, aussi fermé que nécessaire » signifie qu'avant d'ouvrir l'accès aux données, il est indispensable de vérifier si les conditions d'ouverture sont réunies et si l'accès aux données est soumis à une réglementation spécifique ou à un protocole de mise à disposition particulier, voire à une interdiction pure et simple. Les personnes concernées sont les personnes publiques ou les personnes privées délégataires d'une mission de service public. Il existe également un critère d'effectif de 50 ETP minimum. Sont concernées les données en bases ou hors bases, les données produites ou reçues par l'établissement, les données achevées et les données qui ont un intérêt économique, social, sanitaire ou environnemental, peu importe leur date de constitution. En ce qui concerne le partage des données, il s'agit de vérifier l'existence d'un partenariat, d'un protocole de mise à disposition spécifique, ou des conditions de traitement préalable des données (par exemple, une occultation de certaines informations). Des exceptions à l'ouverture en open data sont prévues, dans des cas expressément prévus par les textes, à savoir des éléments dont la consultation ou la communication porterait atteinte, notamment, au secret de la défense nationale, à la conduite de la politique extérieure de la France et à des impératifs de sécurité publique, de sécurité des personnes et des systèmes d'information. Ces vérifications constituent des étapes préalables à la diffusion et à la réutilisation des données via une licence adaptée.

En conclusion, l'établissement organise l'accès aux données, en conformité avec les cadres juridiques et politiques en vigueur, suivant le principe directeur « aussi ouvert que possible, aussi fermé que nécessaire », qui se traduit par des étapes de vérification menant à l'ouverture, à un partage ou à la mise en œuvre d'une exception. Les pratiques et outils incontournables sont les principes FAIR, le plan de gestion des données et les licences. L'objectif est de faciliter la réutilisation des données au profit du public. Cette réutilisation est libre ou gratuite pour les données gratuites. Le réutilisateur doit s'engager à citer les sources, indiquer les dates de dernières mises à jour, à ne pas altérer les données et à ne pas les dénaturer.

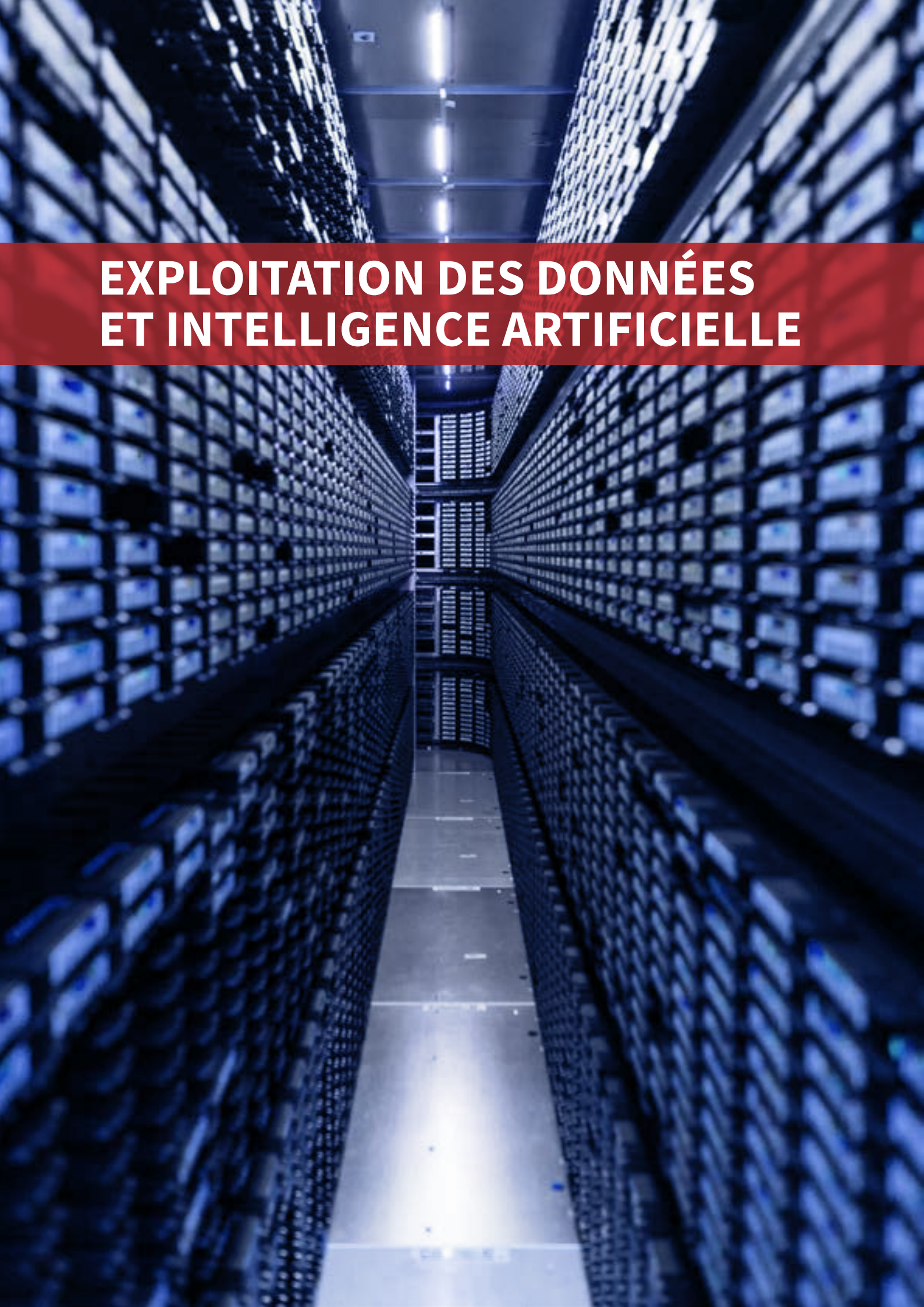
Jacques Van Helden,
Institut francais de bioinformatique (IFB)

Nous avons proposé de partager le cas d'étude relatif au partage, à la protection et à l'ouverture des données de séquençage du projet EMERGEN, qui est le projet national de séquençage du virus SARS-CoV-2. Ce projet, monté dans l'urgence, mobilise un grand nombre d'acteurs : 5 000 plateformes de prélèvement, réparties partout en France, 50 laboratoires dotés de gros équipements pour réaliser du séquençage. Il s'agissait de rassembler toutes ces données dans une base de données et y donner accès, pour la surveillance et la recherche. Santé publique France, le 27 janvier 2021, nous a contacté en nous demandant si l'Institut Français de Bioinformatique (IFB) pouvait monter une base de données et une infrastructure numérique pour accueillir ces données, les traiter et les mettre à disposition. Nous l'avons fait en deux à trois semaines, puis avons continué de progresser pendant deux ans, avec des outils de plus en plus élaborés.

L'analyse du flux de données nous a permis de caractériser les parties de données non sensibles, les parties de données non identifiantes mais relevant de la santé (conditions de prélèvement, par exemple), les données sensibles et les données identifiantes. Nous collectons toutes les données de séquence des virus, mais il nous faut également collecter les données de patients, que nous allons réappareiller avec les données de santé. D'une part, une partie de ce réappareillement se fait dans la bulle HDS (Hébergeur de Données de Santé) de Santé publique France, avec les données de vaccination, d'hôpital, de décès, etc. D'autre part, pour la recherche, un espace HDS est en cours de développement par la DSI de l'INSERM, l'IFD et le Health Data Hub, qui est le dépositaire des données du Système National des Données de Santé (SNDS).

Nous avons ainsi monté une base de données et réalisé une caractérisation très détaillée de toutes les métadonnées. Nous avons réalisé un référentiel des métadonnées, dans lequel nous avons caractérisé chaque champ. Nous avons également réalisé un mapping des bases de données internationales. Une de nos missions était en effet de déposer les données dans deux entrepôts internationaux, GISAID et ENA (*European Nucleotide Archive*).

Depuis le début du projet, les producteurs de données sont confrontés à des enjeux contradictoires, avec une double finalité des données : d'une part, la surveillance, qui justifie une mise à disposition sans délai, et d'autre part, la recherche. Au niveau international, la base de données GISAID permet le partage des données sous licence protective extrêmement restrictive. Elle est maîtrisée par une fondation privée, mais l'OMS (Organisation mondiale de la Santé) et les CDC (*Centers for Disease Control and Prevention*) demandent à tous les pays d'y déposer les données, car il s'agit de l'instrument des neurologues. Enfin, l'ENA est la base de données européennes ouverte. L'une des questions qui se posent est celle de la reconnaissance du producteur. Les laboratoires séquenceurs considèrent en effet que les données qu'ils produisent leur appartiennent, et que les prédateurs de données les publient à leur place. Ce débat peut être géré par le plan de gestion de données, qui peut constituer un instrument de pacification. Nous avons ainsi commencé à créer un PGD. Ceci a supposé de segmenter la donnée, afin de définir le degré de protection et la destination des données. La charte d'accès aux données n'est pas encore définie. La demande CNIL donne quant à elle lieu à un certain nombre d'échanges. Nous disposons à présent d'une ébauche de PGD. Pour le valider, nous devons impliquer tous les acteurs, producteurs de données et coordinateurs du projet (ANRS/MIE, Santé publique France).



EXPLOITATION DES DONNÉES ET INTELLIGENCE ARTIFICIELLE

Jean-Luc Parouty,
CNRS - SIMAP, Chargé de mission Calculs scientifiques

Ce sujet est absolument central, puisqu'il n'est pas possible de faire de l'intelligence artificielle sans données. Commençons par présenter l'histoire du *deep learning*.

En 2007, Jim Gray a décrit quatre phases dans l'histoire de la science : la science empirique, qui a permis de construire des cathédrales, la modélisation, par des équations, qui a permis de construire des objets plus complexes, comme les voitures, l'utilisation de la puissance de calcul par des outils numériques, qui a permis de construire des objets encore plus complexes, et enfin l'extraction de l'intelligence issue des données, qui caractérise une science pilotée par les données. L'intelligence artificielle représente un grand ensemble, qui comporte un sous-ensemble, qui regroupe ce qui se fait par apprentissage. Au sein de ce sous-domaine, l'utilisation de réseaux de neurones artificiels caractérise l'apprentissage profond (*deep learning*).

Au sein du *machine learning*, deux grandes familles d'apprentissage peuvent être distinguées. La première correspond à l'apprentissage supervisé, qui permet d'apprendre à partir d'exemples. À titre d'illustration, un apprentissage est réalisé à partir d'images de chat et de lapin puis, lorsque l'image d'un animal est présentée, l'intelligence artificielle essaie de la catégoriser. L'apprentissage se fait ainsi à partir de données connues. Par opposition, l'apprentissage non supervisé consiste à utiliser des données qui ne sont pas annotées. Cette technique est efficace et présente l'avantage d'être automatique. Une autre grande famille d'usages, du côté de l'apprentissage non supervisé, est constituée par les réductions de dimensions. Il existe dans ce panorama de nombreux types d'algorithmes. Le *deep learning* s'applique pratiquement à tous ces domaines.

Une question centrale, en matière d'intelligence artificielle, réside dans la définition de l'intelligence, qui est cruciale dans l'avènement de ces réseaux de neurones. Selon la définition classique, il s'agit de « la capacité à percevoir ou d'inférer l'information et de la conserver comme une connaissance à appliquer à des comportements adaptatifs dans un environnement ou un contexte donné ». Il existe d'autres définitions ; le Petit Larousse définit l'intelligence comme « l'ensemble des fonctions mentales ayant pour objet la connaissance conceptuelle et rationnelle ». Ces deux définitions conduisent à un affrontement entre deux conceptions de l'intelligence, donc de l'intelligence artificielle. D'une part, l'approche connexionniste consiste à considérer que l'intelligence est le résultat de briques élémentaires que sont les neurones. Ceux-ci sont caractérisés par l'entrée de signaux, qui sont intégrés puis ressortent, en interconnexion avec d'autres neurones, constituant ainsi un grand réseau, ces neurones ayant la capacité de se forger durant l'apprentissage. Les neurones artificiels fonctionnent ainsi. D'autre part, l'approche dite symbolique, beaucoup plus conceptuelle, considère que le processus de la pensée est le résultat de concepts de haut niveau, avec un moteur de référence. Ces deux approches fonctionnent et sont pertinentes. Dans le premier cas, l'approche est inductive. Elle repose sur une entrée et une sortie. À partir d'un ensemble d'observations, un programme est fabriqué, qui reproduit ce modèle. Le modèle est donc le résultat de l'analyse des données, sans intervention de l'intelligence humaine dans l'écriture du programme. Dans l'autre cas, il s'agit d'observer les données et de recourir à des experts dans tous les domaines, qui utilisent leur expertise pour écrire le programme.

Un très bel article de Cardon, Cointet et Mazières de 2018 (La revanche des neurones) dresse l'histoire de l'affrontement de ces deux approches, en exposant le ratio de publications connexionnistes et symboliques à travers le temps. Initialement, dans les années 1940, l'approche était fortement connexionniste, en l'absence d'informaticiens. La première production fut Perceptron, une machine à un neurone, en 1975. L'informatique est ensuite apparue. Dans les années 1960, les ordinateurs ont pris de la puissance. De grandes pointures (Minsky, Papert, Simon, MacCarthy) ont imaginé le concept d'intelligence artificielle et mobilisé tous les budgets des grandes universités. Le Perceptron, qui n'était capable de traiter que des problèmes strictement linéaires, a décliné. Ces acteurs ont annoncé de grands accomplissements : la traduction du russe en américain, la production d'avions autonomes, etc. Ceux-ci n'ont cependant pas pu aboutir suffisamment vite, et leurs recherches ont ainsi été remises en cause. À partir du début des années 1970, un assèchement des recherches dans le domaine de l'intelligence artificielle s'est ainsi fait jour. Nous parlons ici d'un premier hiver de l'intelligence artificielle, où tous les financements se sont taris. Cependant, l'informatique a continué de se développer. Les systèmes experts ont en outre livré de très bons résultats dans un grand nombre de domaines. Quelques équipes de recherche ont également continué de travailler sur les réseaux de neurones. Deux avancées majeures sont intervenues : en 1986, la rétropropagation, avec Rumelhart, qui permet de mettre les neurones en réseau, et ainsi de traiter des problèmes non linéaires, puis en 1989, les réseaux convolutifs (avec Yann LeCun), à savoir la capacité pour des réseaux de neurones de travailler sur des images de façon très efficace. Une autre approche est ensuite apparue, plus rigoureuse et portée par des mathématiciens : l'approche SVM (*Support Vector Machine*), qui a trusté le domaine de la

►►►

recherche, avec des résultats bien plus performants que les réseaux de neurones et les systèmes experts, conduisant au deuxième hiver de l'intelligence artificielle pour les réseaux de neurones.

La progression des puissances de calcul en 50 ans fait apparaître six ordres de grandeur. S'agissant des datasets, ceux-ci sont passés des laboratoires aux images du monde réel. Les approches mathématiques de type SVM n'ont plus été capables de faire face à cette complexité. Un point de bascule symbolique est intervenu en 2012, avec la compétition ImageNet. Un doctorant s'est présenté avec un réseau de neurones, AlexNet, et a divisé par deux le taux d'erreurs. Ce point de rupture s'est propagé et traduit dans tous les domaines de l'intelligence artificielle.

Confrontée au monde réel, la complexité est devenue telle que les approches classiques ont atteint leurs limites. Les réseaux de neurones, de par leur capacité à appréhender de très grands degrés de complexité, sont parvenus à apporter des réponses.

La progression des tailles de modèles fait apparaître une envolée dans les années 2010. Nous avons aujourd'hui des modèles d'une complexité extrêmement importante. L'apprentissage des plus gros modèles coûte des millions de dollars. La taille des datasets suit la même tendance, à savoir une croissance très importante. La tentation est de reproduire un cerveau de type humain, en intégrant des problématiques de plus en plus complexes. Du point de vue du dataset, elle consiste à faire du dataset notre environnement. La question est de savoir si une IA qui scanne internet pour reconnaître des images et du texte devrait payer des droits, et à qui. La Grande-Bretagne a changé sa législation il y a peu afin qu'une IA qui apprend à partir de données publiques ne donne pas lieu au paiement de droits.

En conclusion, l'évolution des paradigmes a conduit à une approche inductive, au cœur de laquelle les données sont centrales. Le modèle qui s'obtient en sortie ne peut en conséquence s'expliquer que par les données utilisées en entrée. Ceci pose des contraintes de qualité, de disponibilité, de transparence, etc. Un escargot a 10 000 neurones, un éléphant 250 milliards, et un être humain 100 milliards. Le corvidé représente une anomalie intéressante, puisqu'il dispose de deux milliards de neurones.

DÉFINITIONS

Data centre
Les data centres (en anglais Data center) sont des infrastructures matérielles (béton, alimentation électrique, climatisation, connexions réseaux, gardiennage, etc.) destinées à héberger systèmes d'information, moyens de calcul et infrastructures de stockage de données. Les data centres peuvent fournir un « hébergement sec » (les moyens sont opérés à distance par leurs propriétaires) ou un certain nombre de services assurés par son personnel (gestes de proximité, exploitation, accès à des espaces de stockage ou des logiciels, etc).

Entrepôt de données
Un entrepôt de données (en anglais Data Repository) est un service en ligne permettant le dépôt, la curation, la description, la conservation, le référencement, la recherche et la diffusion de jeux de données de recherche. Il peut être générique (plateforme RDG) ou thématique (spécifique à une discipline ou un groupe de disciplines).

Mésocentre
Un mésocentre est un ensemble de ressources humaines, matérielles et logicielles qui fournit à la communauté scientifique d'une même région un environnement scientifique et technique propice au calcul et au traitement de données.

Stockage
Le stockage (en anglais Data Storage) est une opération qui consiste à déposer les données sur un support numérique qui peut être un ordinateur personnel, un disque partagé, une bande magnétique, un espace de stockage dans un data centre ou tout organe de dépôt, pour permettre leur exploitation à court terme.
Ce stockage est généralement destiné à l'usage personnel du chercheur et, éventuellement, de son équipe et de ses collaborateurs et ne s'accompagne pas nécessairement d'une mise à disposition ou du référencement des données, du ressort d'un entrepôt.

NOTES ET RÉFÉRENCES

1 **Article processing charges (APC)**
Paiement effectué par un auteur ou son institution pour financer la publication d'un document scientifique dont la consultation est ensuite gratuite pour le lecteur, dès la publication. Les frais de publication en accès ouvert appliqués à la recherche universitaire sont généralement coûteux, ce qui limite la publication en accès ouvert pour les institutions, les universitaires et les étudiants moins fortunés. Le système d'accès ouvert basé sur les APC (système auteur-payeur) fait partie, parmi d'autres controverses, du débat éthique global plus vaste sur l'accès ouvert. Un des problèmes est que si un éditeur réalise des bénéfices en acceptant des articles, il est incité à accepter tout ce qui lui est soumis, au lieu de sélectionner et d'éventuellement rejeter des articles en fonction de leur qualité.

2 **Centre Mersenne** | https://www.centre-mersenne.org
Le centre Mersenne est une infrastructure publique d'édition au service de la communauté scientifique. Elle vise à promouvoir l'édition scientifique et la diffusion de publications (revues, livres, séminaires et colloques) de toutes disciplines scientifiques (mathématiques, physique, statistiques, informatique…), nationales et internationales, engagées dans le libre accès et publiant principalement en LaTeX..

3 **Interview** d'Alain Schuhl « Le CNRS encourage ses scientifiques à ne plus payer pour être publiés », CNRS Info, 7 avril 2022 | https://www.cnrs.fr/fr/cnrsinfo/le-cnrs-encourage-ses-scientifiques-ne-plus-payer-pour-etre-publies

4 **Peer Community In (PCI)** | https://peercommunityin.org
Peer Community In est un processus d'évaluation par les pairs qui jouent le rôle d'éditeurs et émettent des recommandations sur des articles de prépublication ou des articles.

5 **OpenEdition** | https://www.openedition.org
L'infrastructure OpenEdition est un service d'édition numérique pour la communication scientifique en sciences humaines et sociales. Elle rassemble quatre plateformes de publication et d'information en accès ouvert : revues, livres, carnets de recherche et annonces d'événements académiques internationaux.
Elle rassemble quatre plateformes de publication et d'information en accès ouvert : revues, livres, carnets de recherche et annonces d'événements académiques internationaux.

6 **Comptes rendus Académie des Sciences** | https://comptes-rendus.academie-sciences.fr

7 **Interview** d'Alain Schuhl « Il n'y a pas de raison que les scientifiques fassent une cession exclusive gratuite de leurs oeuvres aux éditeurs», CNRS Info, 1er décembre 2022 | https://www.cnrs.fr/fr/cnrsinfo/il-ny-pas-de-raison-que-les-scientifiques-fassent-une-cession-exclusive-gratuite-de-leurs

8 **CPCN (Conférence des Présidents des Sections du Comité national)** | https://www.cnrs.fr/comitenational/struc_coord/cpcn.htm
La CPCN représente les sections et les commissions interdisciplinaires du Comité national auprès des diverses instances décisionnelles ou consultatives, intérieures ou extérieures au CNRS.

9 **Appel de Paris sur l'évaluation de la recherche** (OSEC 2022) | https://osec2022.eu/fr/appel-de-paris/
Ce texte a été préparé par le Comité pour la science ouverte et présenté aux Journées européennes de la science ouverte (Paris Open Science European Conference – OSEC 2022) à Paris les 4 et 5 février 2022, organisées dans le cadre de la Présidence française du Conseil de l'Union européenne, suite à la publication de la recommandation de l'UNESCO sur la Science ouverte et à la publication par la Commission européenne du texte « Vers une réforme du système d'évaluation de la recherche ».

10 **Coalition on Advancing Research Assessment (CoARA)** | https://coara.eu/

11 **Plan données de la recherche du CNRS** (2020) | https://www.science-ouverte.cnrs.fr/wp-content/uploads/2021/01/Plaquette-Plan-Donnees-Recherche-CNRS_nov2020.pdf

12 **CNRS Données de la Rercherche** | https://cat.opidor.fr/index.php/CNRS_Donn%C3%A9es_de_la_Recherche_:_Catalogue_des_entrep%C3%B4ts_et_des_services

13 **OPIDor** (Optimiser le Partage et l'Interopérabilité des DOnnées de la Recherche) | https://opidor.fr/
Le portail OPIDoR met à disposition de la communauté de l'Enseignement Supérieur et de la Recherche un ensemble d'outils et de services facilitant la mise en application des principes FAIR pour rendre les données Faciles à trouver, Accessibles, Interopérables et Réutilisables. Ce portail est mis en place par l'Inist-CNRS.

14 DoraNum | https://doranum.fr/
DoRANum est une plateforme de formation en ligne sur la gestion et le partage des données de la recherche selon les principes FAIR (Facile à trouver, Accessible, Interopérable et Réutilisable), réalisée par l'Inist-CNRS et le GIS « Réseau Urfist » depuis 2015.

15 DataCite | https://datacite.org/
DataCite est une association à but non lucratif dont la mission est d'attribuer des identifiants numériques pérennes, appelés DOI (Digital object identifier, en français « identifiant numérique d'objet ») aux produits de la recherche dans leur définition la plus large.

16 CNRS Research Data | https://entrepot.recherche.data.gouv.fr/dataverse/cnrs
Depuis juin 2023, le CNRS a ouvert un entrepôt générique institutionnel pour partager les données de la recherche à destination des communautés scientifiques qui ne disposent pas d'un entrepôt thématique plus adapté, au sein de la plateforme nationale Recherche Data Gouv,

17 Research Data Gouv | https://recherche.data.gouv.fr/fr

18 Centre de données astronomiques de Strasbourg | https://cds.u-strasbg.fr

19 International Virtual Observatory Alliance | https://ivoa.net/

20 Projet SKA | https://www.skao.int/en

21 ForM@Ter | https://www.poleterresolide.fr
Le pôle Terre solide a pour objectif de faciliter l'accès aux données et contribuer à la création de nouveaux produits et services, en apportant de la valeur ajoutée aux données spatiales et « in-situ » disponibles. Il s'inscrit dans les paysages national et européen en articulation étroite avec les infrastructures en place et en construction. ForM@Ter est conduit dans le cadre de la mise en place de pôles de données nationaux impulsée par le CNRS et le CNES pour les thématiques Terre solide, Océan, Atmosphère et Surfaces continentales./

22 Data Terra | https://www.data-terra.org/

23 Pôle National de Données de Biodiversité (PNDB) | https://www.pndb.fr/

24 SEANOE (SEA scieNtific Open data Edition) | https://www.seanoe.org/html/about.htm

25 GO FAIR | https://www.go-fair.org/

26 BRGM | https://www.brgm.fr/fr

27 GEoNetwork | https://geonetwork-opensource.org/

28 Software Heritage | https://www.softwareheritage.org/?lang=fr

29 Réseau sismologique et géodésique français | https://www.resif.fr/

30 GAIA Data | https://www.gaia-data.org/

31 CLIMERI France | https://climeri-france.fr/

32 HumaNum | https://www.huma-num.fr/

33 CINES | https://www.cines.fr/

34 Nakala | https://www.nakala.fr/

35 OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) | https://www.openarchives.org/pmh/
L'Open Archives Initiative (OAI) a développé et promu des normes d'interopérabilité pour faciliter la diffusion efficace des contenus. L'OAI trouve son origine dans les mouvements de libre accès et de dépôt institutionnel. Au fil du temps, cependant, le travail de l'OAI s'est étendu à la promotion d'un large accès aux ressources numériques pour la recherche en ligne, l'apprentissage en ligne et la science en ligne.

36 RDF (Resource Description Framework) | https://www.w3.org/RDF/
Resource Description Framework (RDF) est un modèle de graphe destiné à décrire formellement les ressources Web et leurs métadonnées, afin de permettre le traitement automatique de telles descriptions. Développé par le W3C, RDF est le langage de base du Web sémantique.

37 EquipEx COMMONS | https://www.cnrs.fr/fr/cnrsinfo/un-financement-pour-lequipex-en-sciences-humaines-et-sociales-commons
Consortium de moyens mutualisés pour des services et données ouvertes en SHS (projet COMMONS) visant à développer la dynamique de science ouverte dans ces disciplines.

38 ELIXIR | https://elixir-europe.org/
ELIXIR est une organisation intergouvernementale qui rassemble des ressources en sciences de la vie provenant de toute l'Europe. Ces ressources comprennent des bases de données, des outils logiciels, du matériel de formation, du stockage en nuage et des superordinateurs. (source Elixir)

39 Core Data Resources | https://elixir-europe.org/platforms/data/core-data-resources
ELIXIR Core Data Resources est un ensemble de ressources de données européennes d'une importance fondamentale pour la communauté des sciences de la vie au sens large et pour la préservation à long terme des données biologiques.(source Elixir)

40 European Strategy Forum on Research Infrastructures (ESFRI) | https://www.esfri.eu/

41 Common European Data Spaces | https://dataspaces.info/common-european-data-spaces/#page-content

42 Commission nationale de l'informatique et des libertés (CNIL)| https://www.cnil.fr

43 Système National des Données de Santé (SNDS) | https://www.snds.gouv.fr/SNDS/Accueil
Instauré par l'article 193 de la loi de modernisation de notre système de santé de janvier 2016, le SNDS est géré par par la Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés (CNAMTS) et a pour objectif de chaîner : les données de l'Assurance Maladie (base SNIIRAM) ; les données des hôpitaux (base PMSI) ; les causes médicales de décès (base du CépiDC de l'Inserm) ; les données relatives au handicap (en provenance des MDPH - données de la CNSA) ; un échantillon de données en provenance des organismes d'Assurance Maladie complémentaire

44 SNIIRAM | https://www.snds.gouv.fr/SNDS/Composantes-du-SNDS
Le SNIIRAM, géré par la CNAMTS, a été créé pour contribuer à la connaissance des dépenses de l'ensemble des régimes d'Assurance Maladie, à la définition, à la mise en œuvre et à l'évaluation des politiques de santé, à l'amélioration de la qualité des soins et à la transmission aux professionnels de santé des informations relatives à leur activité, à leurs recettes et, s'il y a lieu, à leurs prescriptions.

45 High performance computing ou HPC - Calcul intensif | https://www.science-ouverte.cnrs.fr/calcul-intensif/
Le calcul intensif – High Performance Computing (HPC) – consiste à utiliser des super-calculateurs et leur environnement logiciel associé. Des systèmes capables d'effectuer plusieurs centaines de milliards d'opérations de calcul par seconde, en mettant en œuvre des traitements informatiques pour résoudre de façon efficace et dans des temps raisonnables des problèmes complexes issus de la recherche et de l'industrie.

46 Genci | https://www.genci.fr
GENCI est en charge de mettre à disposition des moyens de calcul performants. Il a pour mission, au niveau national et européen, de favoriser l'usage du calcul intensif et de l'intelligence artificielle au bénéfice des communautés de recherche académique et industrielle.

47 Equipex MESONET | https://www.mesonet.fr/
MesoNET est un réseau de mésocentres. L'objectif premier est de mettre en place une infrastructure régionale distribuée, en intégrant au moins un mésocentre par région, institués comme références et relais régionaux. L'infrastructure, intégrée à l'initiative European Open Science Cloud (EOSC), devrait avoir un impact significatif sur l'appropriation par les chercheurs des infrastructures numériques et IA publiques nationales et régionales.

CNRS
DIRECTION DES DONNÉES OUVERTES DE LA RECHERCHE - DDOR

3, rue Michel-Ange 75016 Paris
www.cnrs.fr