FEUILLE DE ROUTE DU CNRS POUR LA SCIENCE OUVERTE

18 novembre 2019

TABLE DES MATIÈRES

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

INTRODUCTION

Le mouvement international pour la science ouverte, initié il y a plus de 30 ans, connaît un développement sans précédent depuis que le web l'a rendu possible à une échelle globale avec des coûts raisonnables. La diffusion de la production scientifique sur internet, son identification et son archivage lèvent les barrières de l'accès pérenne sans remettre en cause, ni la protection des données personnelles, ni la protection de la propriété intellectuelle. Il s'agit d'être « ouvert autant que possible, fermé autant que nécessaire ».
La science ouverte ne favorise pas seulement une approche transversale du partage des résultats de la science. En ouvrant les données, les processus, les codes, les méthodes ou encore les protocoles, elle offre aussi une nouvelle façon de faire de la science.

Plusieurs raisons scientifiques, citoyennes et socio-économiques rendent aujourd'hui incontournable le développement de la science ouverte :

- Le partage des connaissances scientifiques rend la recherche plus efficace, plus visible, moins redondante. L'accès ouvert aux données et aux résultats révolutionne la manière de faire de la recherche, et permet l'utilisation de nouveaux outils. Les outils issus des recherches récentes en intelligence artificielle (fouille de textes et de données, apprentissage automatique) permettent ainsi une recherche plus transversale, facilitant l'interdisciplinarité.
- La science ouverte modifie la façon dont la recherche s'inscrit dans la société en aidant à rétablir la confiance entre le citoyen et le scientifique. L'accessibilité par de nombreux acteurs et par plusieurs disciplines facilite la réponse aux enjeux contemporains (défis multi- et/ou inter-disciplinaires par exemple) et l'intégrité scientifique est renforcée.
- Les chercheurs et les chercheuses paient souvent pour publier, paient souvent pour lire les articles, réalisent gratuitement le travail de reviewing, et les coûts d'abonnement ne cessent d'augmenter. La publication accessible des résultats scientifiques redonne aux scientifiques le contrôle d'un système qui nous a échappé et qui est devenu financièrement insoutenable à cause de l'augmentation constante des coûts de la diffusion scientifique.

Notre pays s'est engagé, il y a un peu plus d'un an, dans ce vaste mouvement de transformation. Présenté le 4 juillet 2018 par la Ministre de l'Enseignement Supérieur et la Recherche, le « Plan National pour la Science Ouverte »[1] vise, selon les termes de Frédérique Vidal, à ce que « les résultats de la recherche scientifiques soient ouverts à tous, chercheurs, entreprises, et citoyens sans entrave, sans délai, sans paiement ».

Dans la foulée, un Comité pour la Science Ouverte (CoSO), présidé par le Directeur général de la recherche et de l'innovation (DGRI), a été institué. Il se décline en quatre collèges : publications, données de la recherche, compétences et formation, Europe et international. Des chercheurs et ingénieurs du CNRS, ainsi que des membres de la DIST (Direction de l'Information Scientifique et Technique), sont présents dans tous les groupes et co-pilotent trois des quatre collèges. Le plan national pour la science ouverte engage les opérateurs de la recherche à se doter d'une politique de science ouverte.

Compte tenu de l'intrication des travaux menés au sein des unités mixtes de recherche par les chercheurs et les chercheuses et les enseignants-chercheurs et les enseignantes-chercheuses, le développement de la science ouverte doit être pensé nationalement et conjointement avec l'ensemble des acteurs de la recherche, et en particulier les autres organismes de recherche et nos partenaires universitaires, en forte cohérence avec l'action du MESRI (MInistère pour l'Enseignement Supérieur, la Recherche et l'Innovation) pour la science ouverte.

La feuille de route du CNRS pour la science ouverte sera mise en œuvre à travers un plan d'action régulièrement revisité et placé sous la responsabilité de la Direction de l'information scientifique et technique (DIST), laquelle mobilisera en particulier les trois unités de services qui sont dans son périmètre :
L'Inist (Institut de l'information scientifique et technique), unité propre de service du CNRS installée depuis 1989 à Nancy qui vient d'être réorganisée en trois départements liés à trois axes de développement de la science ouverte :

- Accéder à l'information scientifique, axe majeur et historique de l'Inist avec les négociations et le portail BibCNRS, mais aussi les nouvelles plateformes ISTEX et PANIST pour les publications ;
- Valoriser les données de la recherche, axe qui devrait s'étoffer à l'avenir et qui profite de l'investissement pionnier de l'Inist pour développer des outils autour des données de la recherche tels que DMP OPIDoR (Optimisation du Partage et de l'Interopérabilité des Données de la Recherche) pour faire des Plans de Gestion des Données, etc. et également pour faire des réservoirs de données pour les « petites » données ;
- Analyser et fouiller l'information scientifique, axe d'avenir pour le développement des outils de Text et Data Mining.

Le CCSD (Centre pour la communication scientifique directe), unité mixte de service, maître d'œuvre de HAL (Hyper Articles en Ligne), les archives ouvertes, et qui a ensuite développé Episciences, plateforme de publications de revues en accès ouvert et d'autres services ;
PERSEE, créée en 2013 pour numériser et valoriser le patrimoine scientifique.

La politique de la science ouverte se fera de manière différenciée entre les différentes disciplines. Elle doit être placée au cœur des instituts du CNRS. Ceux-ci gèrent en particulier de nombreuses infrastructures de recherche productrices de données. Certaines sont des infrastructures thématiques de gestion et de partage des données inscrites sur la feuille de route nationale des infrastructures de recherche, qui jouent un rôle important dans le paysage national, européen et international, tels le Centre de données astronomiques de Strasbourg, la plateforme Huma-Num des humanités numériques, ou l'Infrastructure de recherche data terra (ex-Système Terre). Les instituts ont également développé des initiatives liées aux publications, en particulier la plateforme d'édition de revues et ouvrages en accès ouvert « OpenEdition », le centre Mersenne développé pour la publication en accès ouvert des revues mathématiques et actuellement en phase d'ouverture à toute discipline dans laquelle les articles sont composés en LaTeX. La DIST est en relation permanente avec les instituts thématiques grâce aux correspondants Information scientifique et technique (IST). Un réseau de « correspondants données » est mis en place dans le cadre du nœud français de la Research data alliance (RDA).

La mise en œuvre de la feuille de route du CNRS pour la science ouverte présentée a l'ambition d'accélérer le processus vers la science ouverte en s'appuyant sur des actions concrètes structurées autour de quatre grands objectifs :

1. Garder le contrôle sur notre production scientifique et aboutir à l'échelle de la feuille de route à 100 % des publications du CNRS en accès ouvert ;
2. Développer une culture de la gestion/partage des données chez tous les acteurs du cycle de vie de la donnée : chercheurs et chercheuses, ingénieur-e-s, informaticiens et informaticiennes, documentalistes, bibliothécaires… basée sur la mise en œuvre des principes FAIR (Faciles à trouver, accessibles, interopérables et réutilisables) ;
3. Développer et promouvoir des infrastructures, et des outils permettant la fouille et l'analyse des contenus scientifiques en toute indépendance ;
4. Transformer l'évaluation individuelle des chercheurs et des chercheuses en la rendant compatible avec les objectifs de la science ouverte d'une part et prendre en compte la contribution des chercheurs et des chercheuses à la science ouverte dans l'évaluation d'autre part.

L'ensemble de ce processus sera accompagné par des actions de formation et une stratégie internationale.

1. bit.ly/plannationalSite

1 | LES PUBLICATIONS

Objectif : Les publications scientifiques, produites par le travail des chercheurs et des chercheuses du CNRS, et financées en majorité sur des fonds publics, doivent être à 100 % accessibles (*) et ré-utilisables à l'échelle de temps de la feuille de route. Les droits d'auteur ne doivent pas être cédés.

La stratégie du CNRS pour atteindre 100 % des publications scientifiques en accès ouvert est de favoriser la « bibliodiversité »: une diversité des produits issus de la recherche ainsi qu'une diversité des modes de publications. En effet, il existe plusieurs voies possibles pour publier en accès ouvert. Au-delà des archives ouvertes, les modèles d'éditions dits « vertueux » en accès ouvert, à but non lucratif, se multiplient. Les relations avec les éditeurs conventionnels sont toujours considérées mais dans un contexte de négociations qui vise à réduire les coûts et à favoriser l'accès ouvert.

Action 1 : mener une politique de soutien et de développement de l'archive ouverte HAL conjointement à une politique d'incitation à y déposer les publications scientifiques.

Cela passe par le développement du CCSD, l'Unité mixte de service qui pilote HAL. En 2019, le CNRS a mis des moyens supplémentaires exceptionnels à hauteur de 650 k€ en plus d'un poste externe ouvert au concours de la fonction publique. La pérennisation de cette augmentation des moyens nécessaires à une nouvelle ambition sera assurée en 2020 par un support du fonds national pour la science ouverte, puis par une contribution des utilisateurs suivant un modèle économique pour le développement adapté de HAL relié à ses nombreux partenariats qui a été validé par le Coso (comité pour la science ouverte). Ceci est cohérent avec le Plan national qui a choisi de mettre au cœur de sa politique l'archive ouverte HAL, développée par le CNRS et les autres tutelles INRIA, l'INRA et l'université de Lyon. Cette infrastructure a été adoptée depuis plusieurs années par un grand nombre d'universités et d'organismes de recherche (137 portails institutionnels ont été développés).

Toutes les études sur les archives ouvertes montrent qu'il ne suffit pas d'encourager le dépôt dans les archives ouvertes mais qu'un seuil de 80 % des dépôts ne peut être atteint que par une politique plus fortement incitative. Aussi le CNRS demande à ce que toutes les publications scientifiques issues des recherches financées essentiellement par des fonds publics et pouvant être déposées en archives ouvertes sur la base de la loi Pour une République numérique, soient accessibles dans HAL.

Une première mesure a été prise dès 2019 par le CNRS qui demande à ce que seules les publications présentes dans HAL puissent être signalée dans le Crac (Compte rendu annuel des chercheurs et des chercheuses) 2019. La même mesure pour le compte rendu Ribac (activité annuelle des chercheurs et des chercheuses en SHS) est prévue pour 2020. Cette mesure a été annoncée aux directeurs d'unité par un courrier du Directeur général délégué à la science (DGDS) en date du 19 avril 2019.

L'augmentation du dépôt dans HAL doit s'accompagner d'une amélioration de l'outil HAL pour répondre aux besoins des communautés scientifiques françaises. Le CNRS s'engage pour cela à faciliter le dépôt dans HAL par les chercheurs et les chercheuses. L'interopérabilité avec les autres archives ouvertes internationales comme ArXiv, PubMed Central et RePEc devra être consolidée ou développée. Les questions liées à la modération, aux référentiels, au moissonnage, sont considérées également dans le but d'améliorer l'Infrastructure de Recherche HAL.

Action 2 : recommander l'utilisation des serveurs de preprints, hébergeant des manuscrits soumis à des revues, afin d'offrir des solutions de diffusion rapide en accès ouvert via des plateformes à but non lucratif.

À l'image d'ArXiv précurseur en physique et en mathématiques, de nombreuses communautés ont développé ou commencent à développer des serveurs de preprint destinés à la diffusion rapide des résultats scientifiques parallèlement aux processus de certification assurés par les revues. On peut notamment citer ChemRxiv en chimie, bioRxiv en sciences de la vie et SocArXiv en sciences humaines et sociales (SHS).

Ces plateformes permettent pour chaque manuscrit non encore évalué de garantir l'attribution aux auteurs, la datation, l'identification pérenne, l'hébergement et l'attribution d'une licence définissant les conditions de réutilisation. Elles permettent également parfois la soumission directe du manuscrit à une série de revues et font le lien avec la publication définitive si celle-ci a lieu.

(*) cela ne remet en cause, ni la protection des données personnelles, ni la protection de la propriété intellectuelle, ni tout autre protection nécessaire. Il s'agit d'être « Ouvert autant que possible, fermé autant que nécessaire »

Pour faciliter ces nouvelles modalités de communication scientifique en accès ouvert, le CNRS sélectionnera, d'une part, un certain nombre de plateformes qui seront soutenues en fonction de l'importance pour les communautés considérées, en coordination avec le fonds national pour la science ouverte. Il s'efforcera d'autre part de faire avancer l'interopérabilité entre les différentes plateformes de dépôts. Nous mobiliserons des moyens pour que le CCSD puisse proposer des solutions en coordination avec les responsables de ces plateformes et développe les outils qui permettront aux chercheurs et aux chercheuses de n'avoir qu'un seul dépôt à effectuer.

Action 3 : soutenir des plateformes d'édition électronique qui hébergent des textes en accès ouvert, qui proposent des espaces de publication et de certification.

La finalité est que chaque communauté scientifique, notamment disciplinaire puisse trouver des alternatives pour diriger ses publications vers des systèmes d'édition en accès ouvert, avec des modèles économiques qui peuvent être soutenus par les institutions afin que ce ne soit pas l'auteur qui paie, ni pour publier, ni pour lire.

Des plateformes de publication en accès ouvert proposent différents types d'objet : des revues, des ouvrages, des recommandations, des rapports d'évaluation des referees, des commentaires de lecteurs, des annotations ou d'autres formes de discussion autour des publications. Certaines plateformes sont, de fait, des hébergeurs, au sens où elles accueillent et valorisent du contenu scientifique produit par d'autres organisations (revues, éditeurs, communautés organisées, archives ouvertes) alors que d'autres sélectionnent et produisent directement leur propre contenu. Dans tous les cas, elles accueillent des étapes d'évaluation par les pairs, qui peuvent être plus ou moins ouvertes (open reports, open identities). La grande majorité d'entre elles publie des documents sous licence creative commons, la propriété intellectuelle étant conservée par les auteurs des textes.

Même si certaines ont pour objectif de couvrir l'ensemble des disciplines, elles sont marquées par une histoire et des choix plus restrictifs : par exemple, Copernicus est principalement spécialisé en géosciences, Scipost en physique, le centre Mersenne en mathématiques, OpenEdition en SHS, E-life et F1000 en Sciences de la Vie, Episciences en mathématiques, informatique, SHS, PCI (Peer community In) en sciences de l'environnement. Ces choix peuvent être liés au format d'écriture des auteurs en particulier le choix d'utiliser Latex ou d'autres logiciels de traitement de texte. C'est le cas, par exemple, respectivement du centre Mersenne et d'OpenEdition. Du côté des publications elles-mêmes, différents logiciels sont utilisés pour produire des documents en HTML, PDF, Epub ou XML (par exemple Lodel, OJS).

Le CNRS accompagnera les communautés scientifiques, discipline par discipline, dans l'élaboration de solutions alternatives de publications avec un contexte d'édition et d'évaluation par les pairs comparables aux éditeurs classiques.

Action 4 : soutenir une stratégie documentaire qui vise à diminuer le poids et le coût des abonnements aux revues et augmenter l'accessibilité des manuscrits dans la version éditée.

Le CNRS dépense chaque année plus de 12 M€ HT auprès d'une cinquantaine d'éditeurs pour que ses laboratoires aient accès à la documentation scientifique derrière des barrières d'accès (portail BibCNRS mis en oeuvre par l'Inist). Certains accès pour les unités mixtes du CNRS sont financés et proposés par les partenaires. Les accès peuvent être donnés par des portails des universités, des organismes. Certaines bibliothèques thématiques financent leurs propres abonnements complémentaires (réseau des mathématiques par exemple etc.). Les chercheurs et les chercheuses passent donc de plus en plus de temps pour accéder à la production scientifique. Menées dans le cadre de Couperin, les négociations des contrats avec les éditeurs qui accaparent la plus grande partie du marché doivent aboutir à des coûts de revues à la baisse, et une politique d'ouverture des publications conforme à la loi « Pour une République numérique » notamment. Le paiement de frais de publications pour rendre libre l'accès à un article dans une revue par ailleurs vendue par abonnement est une démarche qui doit être fortement déconseillée, car elle n'est pas conforme au plan national pour la science ouverte, et elle implique des coûts supplémentaires difficiles à suivre et à maîtriser pour une institution. Il s'agit de revues dites hybrides qui ne sont a priori pas vertueuses puisqu'on opère de fait un double paiement. La publication dans de telles revues, sans payer de frais de publication, accompagnée d'un dépôt dans HAL préserve la liberté de choix du canal de publication par les chercheurs et les chercheuses.

Action 5 : demander à ce que toutes les publications issues des travaux de recherche financés par un appel à projet du CNRS soient mises en accès ouvert.

Cet objectif est plutôt une action minimum, concrète et immédiate à mettre en place, cohérente avec le plan national de la science ouverte, et les demandes de l'Europe et de l'ANR dans leurs appels à projets respectifs. L'accès ouvert de la publication pourra se faire selon la diversité des voies possibles énoncées plus haut : archives ouvertes, plateforme d'édition de revues en accès ouvert. La voie archive ouverte est une alternative vertueuse à l'exigence d'APC (articles processing charges ou frais de publication en français) très importants pour des revues dites « prestigieuses ».

2 | LES DONNÉES DE LA RECHERCHE

Objectif : Les données (données brutes, textes et documents, codes sources et logiciels) produites par les chercheurs et les chercheuses CNRS ou avec des moyens mis en œuvre par le CNRS doivent être, dans la mesure du possible, rendues accessibles et ré-utilisables selon les principes FAIR pour une consolidation des connaissances essentielle au développement d'une science plus efficiente. « Les données doivent être aussi ouvertes que possibles, et fermées autant que nécessaire ».

Action 1 : développer une culture de la gestion/ partage des données chez tous les acteurs du cycle de vie de la donnée : chercheurs et chercheuses, ingénieurs et ingénieures, informaticiennes et informaticiens, documentalistes, bibliothécaires, …

Avec la massification actuelle des données numériques produites au cours du processus de recherche, il est essentiel de développer et de partager des bonnes pratiques (FAIR) entre acteurs de la recherche. Par données, on entend ici au sens large des données brutes ou traitées, des textes et documents, des codes sources et des logiciels. Les services OPIDoR développés par l'Inist permettent l'attribution d'identifiants DOI (Digital object identifier) à des jeux de données (via DataCite) et un accompagnement pour la mise en place de Plan de gestion des données (DMP ou Data Management Plan). Les DMP sont demandés lors de la rédaction des projets de recherche par la commission européenne et plus récemment par l'ANR. L'opportunité d'identifier des personnes ressources se retrouve dans le concept de « Data stewardship ». La Mission pour le pilotage et relations avec les délégations régionales et les instituts (MPR) a développé un outil méthodologique de gestion de données qui peut aider à la mise en place d'une politique de partage et gestion des données. Nous mettrons en place un réseau de personnes ressources, responsables de données dans les structures de recherche et les projets, formées à l'utilisation de services particuliers.

Les principes seront mis en place diversement dans des contextes disciplinaires distincts, sans dogmatisme et dans le respect des usages établis. La définition même de la donnée est très variable d'une communauté scientifique à l'autre. Cela comprend une discussion approfondie au sein des communautés sur le cycle de vie de la donnée, sa sauvegarde, son accès, son archivage, sa réutilisation. Les instituts du CNRS et leurs communautés ne sont pas toutes au même niveau d'avancement du développement d'une culture de la gestion et du partage des données. L'enjeu est de sensibiliser tous les acteurs de la recherche pour arriver à la mise en œuvre des principes FAIR.

Action 2 : développer la publication des données (data papers), le dépôt conjoint publications / données et accompagner les chercheurs et les chercheuses dans l'utilisation des outils de gestion des données.

Cet objectif est une façon de s'engager concrètement dans la question du partage des données. Les données sur lesquelles s'appuient les publications doivent être rendues accessibles et réutilisables dès que possible. Cette exigence est déjà mise en œuvre dans le cadre du programme européen H2020 et doit être généralisée à toute donnée produite à partir de fonds publics. Il s'agira donc de faciliter le dépôt concomitant à la publication dans l'archive ouverte des jeux de données sur lesquels se basent la publication (ceci ne signifie en aucun cas qu'il faudrait déposer les données dans les archives ouvertes). Les jeux de données associées aux publications seront déposés dans un réservoir de données approprié, si possible un dépôt thématique qui assurera au mieux la qualité et la diffusion des données. Plus généralement le CNRS encourage la diffusion de toutes données structurées par leur dépôt dans des réservoirs de données thématiques ou généralistes, éventuellement accompagné de la publication de data papers. Pour ce faire, les instituts du CNRS en lien avec (les services de) la DIST, pourront identifier le cas échéant des structures ressources autour d'eux, comme les bibliothèques universitaires, les MSH (Maison des Sciences et de l'Homme) en SHS, ainsi que des personnes ressources à même d'accompagner les scientifiques dans ces nouvelles pratiques. Ces « voltigeurs » en charge des activités de curation de données préfigurent la reconnaissance effective du « data stewardship ». Les instituts, qui ont une politique de soutien aux revues, disposent d'un levier supplémentaire et pourraient inclure à l'avenir un volet sur les données liées aux publications.

Action 3 : soutenir et accompagner les infrastructures de recherche, productrices de données, dans la définition et la mise en œuvre de politiques de données.

Le CNRS est largement engagé avec ses partenaires dans les Infrastructures de Recherche (IR) nationales et internationales, qui représentent les lieux où se créent et s'analysent les données de la recherche : instruments analytiques, infrastructures de calcul, infrastructures de données, observatoires, etc. Pour généraliser l'application des principes FAIR à toutes les disciplines, le CNRS publiera une charte des infrastructures, engageant celles-ci à respecter les pratiques FAIR et des standards de qualité, en affichant des politiques de données concertées avec les communautés scientifiques utilisatrices des infrastructures concernées. Certaines infrastructures (telles que Progedo et Humanum à l'institut des SHS (INSHS)) sont déjà bien engagées dans ce processus, d'autres sont en cours d'accompagnement telles que les IR de Chimie. Le synchrotron SOLEIL a également mis en route une politique de gestion des données. Les exemples sont multiples et devraient tendre à être généraliser. Ces développements doivent être corrélés avec les certifications (de type CoreTrustSeal) dans le cas où les infrastructures prennent elles-mêmes en charge la distribution de leurs données.

Action 4 : soutenir et accompagner des Infrastructures de données - Mettre en œuvre un service coordonné avec les instituts pour favoriser le dépôt des données pour tous les personnels des unités du CNRS

Les infrastructures de données thématiques jouent un rôle national ou international. Certaines sont inscrites sur la feuille de route nationale des infrastructures de recherche. Cela s'inscrit dans la mesure de structuration du Plan national pour la Science ouverte qui préconise de « développer des centres de données thématiques et disciplinaires ». Le CNRS continuera à soutenir ces infrastructures, et soutiendra le développement de nouveaux réservoirs et services de données thématiques. Ce soutien sera conditionné à une évaluation de leur impact, de leur adéquation aux besoins scientifiques, et de la qualité de leur gestion. Une certification CoreTrustSeal sera recherchée.

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. Ce n'est pas encore le cas pour les autres disciplines pour lesquelles la structuration internationale des données de la recherche est moins évidente et/ou établie, et parce que la nature même des données produites (taille, hétérogénéité, complexité, etc.) imposent une réflexion préalable sur la nature même de la donnée. Ainsi, beaucoup d'unités peuvent créer un grand nombre de « petites données » (au sens du stockage de la donnée, par opposition aux « big data »). C'est ce qu'on appelle la longue traîne des données. Le CNRS étudiera l'opportunité de créer un entrepôt de données généraliste accueillant les données dites de longue traîne pour lesquelles des entrepôts thématiques ne peuvent être identifiés. Cette réflexion se placera dans la cadre d'une réflexion au niveau national sur la prise en charge des données de longue traîne et de la mesure du Plan National « développer un service générique d'accueil et de diffusion des données simples ». La réutilisation des jeux de données ainsi collectés ne pouvant être effective que dans un cadre de confiance reconnu, un mécanisme de certification CoreTrustSeal sera réalisé.

Action 5 : créer et afficher un répertoire des dépôts et des services de données dont le CNRS est responsable et auxquels il participe.

Ce répertoire permettrait d'afficher un élément important de la contribution du CNRS à la Science ouverte. Il regrouperait les services de données des infrastructures de recherche cités mis en place dans le cadre de l'action 3, les centres et services de données thématiques (action 4) et l'entrepôt de données généraliste éventuellement créé dans le cadre de l'action 4.

3 | LA FOUILLE ET L'ANALYSE DES TEXTES ET DES DONNÉES

Objectif : Faciliter la fouille des textes et des données avec le développement des infrastructures, des outils et des compétences permettant l'analyse de contenus scientifiques en toute indépendance.

Depuis quelques décennies, le volume de l'information produite par la recherche augmente. Dans un domaine donné, les chercheurs et les chercheuses sont dans l'incapacité de « digérer » cette information produite, quand ils ont les moyens de s'en offrir l'accès. Les nouveaux services d'appui aux chercheurs et aux chercheuses dans le domaine de l'IST doivent évoluer vers l'analyse de contenu (Content mining) en complément de l'offre de signalement (bases de données bibliographiques, catalogues) et d'accès (portails) à la production scientifique. Ces derniers services mobilisent des technologies déjà éprouvées (indexation basée sur des fichiers d'autorité, moteurs de recherche, identifiants, etc.) et correctement maîtrisées par le monde de l'IST. Au contraire l'analyse de contenu nécessite le développement et la mise en œuvre de technologies encore peu maîtrisées par l'ensemble des scientifiques (ontologies, description sémantique, langages de représentation de connaissance, etc.). Ainsi, la fouille de contenu est définie comme l'acte d'explorer à l'aide de programmes, des textes et de données, numériques, pour en extraire de l'information. Ce processus, répété sur des corpus distincts, est susceptible de produire des connaissances scientifiques nouvelles. La fouille de textes et de données est ainsi vue par les chercheurs et les chercheuses comme une extension du « droit de lire ».

Action 1 : soutenir et développer des infrastructures permettant l'analyse de contenus

Il est important d'investir des ressources dans les activités de fouille de contenus (content mining). Récemment, les méthodes de deep learning ont créé de nouvelles possibilités de recherche pour traiter des données massives et de grandes dimensions. Le CNRS a orienté la plateforme ISTEX en ce sens, en portant l'idée que la collection agrégée de données de provenance diverses a permis le développement d'outils et de services d'exploration des contenus, en adéquation avec les demandes des chercheurs et les chercheuses. L'Inist alliant compétences d'infrastructures techniques, terminologiques et de connaissances scientifiques est l'opérateur choisi.

Le projet VISATM (Vers une Infrastructure de Services Avancés en Text Mining) développée à l'Inist et soutenue par le Mesri a permis de créer le lien entre la plateforme Istex et la plateforme européenne OpenMinted. Elle propose un panorama des outils de fouille de textes disponibles. Le projet VisaTM montre la nécessité d'une infrastructure offrant aux chercheurs et aux chercheuses l'environnement technique et scientifique, nécessaire pour mener à bien leur recherche utilisant les technologies de fouille de contenus.

Action 2 : cadre législatif : accompagner, traduire, et informer

La France a reconnu et pris en compte en octobre 2016, dans la loi Pour une République numérique, la nécessité d'autoriser à des fins de recherche et d'enseignement tout contenu licitement accessible. Cette volonté n'a toutefois pas été suivie de décret d'application dans l'attente du vote de la directive européenne pour le droit d'auteur.
L'Europe a voté récemment (mars 2019) le texte de la directive pour le droit d'auteur. Celle-ci crée une exception pour la recherche l'autorisant à pratiquer la fouille de contenus, à des fins de recherche sur tous contenus licitement accessibles.
Sa transposition dans le droit français nécessite d'être vigilant :
1. À l'imposition de contraintes sur les moyens d'accès autorisés pour le text mining.
2. Aux moyens mis en œuvre par les plateformes pour préserver l'intégrité des réseaux et des serveurs.
3. À l'ensemble des données d'usages (interrogation et téléchargement) recueillies par les plateformes et conservées et pouvant éventuellement être diffusées.
4. Aux licences négociées avec les éditeurs et la manière dont leurs clauses prennent en compte le cadre législatif du TDM.

Action 3 : développer l'usage d'outils et de techniques de traitement de données et de visualisation.

Les disciplines communiquent par le biais de textes et de données que ces dernières soient associées aux textes, ou pas. Ces informations sont importantes et nombreuses et difficilement appréhendables dans leur ensemble par l'humain seul. Les possibilités d'extraction d'information sont multipliées par l'usage de ces outils.
Les unités rattachées au CNRS se sont déjà emparées de ces problématiques. Par exemple, l'Inist du CNRS a développé des techniques et des outils de traitement de corpus. Ainsi, l'Inist propose des services d'extraction de contenus. Huma-num, la TGIR (très grande infrastructure de rechercher) des Sciences Humaines et Sociales, propose une liste d'outils pour extraire des informations des textes. L'Institut des systèmes complexes de Paris Ile de France, met à disposition une plateforme permettant, entre autres, l'extraction de contenus et leur visualisation. Ces techniques et outils doivent être connus des chercheurs et des chercheuses et ingénieur-e-s dans les laboratoires et leur usage facilité. Tout type de questionnement peut être traité par ces outils : des questions scientifiques en assemblant des corpus de sciences répondant à une requête donnée ou des questions de pilotages par l'étude d'un ensemble de rapports.

4 | L'ÉVALUATION INDIVIDUELLE DES CHERCHEURS ET DES CHERCHEUSES ET LA SCIENCE OUVERTE

Objectif : Repenser l'évaluation individuelle des chercheurs et des chercheuses avec d'une part l'utilisation d'une évaluation compatible avec les objectifs de la science ouverte et d'autre part la prise en compte de la contribution des chercheurs et des chercheuses à la science ouverte dans l'évaluation.

De nombreuses études et publications indiquent que le système actuel d'évaluation de la recherche est le frein principal à une transition vers la science ouverte. Il y a deux aspects à cela :

1. Les évaluations sont aujourd'hui largement basées sur des bibliographies. Apprécier la qualité des productions sur la base du prestige de la revue ou de l'éditeur revient à déléguer la responsabilité de l'évaluation aux « referees » mobilisés par les éditeurs. Dans certaines disciplines, ce prestige est basé sur des indicateurs « fermés et invérifiables » comme le facteur d'impact des revues. Dans ce cas, la responsabilité de l'évaluation est déléguée à des algorithmes « propriétaires » basés sur des données « fermées » ;
2. L'énergie qui doit être déployée et les ressources engagées par les chercheurs et les chercheuses pour que leur production soit ouverte autant que possible ne sont en aucune façon valorisées par les procédures d'évaluation.

Un certain nombre d'initiatives de terrain ont été prises. Parmi celles-ci, citons l'appel de Jussieu[1], la Déclaration de San Francisco[2] sur l'évaluation de la recherche (DORA), rédigée en 2012 par un groupe de directeurs et d'éditeurs de revues savantes, signée par de nombreuses organisations et individus.

Le CNRS l'a signée le 14 juillet 2018. Il s'agit d'un engagement, lequel consiste à éviter le recours à la bibliométrie et à préférer une évaluation plus qualitative, ainsi qu'à prendre en compte toute la variété des types de productions de l'activité de recherche.

Les sections et commissions du Comité National doivent déterminer et publier les critères qu'elles emploient pour leurs travaux d'évaluation des chercheurs et des chercheuses. Cependant, le CNRS souhaite que certains critères communs soient respectés par toutes. Certains de ces critères s'inscrivent dans le contexte de la mise en œuvre des politiques de science ouverte.

1. https://jussieucall.org/
2. bit.ly/SanFransiscoDORA

Dans ce cadre il convient que chaque section et chaque commission inscrive dans ses critères d'évaluation les quatre principes suivants :

1. **Ce sont les résultats eux-mêmes qui doivent être évalués, et non pas le fait qu'ils aient pu être publiés dans une revue prestigieuse ou autre média réputé : Les membres du CoNRS (Comité National de la Recherche Scientifique) doivent assumer la responsabilité de leur jugement et ne s'en remettent ni aux évaluations anonymes des éditeurs ni aux algorithmes. Ceci doit transparaître dans les rapports d'évaluation.**
2. **Pour chacune des productions citées dans les dossiers d'évaluation les chercheurs et les chercheuses doivent en expliquer la portée, l'impact, et la contribution personnelle qu'ils y ont apportée : L'exhaustivité de la liste des productions est inutile.**
3. **Tous les types de production doivent pouvoir être des éléments de l'évaluation[3] : En particulier, dans tous les cas où cela a un sens, les données sous-tendant la publication ainsi que le code source nécessaire à la production des résultats doivent pouvoir être fournis. Les « préprints » et autres documents de travail sont des productions acceptables pour l'évaluation. Il en va de même pour les « data papers » (« articles de données »).**
4. **Toutes les productions citées dans les dossiers d'évaluation doivent être accessibles dans HAL ou éventuellement dans une autre archive ouverte[4] : Il s'agit bien des productions elles-mêmes et non de leurs références. Il est normalement inutile de les fournir dans le dossier : le lien actif vers l'archive doit suffire.**

3. Voir les guides des produits de la recherche du HCERES : www.hceres.fr

4. Trois exceptions à cette règle sont recevables :
- Les résultats trop récents peuvent être sous embargo. Auquel cas ils doivent quand même avoir été déposés dans HAL, avec une durée d'embargo ne dépassant pas ceux prévus par la loi (6 mois en STM, et 12 mois pour les SHS). Ils sont alors fournis par un lien privé dans HAL (ou alors dans le dossier).
- Pour les recrutements, cette règle ne peut pas être absolue pour les candidats exerçant à l'étranger dans des institutions étrangères ou internationales, ou des institutions privées.
- Le type de production peut ne pas être accepté dans HAL.

5 | LA REFONDATION DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE POUR LA SCIENCE OUVERTE

L'information scientifique et technique s'est longtemps fondée sur des dispositifs privés, possédés par des entreprises à but lucratif faisant chèrement payer leurs services. Outre son coût économique, cette configuration limitait le contrôle des données, la capacité des opérateurs publics à les enrichir, les fusionner et les réutiliser.
La science ouverte, caractéristique par exemple, de l'Open citation initiative, doit permettre de changer cette situation et de rendre plus visible la recherche menée au CNRS tout en rationalisant* les processus d'acquisition des informations pour faciliter la tâche des chercheurs et des chercheuses et des personnels de recherche (limitation des enquêtes et des sollicitations multiples des unités) et aider au pilotage.

Action 1 : développer l'adhésion des chercheurs et des chercheuses à Orcid

Le CNRS a adhéré à Orcid en Mai 2019 en tant qu'institution et souhaite que l'usage de cet identifiant unique des personnels scientifiques soit largement adopté. Aussi, le CNRS mènera des actions afin d'inciter ses personnels scientifiques à obtenir un numéro ORCID et créer leur profil pour gagner en visibilité internationale. En contrepartie, il s'engage à clarifier les processus d'échange d'informations entre Orcid et les Systèmes d'Informations du CNRS « afin de leur simplifier la vie ».

Action 2 : travailler sur de nouveaux indicateurs bibliométriques

Le marché des bases de données sur les publications, les brevets ou d'autres productions scientifiques est en pleine transformation avec d'une part l'émergence de concurrents aux acteurs bien installés (Web of science Scopus), d'autre part la volonté de financeurs de la recherche, d'institutions de recherche et d'éditeurs de rendre plus accessibles et utilisables les métadonnées de ces objets, en complément de l'accès ouvert aux productions elles-mêmes. De plus, ces nouveaux acteurs, plutôt que de présenter une sélection des productions sur la base du « prestige » de ses lieux de certification, visent à donner des images les plus exhaustives possibles de l'état de la production, quel que soit le type d'objet considéré (article, ouvrage, conférence…) et la langue d'expression. Dans ce contexte, le CNRS réévaluera régulièrement ses besoins en matière d'outils et favorisera l'adoption et la diffusion d'outils en cohérence avec sa politique générale de science ouverte (signature de DORA, appel de Jussieu sur la bibliodiversité).

(*) Saisie par les détenteurs de l'information, pas de saisie en double...

6 | LA FORMATION ET LES COMPÉTENCES

Objectif : Accompagner et former tous les personnels de la recherche, en particulier les doctorants, et également proposer la formation des professionnels d'appui à la recherche en relation avec les branches d'activité professionnelle et les réseaux métiers concernés.

L'objectif de l'adoption et de l'appropriation des principes de la Science ouverte et des nouvelles pratiques liées à l'exercice de la recherche sera atteint si la communauté scientifique est accompagnée et aidée dans cette démarche. La compréhension des nouveaux usages et au final l'aide fournie dans l'accompagnement de nouvelles pratiques doivent passer par la compréhension de nouveaux enjeux de la recherche scientifique ouverte.

Les outils, les services et les infrastructures, mis en place par le CNRS notamment, sont des accélérateurs essentiels de la science ouverte. Toutefois, la simple présence de ces éléments ne signifie pas que la science ouverte deviendra la norme et sera mise en pratique. Pour ce faire, les chercheurs et les chercheuses et le personnel de soutien doivent posséder les connaissances et les compétences nécessaires pour adhérer aux nouvelles pratiques scientifiques, pour connaître et utiliser les outils, les services et l'infrastructures et pour accompagner leurs collègues pendant la phase de transition.

Compétences « science ouverte » **« Open Science skills »**[1]

Action 1 : développer les compétences et l'expertise nécessaires pour la publication en libre accès

- Plateformes ;
- Archives ouvertes (HAL et thématiques) ;
- Licences.

Action 2 : développer les compétences en matière de gestion de données de la recherche

- Services d'accompagnement relatifs à la production, l'utilisation/réutilisation des données, leur diffusion selon le principe « ouvertes autant que possible, fermées si nécessaires » ;
- Respect des principes FAIR ;
- Aide à la création de plan de gestion de données ou data management plan en anglais ;
- Travail sur les métadonnées, standards d'interopérabilité, etc.

Compétences techniques, en particulier en ingénierie des données, en sciences des données et en gestion des données.

Action 3 : développer les compétences « scientifiques » permettant une conduite de la recherche ouverte, y compris des compétences en intégrité de la recherche, en éthique et en droit.

Développer les compétences pour agir au sein et au-delà de sa propre communauté scientifique et disciplinaire.
Une offre significative des formations qui tendent à développer des compétences en science ouverte existe déjà. Le plan de futures formations doit se concentrer sur l'amélioration de la qualité et de la pertinence des compétences. Pour faciliter cette tâche, il sera nécessaire d'offrir et de promouvoir des cours de formation traditionnels et/ou en ligne et pour les scientifiques adaptés au niveau de leur carrière.

Action 4 : développer les compétences d'accompagnement dans les laboratoires pour l'analyse et la fouille des résultats

Le CNRS s'implique dans le développement des compétences nécessaires à l'usage d'outils permettant l'analyse des textes et des données et la visualisation des résultats de ces analyses. Il s'agit ainsi d'éveiller aux nouvelles pratiques documentaires et de renforcer la communauté d'utilisateurs de ces outils et technologies.

1. « Providing researchers with the skills and competencies they need to practice Open Science »- European Commission Open Science skills Working Group – juillet 2017

7 | LE POSITIONNEMENT INTERNATIONAL

Objectif : Prendre sa place au sein des organisations internationales afin de définir et partager les stratégies ainsi que les bonnes pratiques pour la science ouverte (tant au niveau des publications scientifiques que des données de la recherche).

La politique du CNRS s'inscrit dans un environnement international. La DIST est engagée dans de nombreuses actions et collaborations internationales, en tant que telle ou par l'intermédiaire des unités ou des individus. Au sein du CoSO, le Collège Europe & International a pour vocation de coordonner les acteurs français sur l'ensemble des actions, la veille et la prospective de manière transversale aux autres collèges. Le CNRS y prend une part très active.

Action 1 : aligner les positions CNRS avec le cadre européen et international de la Science ouverte sur le thème des données

La Science ouverte est une réflexion très largement internationale portée de façon descendante tant par les instances de politique scientifique (DG Connect, DG Recherche de l'Union Européenne) que par les structures de financement (Science Europe,…)), et de façon ascendante par les communautés scientifiques elles-mêmes (RDA, GO FAIR, Codata,…). La Research Data Alliance (RDA) comprend près de 9 000 membres de 133 pays qui sont regroupés en 88 groupes de travail ou d'intérêt. Elle a vu en 2019 la fondation de son « nœud français » par les experts du CNRS.
La déclinaison nationale de ces efforts s'organise autour du MESRI (CoSO, SPSO (secrétariat permanent pour la science ouverte) et ses différents collèges) en liaison avec le département du pilotage des infrastructures de recherche. Le CNRS prend d'ores et déjà toute sa place dans ces dispositifs pour y porter un point de vue multidisciplinaire et renforcera l'engagement des instituts qui portent des initiatives d'ouverture des données produites par leurs communautés scientifiques.
S'inscrire dans la mise en place de l'EOSC et des services de données qui s'y installent est un enjeu fort pour le CNRS.

Action 2 : soutenir les initiatives qui travaillent à définir les éléments de FAIRisation des données

La FAIRisation des données demande un travail au niveau des disciplines pour définir les formats de données et la manière de les décrire. Il faut aussi des travaux plus généraux sur les aspects technologiques et sociologiques du partage des données. Ces travaux doivent être menés au niveau international. Comme le note le Plan national pour la Science ouverte, la RDA décrit les bonnes pratiques du partage des données. Elle peut abriter les discussions disciplinaires lorsqu'il n'y a pas de cadre international pour celles-ci. Le CNRS soutiendra la participation de ses équipes à la RDA et aux initiatives disciplinaires.

Action 3 : communiquer avec nos partenaires européens et internationaux sur les stratégies de publication scientifique en accès ouvert

L'échange de l'information scientifique, le partage des résultats de la recherche à travers les publications scientifiques n'a de sens qu'au niveau international. La science ouverte ne peut avancer que de façon corrélée au niveau international. Le CNRS prend toute sa place dans cet environnement mondial.

Knowledge-exchange est une collaboration originale entre six partenaires de six pays européens qui valorisent la science ouverte et ont un rôle important - bien que différent selon les partenaires - dans la réalisation, le soutien ou le financement des infrastructures et des services numériques pour la recherche et l'enseignement supérieur. Avec ses cinq partenaires (JISC [UK], SURF [NL], DFG [GER], CSC [FIN] et DAFSHE [DK]), le CNRS, par l'intermédiaire de la DIST, peut ainsi co-produire des études et préconisations, échanger sur les pratiques, voire se coordonner, en impliquant des experts CNRS ou d'autres institutions. A titre d'exemple, un rapport sur l'usage des preprints vient d'être produit par Knowledge-Exchange.

Le CNRS, par l'entremise de la DIST et du CCSD, est membre actif de la confédération mondiale des archives ouvertes COAR (Confederation of Open Archive Repository), instance de première importance de coordination des initiatives visant à ce que la recherche conserve la maîtrise des informations qu'elle produit. Les standards d'interopérabilité des réservoirs de publications du futur s'y élaborent, permettant à de nouveaux services de se développer. L'assemblée générale 2019 a été organisée par le CCSD à Lyon.

Le CNRS participe régulièrement aux conférences de Berlin sur l'alignement des stratégies pour favoriser l'accès ouvert, qui ont réunis en décembre 2018 des participants de 37 nations et 5 continents (Chine, Afrique du Sud, Californie, Allemagne, etc.). Un texte de convergence a été produit entre l'appel de Jussieu qui prône la bibliodiversité, et le plan allemand OA2020. L'enjeu est d'aligner les critères de soutien aux journaux en accès ouvert, aux plateformes d'édition et aux infrastructures, et de coordonner les efforts d'investissement à un niveau stratégique. Notons que le plan S a modifié ses recommandations pour tenir compte des diverses voies de publication en accès ouvert sous la pression française notamment, et en organisant une large consultation internationale.

Références

Livre blanc une science ouverte dans une république numérique, DIST, CNRS (2016) :
bit.ly/booksopenedition

Plan National pour la Science ouverte (4 juillet 2018) :
bit.ly/plannational

Turning FAIR into reality, Final report and action plan from the European Commission expert group on FAIR data, European Commission (2018) :
bit.ly/turningFAIR

La déclaration de San Francisco DORA :
bit.ly/SanFransiscoDORA

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