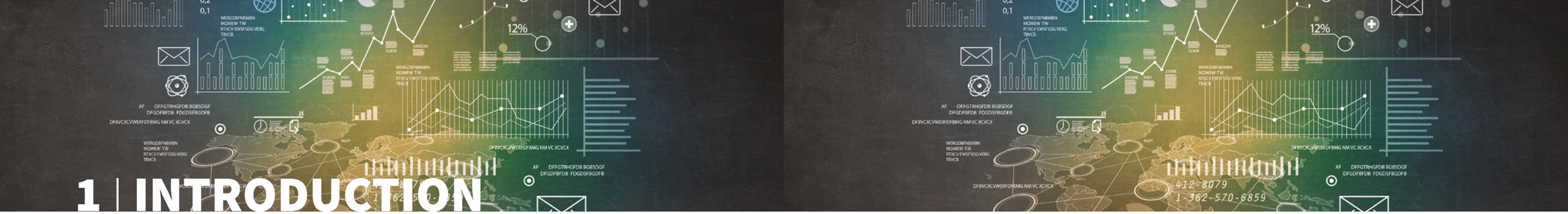
PLAN DONNÉES DE LA RECHERCHE DU CNRS

Novembre 2020

TABLE DES MATIÈRES



1 | INTRODUCTION

Ce plan s'inscrit dans la nécessité d'accélérer le développement vers la Science Ouverte. L'ensemble des données issues de la recherche au sens large telles qu'elles sont définies dans le Plan National de la Science Ouverte[1] (PNSO), est ainsi concerné par ce plan.

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. L'ouverture des données s'entend selon l'expression de la communauté européenne « ouvert autant que possible, fermé autant que nécessaire ». La décision d'ouverture ou de protection des données de la recherche doit être prise avec les services compétents du CNRS : les Services partenariat et valorisation pour la propriété intellectuelle, la Délégation à la protection des données pour les données à caractère personnel et la Direction de la sûreté pour les questions relatives à la souveraineté. Ainsi, ce plan et les actions qu'il propose traitent des données ayant vocation à être ouvertes, que ce soient les données brutes ou retraitées dans tous leurs formats, les textes et documents, également les logiciels, les algorithmes, les protocoles et les « *workflows* ». Ce plan inclut l'écosystème des données, c'est-à-dire les aspects liés aux infrastructures numériques ainsi que l'ensemble des services, en particulier dans le contexte de la mise en œuvre des principes FAIR[2] (Faciles à trouver, accessibles, interopérables et réutilisables).

La diversité des sujets autour des données de la recherche est illustrée par leur cycle de vie, constitué de 6 étapes :

- création ou collecte
- traitement
- analyse
- préservation
- partage
- réutilisation

Il est important de comprendre chacune de ces étapes afin de mettre en place une gestion adéquate des données de la recherche tout au long de ce cycle. Celui-ci comprend plusieurs niveaux d'intervention (par exemple la production des données y compris par le calcul, le stockage des données, la FAIRisation des données impliquant le traitement et la conservation des données).

La question du stockage (transitoire, long terme) est différente de celle de l'archivage. Le stockage fait référence à la persistance des données, leur identification, leur indexation et l'optimisation de leur accès en vue de traitements fréquents et intensifs. L'archivage fait référence à la conservation des données pour des raisons légales ou historiques. Les technologies de stockage et d'archivage sont généralement différentes. Même si le matériel peut être le même, la rigueur et le systématisme des procédures ne sont pas les mêmes.

Il existe de multiples façons de caractériser les données de la recherche, en fonction de leur type, de leur mode de production, des différents stades de leur transformation et des différents modes de présentation à l'utilisateur. Parmi les types de données, on distingue les données numériques ou symboliques, les textes, les images, les graphes, les sons, etc. Les données sont le résultat d'observations, de calcul ou de transformations diverses. Les données peuvent être à un stade brut, curé, intégré ou agrégé. Ceci comprend les données traitées et les produits dérivés de données, les données inclues dans les publications ou constituant des publications spécifiques, appelés « *data papers* » en anglais.

Il est nécessaire de rendre les données de la recherche réutilisables, voire partageables, afin que d'autres personnes, qui ne sont pas à l'origine de leur création, puissent les réutiliser.

Notons une différence de définition en fonction des communautés sur les termes de « stockage des données de la recherche » et d'« entrepôt des données de la recherche ». Pour certains, entreposer ses données, par exemple sur un disque dur de laboratoire, ne signifie pas nécessairement d'avoir fait un traitement des données, ou la création de métadonnées associées qui permettent à d'autres personnes de réutiliser ces données. Pour d'autres communautés (notamment celle de la science ouverte), l'entrepôt des données peut être défini comme un service en ligne, destiné à gérer la description d'ensemble des données en vue de leur préservation et de leur réutilisation. Dans cette définition, l'entrepôt des données de la recherche ne comprend pas l'espace physique ou l'infrastructure matérielle qui contient les données. Nous retiendrons cette définition dans ce document.

L'archivage et la curation des données, plus largement la gestion et le partage des données, offrent de nombreux intérêts qui ne sont pas rappelés ici de façon exhaustive. La mise à disposition des données attachées à une publication scientifique est un élément constitutif de la compréhension et de la validation de résultats scientifiques ainsi qu'une base à la reproductibilité des processus qui ont conduit à ces résultats. Par ailleurs, réutiliser des données existantes plutôt que refaire les calculs ou les expériences qui les ont produites, permet un meilleur usage de l'investissement public. Enfin, de nouvelles connaissances peuvent émerger du croisement de données issues de communautés très différentes et peuvent conduire à des thématiques de recherche originales, à condition d'en assurer le partage dans un contexte de recherche, qui implique un niveau de qualité, de contextualisation, voire de validation par les pairs.

Ce plan s'appuie sur les réflexions qui ont conduit à la rédaction en janvier 2018 d'un livre blanc des données au CNRS[3]. Une analyse détaillée, institut par institut, a alors conclu qu'il était urgent de promouvoir une véritable « culture de la donnée », de doter le CNRS d'une stratégie forte pour répondre aux besoins des communautés en matière de plateformes pour l'analyse de données à grande échelle et de mettre en place une politique de gestion, de valorisation et de pérennisation des données.

1. PNSO, 4 juillet 2018 : https://www.enseignementsup-recherche.gouv.fr/cid132529/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html
2. *Turning FAIR into reality, Final report and action plan from the European Commission expert group on FAIR data, European commission* (2018) : https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf
3. *Le livre blanc sur les données au CNRS, État des lieux et pratiques*, Mission Calcul – Données, janvier 2018 : http://www.cocin.cnrs.fr/IMG/pdf/livre_blanc_donne_es_2018.pdf

2 | PRINCIPES GÉNÉRAUX

Avec la loi numérique de 2016 dite Loi Lemaire[1], le cadre légal des données de la recherche produites dans un contexte de financement sur fonds publics a évolué vers un principe de mise à disposition en données ouvertes, mais il reste relativement méconnu d'un grand nombre de scientifiques.

Au-delà des obligations légales, la feuille de route du CNRS pour la science ouverte[2], adoptée en novembre 2019, à la suite des objectifs fixés par l'Europe et par le MESRI en matière de science ouverte, comprend d'ores et déjà un volet « données de la recherche ». Elle fait de la mise à disposition des données de la recherche, l'un des ressorts des avancées futures de la connaissance.

L'explosion du volume et de la diversité des données de la recherche, les développements méthodologiques du Big Data et les nouvelles possibilités d'analyse offertes par l'Intelligence Artificielle (AI), impliquent que le CNRS se dote, en lien avec ses partenaires, d'une politique et d'une stratégie prenant en compte les principes FAIR et le principe d'ouverture des données (sauf exceptions légitimes).

Le CNRS, qui est un des acteurs européens majeurs de la production des données de la recherche, doit développer une stratégie et une politique volontaristes et lisibles, ainsi que des services de données de qualité au service de la recherche. La démarche proposée se situe dans le cadre de l'initiative européenne *European Open Science Cloud* (EOSC)[3]. Elle correspond naturellement à promouvoir un « EOSC français », qui sera une contribution et la porte d'entrée logique à l'initiative européenne. Elle s'inscrit également au niveau international au travers de l'implication du CNRS dans des très grands instruments, systèmes d'observation et infrastructures de données.

Le Plan données de la recherche est avant tout piloté par les besoins des scientifiques et prendra en compte les contextes disciplinaires. L'articulation entre le travail réalisé dans les instituts et une gouvernance transverse des données, est un point clé et stratégique pour la réussite de la mise en œuvre du Plan données de la recherche. Cela comprend plusieurs éléments :

- le travail des instituts, qui prennent en compte la diversité des approches et des communautés et développent des stratégies nationales pour organiser leur communautés autour des données ;
- la volonté de faire bénéficier les communautés les moins avancées, des connaissances et de l'expérience des plus avancées sans pénaliser ces dernières ;
- le « ruissellement » des communautés dont la donnée est un objet de recherche vers les communautés « utilisatrices ».

Le plan d'actions devra prendre en compte la nécessité de faire évoluer les pratiques et les mentalités tout autant que le développement des outils adéquats pour la gestion, le partage, la préservation à long terme et la diffusion des données de recherche en conformité avec les principes FAIR. La question spécifique des moyens, notamment les moyens humains ainsi qu'un plan de formation, font partie intégrante du Plan données de la recherche.

1. Loi pour une République numérique (2016) : www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746/
2. La feuille de route Science Ouverte du CNRS : https://www.science-ouverte.cnrs.fr/wp-content/uploads/2019/11/Plaquette_Science-Ouverte_18112019.pdf
3. EOSC : https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

3 | OBJECTIFS

L'indispensable évolution des pratiques et des mentalités relève de l'organisation générale des activités de recherche, au niveau individuel, mais aussi et surtout au niveau collectif. Les initiatives dans ce sens, portées par le CNRS, les instituts, les communautés scientifiques ou les infrastructures de recherche, doivent inciter toutes les structures de recherche à se doter de « politiques des données ». Le Plan données de la recherche du CNRS s'articulera ainsi autour de trois objectifs principaux :

• *Diffuser une culture de données FAIR*

L'objectif global est d'adopter de bonnes pratiques en conformité avec les principes FAIR pour accélérer l'accès et le croisement des données et leur utilisation/réutilisation par les humains et par les machines, contribuer à leur fiabilisation (qualité et véracité des données) et à la reproductibilité des résultats de recherche. Dès la conception des projets de recherche, le cycle de vie des données doit être pris en compte : de leur production à leur libre mise à disposition quand cela est possible (données ouvertes ou à usage restreint), en passant par leur stockage, leur curation, leur analyse et modélisation et l'archivage intermédiaire voire à long terme. Il est clair cependant que la qualité et la reproductibilité des données dépendent aussi d'autres facteurs qui ne sont pas considérés ici, tels que la qualité des sources ou des processus d'analyse. Afin d'accompagner cette diffusion, l'offre de formation doit être complétée tant la demande est croissante.

• *Faire connaître les services et les outils existants*

Services, « *workflows* », entrepôt des données, portails de référencement et d'accès, technologies, normes/standards, etc. qui facilitent la prise en charge par les communautés, des données de recherche tout au long de leur cycle de vie et la mise en œuvre des principes FAIR. Partant du constat que toutes les communautés de recherche n'ont pas atteint le même degré de maturité dans cette prise en charge et in fine dans leur mise à disposition pour l'ensemble de la communauté, le Plan données de la recherche du CNRS doit servir de levier à chaque fois que cela est possible.

• *Accompagner la création de nouvelles pratiques, de nouveaux services et de nouveaux outils*

L'archivage et la curation des données nécessitent une réflexion en amont afin de savoir quelles données conserver, et comment contrôler leur intégrité, leur qualité et leur véracité. Le CNRS pourrait également soutenir et mener des expériences pilotes pour l'organisation de plateformes distribuées de stockage, d'indexation et de curation de données, avec un ensemble de services fondamentaux.

Seconde salle informatique du Centre de calcul de l'Institut de physique nucléaire et de physique des particules (CC-IN2P3).
© Cyril FRESILLON / CC IN2P3 / CNRS Photothèque

4 | POLITIQUE DES DONNÉES

La politique des données du CNRS doit être explicite et en phase avec les besoins de la recherche et des scientifiques. Les communautés scientifiques ne sont pas toutes au même degré de maturité. Le partage des données est une pratique déjà largement développée dans certaines disciplines ; ainsi des entrepôts disciplinaires, le plus souvent internationaux, se sont développés depuis plusieurs décennies notamment en astronomie. D'autres communautés n'ont pas encore de dispositif pour la prise en charge de leurs données, ni même de réflexions poussées sur le sujet.

Actuellement, nombre de données sont irrémédiablement perdues au cours du temps alors qu'elles pourraient être réutilisées. Le Plan données de la recherche du CNRS doit s'appuyer chaque fois que cela est possible sur l'existant. Ainsi la communauté des astronomes dispose depuis 1972 d'un centre au service du partage des données, le Centre de données astronomiques de Strasbourg[1]. Plus récemment, les SHS, via la TGIR Huma-Num[2], ont mis en place un outil de prise en charge des données, de leur production à leur archivage à long terme le cas échéant, et à leur diffusion.

Prenant en compte les principes FAIR et la nécessité de l'ouverture des données, la politique des données du CNRS devra inclure les points suivants :

- Le soutien aux communautés pour les inciter à FAIRiser leurs données très en amont, ou à défaut lorsqu'elles s'apprêtent à les stocker ou à les publier. La question de la gestion des données n'est pas seulement le volume de stockage nécessaire, mais aussi la façon dont les données sont traitées (curées, enrichies, structurées) et identifiées avec les services permettant de pouvoir les retrouver, les réutiliser et les partager. L'évaluation des pratiques de la recherche incorporera progressivement des critères sur la FAIRisation des données.

- La recommandation de déposer les données dans un entrepôt en accès libre. Ceci est particulièrement important avec la multiplication des incitations à déposer les données dans des entrepôts gérés par les éditeurs des journaux, etc. Les entrepôts ne sont pas nécessairement disciplinaires, mais quand on examine s'ils sont de confiance, on cherche s'ils servent leur communauté cible (qui peut être plus large qu'une discipline) selon les pratiques et les attentes de celle-ci.

- Le CNRS constituera à l'intention des chercheurs et des chercheuses un annuaire des entrepôts et des services de données existants, avec en particulier l'objectif d'aller vers la certification des entrepôts et services de données. Un entrepôt doit avoir un rôle de curation et de préservation des données, et les principes FAIR sont un objectif dans le contexte Science Ouverte. La certification de base *CoreTrustSeal* explicite les critères pour un entrepôt « de confiance », ce qui permet de travailler à améliorer les pratiques en se basant sur les critères, sans nécessairement aller jusqu'à soumettre un dossier de certification.

- La possibilité de mettre en place des périodes propriétaires ou d'embargo « raisonnables » sur les données (par exemple six mois à deux ans) prenant en compte les pratiques disciplinaires.

- L'encouragement à la réutilisation des données pertinentes, s'il en existe, plutôt que la création de nouvelles données.

La stratégie et la politique du CNRS en matière de données de la recherche devront être articulées avec celles du MESRI et de ses partenaires de l'ESR. Le CNRS désignera ainsi un administrateur ou une administratrice des données tel que préconisé par le PNSO pour représenter le CNRS dans un réseau en cours de mise en place par le MESRI.

Le CNRS devra assurer sa participation aux forums nationaux, européens et internationaux de discussion des politiques de la science ouverte, du calcul et des données de la recherche. La participation du CNRS à la *Research Data Alliance* (RDA) et au *European Open Science Cloud* (EOSC) fait partie intégrante de sa politique des données de la recherche, coordonnée avec le MESRI. Le CNRS pilote le nœud français de RDA soutenu financièrement par le MESRI. EOSC comprend des problématiques de calcul, données de la recherche et science ouverte. Au niveau international, il est prévu que le CNRS adhère à l'association EOSC parmi les premiers membres pour fin 2020.

1. CDS : https://cdsweb.u-strasbg.fr/index-fr.gml
L'exemple est révélateur : le CDS emploie 1/3 d'astronomes, 1/3 d'informaticiens et 1/3 de documentalistes.
2. www.huma-num.fr

5 | GOUVERNANCE DES DONNÉES

Le CNRS doit se doter de moyens pour agir de manière coordonnée sur l'ensemble du continuum, qui va du calcul aux publications, en passant par les données massives, la longue traîne de données, les infrastructures matérielles et logicielles, les services aux utilisateurs, le référencement et la documentation. Toute division serait un frein à la mutualisation et à la transversalité des actions.

La gouvernance des données doit prendre en compte les trois aspects suivants :

Scientifique :
De nouvelles découvertes voire de nouveaux thèmes de recherche peuvent émerger de la réutilisation des données et le partage des bonnes pratiques entre les instituts.

Technique :
Il s'agira de promouvoir et de mettre en œuvre la gestion FAIR des données de la recherche, de stimuler les réseaux métiers et groupes inter-réseaux et de définir des nouveaux métiers.

Économique :
De nouveaux moyens mutualisés, en particulier humains, seront nécessaires. Il faudra évaluer les conséquences sur les coûts (notamment des économies à plus long terme) et sur les impacts socio-économiques potentiels des données FAIR.

Par ailleurs, « aussi ouvert que possible, fermé autant que nécessaire » devra être décliné avec précision, en concertation avec la Direction des affaires juridiques, la Direction générale déléguée à l'innovation, la Délégation à la protection des données, le Fonctionnaire sécurité défense et la Direction de la sûreté.

Au service des nombreuses communautés scientifiques, la gouvernance des données au CNRS devra travailler avec les partenaires de l'ESR qui contribuent notamment à de nombreuses infrastructures productrices de données et à de nombreux services. Ce lien pourra être assuré par l'administrateur ou l'administratrice des données, évoqué plus haut et représentant le CNRS dans un réseau inter-établissements en cours de mise en place par le MESRI selon le Plan National pour la Science Ouverte. Il faudra aussi travailler en lien avec les projets et organisations internationales en coordination avec les instituts. Ces derniers devront être fortement impliqués dans la gouvernance des données.

Un « modèle économique durable » devra être établi pour soutenir toutes les composantes de l'écosystème de données FAIR : curation, mise à disposition, stockage, voire archivage à long terme des données. L'archivage et les services de données ont des coûts financiers et humains qu'il faut chiffrer et prendre en compte, notamment au-delà de la durée des projets qui ont engendré les données. La fédération, la mutualisation de certains services de données, doit permettre d'optimiser et de réduire ces coûts. La définition d'un modèle économique est aussi un point clé de la participation à l'EOSC puisqu'il conditionnera l'implication des acteurs. Un autre aspect du modèle économique concerne le coût[1] qu'il y aurait à ne pas rendre les données FAIR et ceci y compris dans le domaine de l'innovation.

L'intérêt d'une gouvernance unique des données, unique au sens inter-instituts, est d'apporter des moyens supplémentaires sur des actions transverses. Elle doit donc disposer de moyens humains et financiers dédiés, pour sa propre capacité d'action, et pour renforcer les structures dans lesquelles ces moyens seraient affectés (datacenters nationaux ou régionaux, INIST[2], CCSD[3], …) et promouvoir la mutualisation des ressources et des expertises, comme le fait aujourd'hui MICADO[4] pour le calcul, et la DIST[5] pour la science ouverte. Les dispositifs disciplinaires gérés par les instituts n'ont pas vocation à changer de fonctionnement. ▶▶▶

1. Une étude publiée par la Communauté européenne a tenté d'évaluer ces coûts : *Cost-benefit analysis for FAIR research data Cost of not having FAIR research data - Study DOI: 10.2777/02999* : https://www.ouvrirlascience.fr/wp-content/uploads/2019/03/Cost-Benefit-analysis-for-FAIR-research-data_KI0219023ENN_en.pdf
2. Institut de l'information scientifique et technique du CNRS : https://www.inist.fr/
3. Centre pour la communication scientifique directe : https://www.ccsd.cnrs.fr/
4. Mission « Calcul – Données » du CNRS
5. Direction de l'information scientifique et technique du CNRS

Une nouvelle Direction fonctionnelle des données ouvertes de la recherche, DDOR, rattachée à la Direction générale déléguée à la science (DGDS) aura pour mission de proposer et d'accompagner la mise en application d'une politique et d'une stratégie pour l'ouverture des données au CNRS, en intégrant l'ensemble des dimensions du sujet. De façon globale, la DDOR définit et met en œuvre la stratégie pour la science ouverte, élargie à toutes les questions afférentes aux données de la recherche, y compris aux thématiques propres des infrastructures numériques. Le cap de la DDOR est fixé par la Feuille de route pour la Science Ouverte établie en 2019 et par ce Plan des données de la recherche du CNRS d'octobre 2020. Cette direction issue de la fusion de la DIST et de MICADO fournira ainsi un cadre d'aide au traitement des questions liées à l'ouverture des publications scientifiques, à la gestion et au partage des données de la recherche, à la problématique des données massives, à leur stockage et aux infrastructures numériques. Les instituts du CNRS joueront un rôle moteur au sein du comité de pilotage de cette nouvelle direction. Celui-ci sera constitué d'un représentant ou d'une représentante de chaque institut à un niveau décisionnaire, principalement des DAS (Directeurs scientifiques adjoints).

En complément à ce comité de pilotage et afin de traiter les questions reliées à la fermeture des données et en particulier du partage des bonnes pratiques pour l'aide à la différentiation entre les données ouvertes et celles à protéger, sera mise en place une cellule comportant la direction de la DDOR, le Fonctionnaire sécurité défense et des représentants de la Direction générale déléguée à l'innovation, de la Délégation à la protection des données, et de la Direction de la sûreté. Lorsque cela sera nécessaire, les membres de cette cellule seront invités aux réunions du comité de pilotage.

6 | PLAN D'ACTIONS

Un plan d'actions immédiates est esquissé ici. Il ne peut en aucun cas être exhaustif à ce stade et devra être co-construit avec les instituts. L'une des premières missions du comité de pilotage de la nouvelle direction, sera de revoir l'ensemble de ces actions, les valider, en proposer de nouvelles, les hiérarchiser et organiser leur mise en œuvre. Certaines relèvent de l'adéquation aux normes et standards (y compris FAIR), d'autres de la politique incitative des instituts, ou encore de la pratique quotidienne des chercheurs.

Certains points ne sont volontairement pas abordés dans ce texte car ils doivent faire l'objet de travaux plus approfondis. Un dispositif incitatif pour encourager les chercheurs à partager leurs données devrait être étudié. L'entrepôt de données n'est pas la seule option : des infrastructures de fédération ou de publication-souscription permettent aussi de partager les données sans passer par un entrepôt. L'incitation à partager doit s'accompagner d'une incitation à réutiliser les données, avec des services d'ingénierie adaptés pour faciliter cette réutilisation.

Le plan d'actions proposé ici est une illustration concrète des problèmes immédiats soulevés autour de la gestion des données et qui nécessitent une approche transverse, ce qui n'exclut pas de bien tenir compte des spécificités des pratiques disciplinaires et des politiques internes des instituts.

1. Favoriser l'émergence de bonnes pratiques

Les principes FAIR constituent un fil conducteur qui traverse les actions et les outils présentés dans cette section, qu'il s'agisse de Cat OPIDoR[1] mis en œuvre par l'INIST, des plans de gestion des données ou des entrepôts intégrant des démarches de certification. Les cahiers de laboratoire constituent également une source permettant d'atteindre les niveaux de documentation requis dans une optique FAIR.

Soutien aux communautés scientifiques pour la définition des éléments spécifiques de la gestion de leurs données. Cela comprend la définition des critères de sélection des données à FAIRiser, en prenant en compte entre autres, la valeur scientifique et l'impact pour la recherche actuelle et émergente, les principes utilisés par les archivistes et le fait que des observations peuvent être impossible à reproduire. Il faut également les soutenir dans le développement et la maintenance de leur cadre disciplinaire de partage et de FAIRisation des données dans un contexte international, notamment au travers des instituts associés à la gouvernance transverse du dispositif. Soutenir les communautés, c'est les accompagner à s'organiser et à se structurer autour des données et des pratiques de données. C'est une mission essentielle des instituts qui apportent de surcroît une vision nationale et internationale. Le dialogue avec les communautés, la compréhension de leurs besoins et de leurs pratiques est bien le rôle des Instituts.

Cartographie des données et des outils et services pour la gestion et le partage des données. Une cartographie des données produites par les unités de recherche ayant le CNRS pour tutelle est indispensable. En cours de constitution via un questionnaire progressivement diffusé auprès des directeurs ou des directrices d'unités par la DIST, elle permettra d'identifier les lieux de production des données (infrastructures et/ou expérience de laboratoires), les lieux de stockage des données, la quantité de données produites, les politiques de partage et de conservation des données au sein des unités, etc. Il faudra aussi disposer d'une cartographie de l'offre existante de structures et services de gestion des données et de leurs métadonnées, aussi bien au sein du CNRS que de ses partenaires de l'ESR. Cette action est notamment en cours pour les entrepôts et les services dont le CNRS est responsable. Ce travail doit également permettre d'identifier dans un deuxième temps, pour chaque institut, les besoins des chercheurs et des chercheuses.

***Data Management Plan* (DMP).** L'obligation dans les appels à projets européens et dorénavant de l'ANR[2], de produire un plan de gestion des données pour tout projet de recherche déposé, constitue une bonne opportunité pour inciter les chercheurs et les chercheuses à prévoir les conditions d'entreposage, de curation et de diffusion de leurs données tout au long de leur cycle de vie. Il s'agit de les accompagner dans cette démarche. L'INIST y contribue avec l'outil DMP OPIDoR[3] qui facilite la rédaction de ces plans et avec l'organisation de « l'OPIDoR tour » dans l'ensemble des régions. Cet accompagnement doit s'appuyer sur les communautés scientifiques et des composantes locales, qui pourraient être coordonnées par un réseau métier au CNRS. Une généralisation et l'adoption d'un plan de données standardisé pour tous les projets (hors ANR, Horizon Europe) permettraient de faciliter la mise en œuvre des principes FAIR et une bonne gestion des données. Le plan de gestion de données évoluera pour faciliter sa rédaction en utilisant toutes les possibilités technologiques permettant l'extraction et la réutilisation des données des différents systèmes d'information. On parle de DMP « actionnables par des machines », c'est-à-dire remplis et exploités automatiquement.

Identifiants pérennes pour les jeux de données de la recherche. Il est essentiel de développer l'usage des identifiants pérennes pour les données, les logiciels et les publications. L'INIST contribue à cette action par l'attribution de DOI (*Digital Object Identifier*) aux données de recherche pour le CNRS et plus largement pour l'ESR en tant qu'agence nationale *Datacite*. Il y a actuellement plus d'une centaine d'utilisateurs – au sens de structures de recherche au CNRS – et à l'ESR, tel que INRAE, Ifremer, ESRF (E*uropean synchrotron research facility*), CDS (Centre d'astronomie de Strasbourg), Observatoire Midi Pyrénées, CDSP (Centre de données socio-politiques), etc. *Software Heritage* a fait des propositions récentes pour les logiciels[4].

Cahier de laboratoire électronique (CLE). Au-delà de ses fonctions liées à la propriété intellectuelle ou à la preuve d'intégrité scientifique, le cahier de laboratoire peut constituer un outil de prise en charge des données en vue de leur conservation et diffusion, via le cahier lui-même ou via les liens qu'il contient. Il est nécessaire de prendre position en faveur du principe d'adoption du cahier de laboratoire électronique au CNRS (remplacement des cahiers papier) et de faire des recommandations/obligations sur les caractéristiques de l'outil adopté, en termes de conservation, traçabilité, pérennité et accessibilité des données et de l'outil. Il faudra étudier les solutions choisies par nos partenaires français (Inserm, Ifremer, Inrae, Cnes, …), aux niveaux européen et international. Un groupe de travail piloté par la Mission pilotage et relations avec les délégations régionales et les instituts (MPR) fait ce travail au CNRS.

Certification des dispositifs de prise en charge des données de la recherche (notamment le *CoreTrustSeal*[5]). La certification des entrepôts et services de données, citée comme un objectif dans le Plan National pour la Science Ouverte, permet d'assurer qu'un centre de données est « de confiance », en examinant la manière dont il met en œuvre l'ensemble de la chaîne liée aux données, de leur ingestion à leur dissémination et à leur préservation. Elle peut aussi s'entendre dans le cadre de réseaux de centres de données, par exemple ceux des Pôles de données thématiques de l'IR Data Terra[6], ou ceux de l'infrastructure européenne CLARIN[7]. Le CNRS pourra s'appuyer sur les activités de soutien à la certification mises en place par le Nœud National RDA France[8].

1. Cat OPIDoR : https://opidor.fr/reperer/
2. GENCI imposera prochainement un DMP pour les demandes de ressources sur les machines de calcul nationales.
3. DMP OPIDoR : https://dmp.opidor.fr/
4. Identifiants pour les logiciels https://www.softwareheritage.org/2020/05/26/citing-software-with-style/
5. *CoreTrustSeal : CoreTrustSeal Requirements* v02.00-2020-2022 (doi:10.5281/zenodo.3638211)
6. Data Terra : https://www.data-terra.org/
7. CLARIN : https://office.clarin.eu/v/CE-2013-0095-B_checklist-v7_3_1.pdf
8. RDA France : https://www.rd-alliance.org/groups/rda-france

2. Favoriser l'émergence de nouveaux outils

Certaines communautés produisent des données sans DMP (typiquement via des expériences de laboratoire) et n'ont pas de « tradition » du cahier de laboratoire. Ces données, souvent mal référencées, sont généralement définitivement perdues après les publications correspondantes, au départ du doctorant qui les a produites ou lors des changements du matériel de stockage. Plus prosaïquement, il n'existe pas aujourd'hui d'offre de stockage générique pour les laboratoires, sans même parler de référencement ou de FAIRisation. Si un laboratoire pose la question « où pourrais-je stocker mes données ? », il n'obtient pas de réponse la plupart du temps et finit par s'équiper lui-même ou, au mieux, se voit proposer des solutions spécifiques au coup par coup. Comment initier et encourager une démarche de prise en charge de ces données ?

Un outil électronique, « répertoire des recherches » du laboratoire ou cahier de laboratoire informatisé, pourrait faire le lien vers le serveur hébergeant les données, inviter à l'usage de formats de données interopérables, imposer une documentation minimale de la donnée en vue de sa diffusion éventuelle. Un tel outil devra être simple, interopérable (et bon marché) et pouvoir servir, le cas échéant, de base à un futur DMP ou faciliter un dépôt dans un entrepôt à l'issue de la recherche.

Une alternative pourrait être un entrepôt institutionnel CNRS pour les nombreuses données dites de « longue traîne » souvent très diverses et de petits volumes, mais dont le volume et la diversité croissent rapidement avec le développement des techniques d'imagerie et les capteurs distribués. La question d'un tel entrepôt devrait être posée en coordination avec l'initiative en cours du MESRI qui étudie la faisabilité d'un entrepôt national de données « simples » à travers un groupe de travail piloté par J.C. Desconnets de l'IRD. De leur côté, INRAE, IRD, Cirad, etc. ont déjà choisi la solution d'un entrepôt institutionnel, tandis que d'autres établissements ont fait d'autres choix ; il s'agit ici de rechercher une cohérence nationale et une flexibilité permettant de simplifier le travail de dépôt, de référencement et de mise à disposition des données, tout en respectant les différentes pratiques de recherche.

Cet entrepôt générique pour les données de la recherche de la longue traîne, s'il devait être acté, devrait se faire en complémentarité avec les dispositifs existants, que cela soit des infrastructures de recherche qui ont, pour les plus grandes d'entre elles, la capacité de développer leurs propres outils dans un cadre national et/ ou international, ou bien des infrastructures de données qui prennent en charge la FAIRisation des données, et qui sont souvent thématiques, et en prenant en compte les dispositifs mis en place localement par les universités ou au niveau régional. Il y a une nécessité d'associer étroitement calcul, stockage et traitement des données en particulier lorsque de très gros volumes sont produits. Des centres nationaux peuvent jouer un rôle privilégié, comme le fait déjà le CC-IN2P3[1] qui est un centre thématique dédié à la communauté de l'IN2P3. Ils sont également en mesure d'offrir les ressources et services de données et de calcul avec les capacités nécessaires à des traitements et des analyses de plus en plus lourds.

Certaines TGIR archivent les données sur le long terme et les distribuent après une période d'embargo, notamment en astronomie et en sciences de la planète Terre. D'autres n'assurent pas le stockage et la curation de ces données. La mise à disposition des données de la recherche à des scientifiques qui n'utilisent pas les instruments et les systèmes d'observation qui ont produit ces données reste à développer dans la plupart des cas, en particulier pour des applications inter et/ou transdisciplinaires.

Le référencement est un point complémentaire aux stratégies de stockage. Il s'agit de disposer d'une plateforme logicielle de services pour faciliter la visibilité et l'accès aux jeux de données par des outils proposant des fonctionnalités de recherche conviviales. La qualité et la pertinence des métadonnées est fondamentale pour garantir une bonne accessibilité aux données.

3. RH - Évaluation et gestion des carrièress

Le CNRS doit mettre en place une réponse coordonnée face aux nouveaux besoins en termes d'expertise, de moyens humains et de reconnaissance de ces nouvelles activités transdisciplinaires sur les données de la recherche. Il pourrait être aussi judicieux de créer un système de valeurs qui encourage les chercheurs à publier leurs données.

Inventorier tous les rôles joués par les personnels dans la FAIRisation des données et les compétences nécessaires dans ces rôles, pour nourrir et faire évoluer les référentiels d'évaluation des chercheurs et des chercheuses et les référentiels de postes IT (Referens[2]). Cette action conduira à identifier de nouveaux profils et de nouveaux métiers.

1. Centre de calcul de l'Institut national de physique nucléaire et de physique des particules du CNRS : https://cc.in2p3.fr/
2. RÉFérentiel des Emplois-types de la Recherche et de l'ENseignement Supérieur (REFERENS) : https://data.enseignement-sup-recherche.gouv.fr/pages/referens/

Cibler les recrutements grâce aux nouveaux profils identifiés et bénéficier ainsi de moyens humains supplémentaires pour accompagner la gestion et le partage des données. Les moyens humains sont majoritairement IT (*data stewards*) mais peuvent être aussi des chercheurs ou des chercheuses (*data scientists*). Le besoin immédiat en ingénieurs est très fort. Il peut faciliter le partage des expériences et des efforts.

Accompagner une évolution profonde de l'évaluation des scientifiques. Il faudra agir pour que les investissements en support à la FAIRisation et au partage des donnée soient considérés dans l'évaluation. Cela concerne des tâches d'intérêt collectif comme le partage des données sous un mode FAIR, la participation à la définition des pratiques et des standards au niveau disciplinaire et interdisciplinaire, la participation à leur mise en œuvre comme expert scientifique, la participation aux tâches scientifiques nécessaires au partage des données dans les infrastructures de recherche et les infrastructures de données… Au niveau individuel devront être valorisés la production et le partage (lorsque c'est possible) de données et plus généralement d'objets numériques (logiciels) FAIR, dans les projets et la pratique de la recherche, la pratique du DMP, la réutilisation des données dès qu'elle permet de nouvelles avancées de la connaissance...

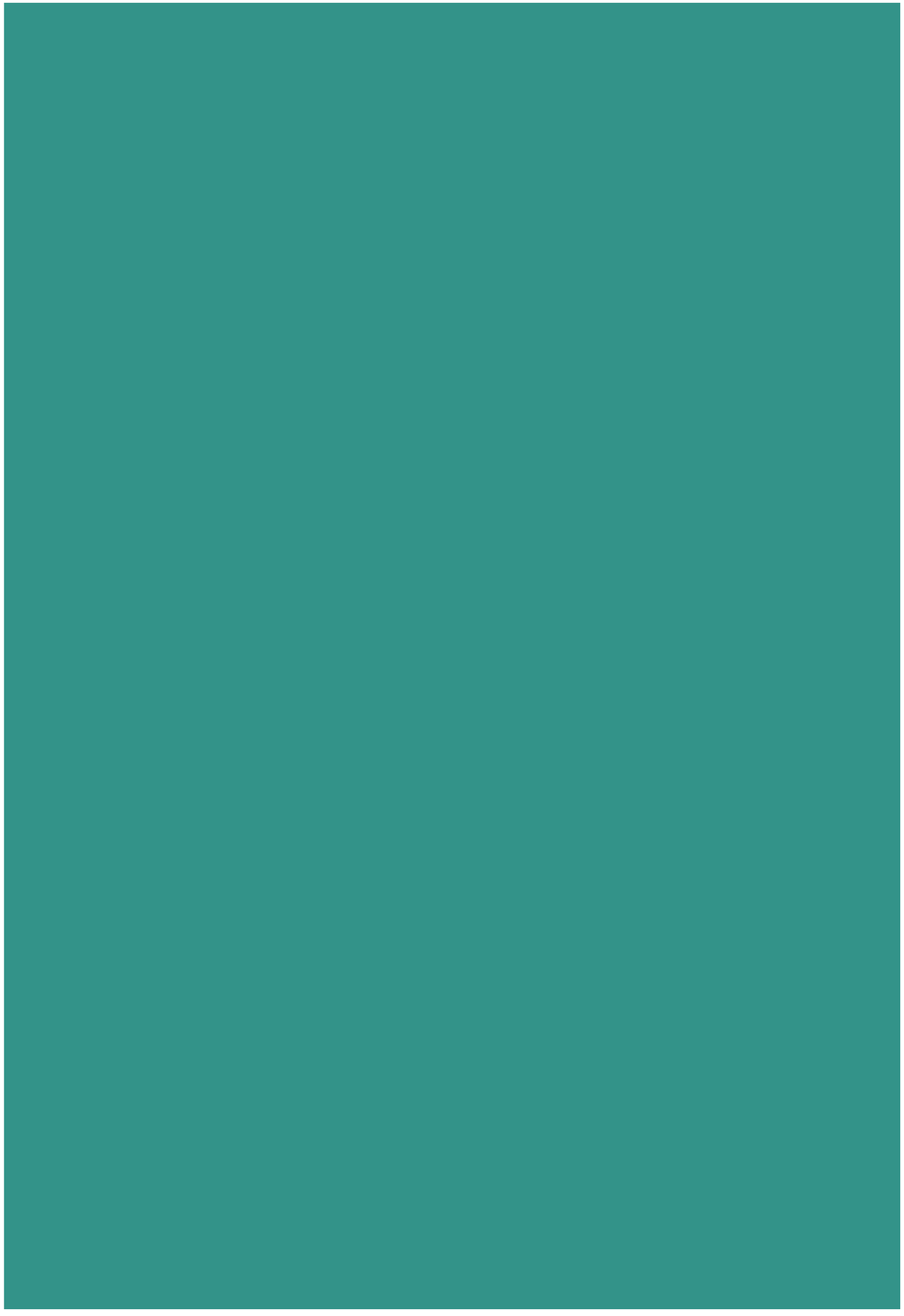
4. Formation

Formation et communication sont des éléments essentiels de ce plan d'actions. Il s'agit en premier lieu de développer les compétences en matière de données FAIR tant en termes de gestion de données (principes FAIR de gestion, partage, archivages intermédiaire et pérenne) que de science des données (manipuler, traiter, analyser les données de recherche). L'INIST participe à cette action avec des formations à distance via la plateforme DoRANum[1] et en présentiel sur le volet gestion, partage, archivage des données et plan de gestion de données.

La culture de la donnée doit être développée par la formation permanente pour tous les types de personnels afin de les sensibiliser au sujet, leur faire connaitre et apprendre à utiliser les standards et plus généralement les différents outils liés aux données (DMP, entrepôts, langages, etc.)

Il est important d'identifier des succès et de les partager, pour convaincre les communautés encore peu impliquées d'adopter le partage des données. Leur sensibilisation reposera sur des échanges de bonnes pratiques et les leçons tirées de l'implantation de FAIR et de la définition des limites légitimes de l'ouverture des données. L'un des aspects est de collecter des exemples concrets et des guides d'implantation. Il faut mener le même type d'action sur la mise en œuvre des DMP. Ces échanges peuvent se faire dans le cadre des champs disciplinaires (instituts) ou via les réseaux métiers (MITI[2]), et en encourageant les personnels impliqués dans le partage des données à participer à la RDA.

1. DoRANum : https://doranum.fr/
2. Mission pour les initiatives transverses et interdisciplinaires du CNRS



© AdobeStock/ra2studio

CNRS
3, rue Michel-Ange
75794 Paris Cedex 16
01 44 96 40 00
www.cnrs.fr

Impression : CNRS IFSeM secteur de l'imprimé
Novembre 2020