INRAE DipSO
Direction pour la Science Ouverte

Principes pour la gouvernance des données

10 décembre 2020

Contexte

Les pratiques de recherche ont profondément évolué au cours des dernières années, avec un accroissement des exigences liées à l'accessibilité aux données et à la reproductibilité des résultats. Par ailleurs, l'intérêt stratégique des données et des développements technologiques permettant des approches nouvelles autour de celles-ci (data science, intelligence artificielle, …) sont devenus des préoccupations centrales des établissements d'enseignement supérieur et de recherche et de leurs partenaires publics comme privés.

Au sein de consortia nationaux ou internationaux, des plateformes historiques sont disponibles pour stocker et partager des jeux de données suivant des standards établis par la communauté scientifique internationale (NCBI, EMBL, etc.). Mais des entrepôts d'un nouveau type apparaissent dans le paysage pour héberger des données parfois sans standard partagé largement (Dryad, Zenodo, Data Terra, etc.), ainsi que pour partager des résultats sous forme de *pre-print* (Hal, arXiv, BioRxiv, etc.). Avec ces nouveaux instruments, de nouvelles pratiques de collaboration sont encouragées et se développent (Peer-Community In, F1000research, etc.). Cette dynamique est internationale renforçant ainsi les exigences de diversité, quantité, qualité et d'intégrité des données et des résultats de la recherche scientifique.

En France, elle s'inscrit dans le cadre légal et réglementaire en faveur d'une transparence accrue via le libre accès et la libre réutilisation des données publiques au bénéfice de la société, tels qu'ils sont prévus par la Loi pour une République Numérique d'octobre 2016. Le « Plan national pour la science ouverte », lancé en 2018 par le Ministère de l'enseignement supérieur, de la recherche et de l'innovation - piloté en inter-établissement par le « comité pour la science ouverte » (CoSo[1])- accompagne l'intégration de cette nouvelle donne dans les pratiques scientifiques.

Dans ce contexte, les agences de financement de la recherche publique françaises et européennes intensifient leurs attentes pour rendre accessibles et réutilisables les données et les publications issues des projets qu'elles co-financent. Via la rédaction de « plans de gestion de données », les financeurs institutionnels encouragent les bonnes pratiques de gestion et de partage des données (rendre les données F.A.I.R.[2]), dans le respect du droit. C'est le cas de l'Agence Nationale de la Recherche en France. Pour supporter ces nouvelles pratiques, des e-infrastructures telles que l'European Open Science Cloud (EOSC), soutenues par la Commission européenne, sont en cours de développement. Les nouveaux contrats de subvention européens Horizon Europe prévoient à ce titre des engagements formels à la FAIRisation des données et à leur stockage et partage sur des plateformes dédiées.

Ainsi, INRAE s'engage dans une politique de science ouverte dans le respect des dispositions législatives et contractuelles considérant les opportunités qu'elle offre pour :

- Le partage des savoirs et de leur validation,
- L'intégrité des pratiques scientifiques incluant la reproductibilité des résultats,
- La réutilisation des données produites pour élaborer de nouvelles connaissances à partir de résultats validés, et une meilleure maitrise des couts de la recherche,
- L'encouragement à la collaboration pour associer et développer les meilleures compétences,
- La valorisation des résultats de la recherche pour les sphères académique, économique, de politiques publiques et sociétale.

[1] https://www.ouvrirlascience.fr/college-donnees-de-la-recherche/

2 Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 http://dx.doi.org/10.1038/sdata.2016.18 (2016)

Nous retiendrons ici la définition des données de la recherche proposée par le Plan national pour la science ouverte : « les enregistrements factuels (chiffres, textes, images, son, vidéo…) utilisés comme sources primaires pour la recherche et qui sont habituellement acceptés par la communauté scientifique comme étant nécessaires pour valider les résultats de la recherche ». Les codes sources font également partie des données de la recherche.

Si l'ouverture de l'accès aux données de la recherche constitue une valeur fondamentale de la recherche scientifique, elle peut comporter des risques de différentes natures : éthiques ou déontologiques (y incluant les données à caractère personnel), démocratique, juridique, financier, économique, réputationnels, concurrentiels, …. Qu'il s'agisse de nouvelles connaissances ou du développement de nouveaux produits ou services, la création de valeur à partir des données ouvertes ainsi encouragée peut être réalisée par des tiers concurrents au détriment de notre institut. Il est donc indispensable d'analyser l'impact de l'ouverture (voir par exemple cet article sur les données de biodiversité). Par ailleurs, la minimisation de l'impact environnemental inhérente à l'accroissement exponentiel de la quantité des données produites par et pour la science est au cœur des préoccupations d'un établissement comme le nôtre.

Ainsi, « Gouverner » les données signifie que des (bonnes) décisions sont prises à chaque étape du cycle de vie des données (de leur production ou réutilisation à leur valorisation) et que les responsabilités de chacun sont définies dans ce processus décisionnel.

Cette note vise à énoncer les principes en matière de « gouvernance des données » en définissant les rôles et responsabilités des acteurs. Elle constitue un guide permettant d'identifier les questions qu'il faut se poser afin de mieux gérer et valoriser les données sur la base de critères scientifiques, juridiques, économiques, techniques et de politique scientifique. Elle présente également un processus pour leur instruction en vue d'aboutir à une décision sur les modalités d'ouverture. Elle est accompagnée de fiches techniques consultables sur le site datapartage.inrae.fr fournissant une information pratique.

Quatre principes fondent les bases de la politique de l'institut et régissent les décisions à prendre lors du cycle de vie des données. Ils forment les axes d'un système de décision cohérent en étudiant une à une les dimensions scientifique (valeurs déontologiques et éthiques), technique (gestion des données mettant en avant les bonnes pratiques à respecter), réglementaire (respect de la législation et des engagements contractuels), et d'innovation (valeur produite pour la société). C'est à l'issue d'un examen approfondi selon ces principes que le degré d'ouverture pourra être décidé.

Principes

1. Principe 1 : Il faut partager et réutiliser les données en respectant les valeurs de la science

Les fondements déontologiques de la recherche sont profondément renforcés par le concept de Science Ouverte qui insiste sur la disponibilité des données utilisées pour produire des résultats scientifiques. Enrichir une connaissance collective en partageant ses résultats est au cœur de l'éthique du chercheur. En partageant les données, le chercheur facilite la reproduction de ses résultats, preuve de son intégrité scientifique. Ainsi, la recherche et la démarche scientifique reposent plus que jamais sur des valeurs de rationalité et de rigueur collectivement définie (peer reviewing). Ces données issues de la recherche possèdent une valeur indéniable pour la recherche, mais sa valeur est également à considérer dans les contextes de l'expertise, de l'appui aux politiques publiques, et de la formation.

La responsabilité de l'ouverture des données reste portée en premier lieu par le chercheur qui doit en particulier :

- Veiller au cadre éthique (valeurs sociétales, valeurs d'établissement, …) et déontologique (rigueur scientifique, reproductibilité, …) dans lequel s'inscrit la production et le partage des données.
- Déposer ses données dans des entrepôts disciplinaires ou dans http://data.inrae.fr afin de leur attribuer un identifiant unique (ex : DOI, URI, etc.) pour permettre de les citer.

- Décrire clairement les données sous forme de métadonnées pour contribuer à les rendre trouvables (findable) et intelligibles pour les réutilisateurs potentiels. La qualité des métadonnées, comme celle des données doivent être garanties par le chercheur.
- Pouvoir valoriser son travail en préservant sa capacité à le poursuivre. Ainsi lorsque les données sont « ouvrables » selon les critères définis par les autres principes, il est possible d'en différer l'ouverture afin de protéger une exploitation scientifique des données par les partenaires du projet ou pour des raisons de sécurité publique. La durée du délai doit toutefois correspondre aux règles éthiques et déontologiques, aux pratiques des communautés, ainsi qu'aux exigences des financeurs, des partenaires et des autorités publiques.

Dans le montage des projets de recherche, notamment européens, la réutilisation de jeux de données produites par d'autres est encouragée. C'est une garantie de rigueur scientifique, mais aussi d'économie importante. Il convient donc de rechercher des données, à l'instar d'une recherche bibliographique ou de brevets, de s'assurer de leur possibilité de réutilisation et de citer les jeux de données conformément à la licence qui leur est associée.

2. Principe 2 : Les données doivent être gérées en vue de les rendre F.A.I.R

Qu'elles soient destinées à être librement accessibles (ouvertes) ou à usage restreint, les données doivent être considérées comme un patrimoine scientifique commun, à l'échelle de l'établissement, qu'il convient de gérer pour les rendre F.A.I.R. Cela suppose d'adopter des bonnes pratiques dès leur création pour être capable de gérer correctement leur partage, leur destruction ou leur archivage. Cette bonne gestion est un gage de qualité qui contribue à la reproductibilité des résultats et, le cas échéant, est également source d'économies[3].

Une bonne gestion des données consiste à :

- Réaliser des plans de gestion des données (« data management plan ») ou des plans de partage de données (« data sharing plan »[4]). Ces outils contribuent à la qualité des projets scientifiques et des résultats obtenus en permettant, dès la conception du projet de se poser les bonnes questions sur les modalités de gestion et de valorisation des données, en abordant des aspects techniques, juridiques, éthiques, scientifiques et économiques (dont l'estimation des coûts de gestion et de FAIRisation qui doivent être insérés dans les budgets de projets).

- Mettre en œuvre les principes F.A.I.R [5] requis par la plupart des financeurs de la recherche publique nationale et européenne. Ceci est d'autant facilité par une prise en compte dès l'étape de production des données. Les principes de base sont les suivants :
  - Quel que soit le mode de persistance des données, celles-ci doivent être identifiées de manière non ambiguë (notamment avec des DOI), suffisamment documentées, dans des formats et des systèmes non propriétaires.
  - Pour les systèmes d'information scientifiques, il faut s'assurer de :
    - Une architecture qui garantit l'indépendance des données vis à vis des traitements et des interfaces utilisateurs pour favoriser leur évolutivité.
    - Une gestion des droits d'accès avec des profils adaptés aux différents usages, dont des usages gratuits et anonymes

[3] Cost of not having FAIR Data – Study, European Commission, janvier 2019, http://dx.doi.org/10.2777/02999

[4] Par exemple celui de l'ERC http://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-erc-tpl-oa-data-mgt-plan_en.odt

[5] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1). https://doi.org/10.1038/sdata.2016.18

3. Principe 3 : Les données doivent être « aussi ouvertes que possible, aussi fermées que nécessaire »

L'accès aux données publiques et les modalités de leur libre réutilisation sont actuellement principalement organisés par les dispositions de la Loi pour une République numérique et de la Loi Valter. Ces lois marquent un tournant majeur car elles indiquent, en substance, que **sauf exception, les données produites par une administration dans le cadre de sa mission de service public** sont par défaut ouvertes et réutilisables gratuitement. L'administration doit également **publier son catalogue des données.** Les **exceptions à l'ouverture** sont liées à d'éventuelles clauses partenariales limitantes et/ou à la nature de certaines données, notamment lorsqu'elles font courir un risque pour autrui. C'est donc un signal fort qui nécessite un renforcement de nos pratiques d'ouverture.

En combinant les cadres juridiques en matière de données, un schéma de diagnostic en trois zones est proposé :

- Zone verte : **Il n'existe pas de texte juridique** (contrat ou loi) qui limite l'usage des données, et les données ne sont pas soumises à une réglementation spécifique. Les conditions de l'accès et de réutilisation gratuites des données par des tiers sont alors réunies.

- Zone rouge : Il existe un texte juridique (contrat ou loi) qui **interdit l'usage** et la gestion des données par des tiers, ou elles sont soumises à une réglementation limitante spécifique (données à caractère personnel -RGPD, nature sensible des données, questions de sécurité nationale, etc.). L'accès et la réutilisation des données sont interdits.

- Zone orange : Il s'agit des cas intermédiaires à approfondir. Par exemple, il existe un texte juridique (contrat ou loi) qui **peut limiter l'usage** et la gestion des données par des tiers, ou la nature des données peut justifier de limiter l'accès et la réutilisation.

Ce schéma diagnostic est fondé sur des arguments juridiques (voir fiche technique « juridique » et le logigramme associé[6]). Notons que si les données sont mises à jour régulièrement, les modalités de leur mise à jour et d'accès aux nouvelles versions doivent être clairement définies ainsi que la mise à jour des licences le cas échéant.

4. Principe 4 : L'ouverture des données contribue à l'innovation et à la création de valeur pour la société

Les données (fichiers, bases de données et codes sources de logiciels) répondant aux critères juridiques d'ouverture doivent être rendues accessibles et réutilisables gratuitement. Il s'agit d'encourager la création de valeur par les acteurs du secteur privé ou public, français ou étranger. Un service d'accompagnement à l'utilisation des données (formations, conseils) peut éventuellement être facturé par l'institut pour couvrir ses frais.

Ainsi les données peuvent être utilisées par des tiers pour créer de la valeur à partir de :

- Création de services sur les données en intégrant les données avec d'autres informations,

[6] Lien vers la fiche technique juridique.

- Intégration de codes informatiques dans des applications professionnelles ou « grand public »,
- Applications en intelligence artificielle entrainées sur les données mises à disposition.

L'ensemble peut alimenter des systèmes d'aide à la décision en agrégeant données, codes informatiques, et intelligences artificielles générant ou supportant une activité économique pour de grands groupes, mais aussi des petites structures comme des start-ups, ou encore contribuer à l'appui aux politiques publiques auprès de ministères, agences publiques ou collectivités territoriales.

L'organisation de Hackathons, Challenges, Datathon, Bring Your Own Data peut permettre d'animer un écosystème d'utilisateurs et de comprendre leurs besoins et les innovations qui en découlent. C'est aussi l'opportunité de recueillir de nouvelles idées pour nos projets sur le plan scientifique ou technologique.

Le processus de décision

Sur la base de ces principes, un processus de décision est construit en identifiant les différents acteurs, leurs rôles et leurs responsabilités. Ce processus a vocation à être précisé ou ajusté au regard des retours d'expérience à venir, en s'attachant à le laisser le plus fluide possible et à s'assurer que les objectifs visés en matière d'ouverture soient atteints.

Quatre principes conduisant à une décision

Ces quatre principes forment une logique d'ensemble. Ils doivent être examinés tour à tour pour permettre une décision éclairée. Ils permettent d'appréhender le sujet de l'ouverture sur l'ensemble de ses facettes en tenant compte des points de vue des acteurs (scientifiques, société civile, économiques et politiques). C'est en examinant les conditions d'ouverture suivant ces 4 dimensions, que la décision finale peut/doit être obtenue, chaque principe posant ou non d'éventuelles restrictions et indiquant un degré d'ouverture envisageable. Sans les mettre en contradiction les uns avec les autres, ces éléments à prendre en compte permettent de déterminer les options les plus appropriées en matière de gestion et diffusion des données.

Au travers de ces principes, la responsabilité individuelle des chercheurs d'un établissement public comme INRAE est mise en avant, mais la dimension collective de la réflexion éthique et déontologique reste nécessaire. En effet même si les données appartiennent la plupart du temps aux établissements qui ont dégagé les ressources et les compétences nécessaires pour les produire, les chercheurs à l'origine de la production de ces données restent les garants de leur qualité, de leur bonne utilisation, de leur gestion et de leur FAIRisation. Leur rôle est central pour choisir une diffusion appropriée à leur communauté. Les risques d'une réutilisation erronée des données, de bonne foi, voire avec une intention de malveillance ou de désinformation volontaire, seront d'autant plus limités que le choix des supports des données ouvertes aura été rigoureux (support reconnu et connu, reviewing par les pairs, informations contextuelles complètes et solides).

Finalement, il en résulte que l'ouverture des données se place sur un spectre du plus fermé au totalement ouvert (Figure 1). Ainsi, on distingue :

1. Des données fermées : Ce sont des données à usage strictement interne à l'institut.
2. Des données partagées : Il s'agit de données dont l'usage par des personnes externes à l'institut est possible sous certaines conditions. Cet usage est donc restreint et peut être contrôlé via une authentification ou un dispositif qui permet d'identifier le réutilisateur.
3. Une modalité de contenu : Il est possible de n'ouvrir qu'une partie des données, ou de n'ouvrir qu'après traitement des données (bruitage, modification d'échelle, anonymisation …), si tant est que cette information garde du sens sur le plan scientifique.
4. Des données ouvertes : Ce sont les données réutilisables par tous, gratuitement, et sans aucun contrôle.

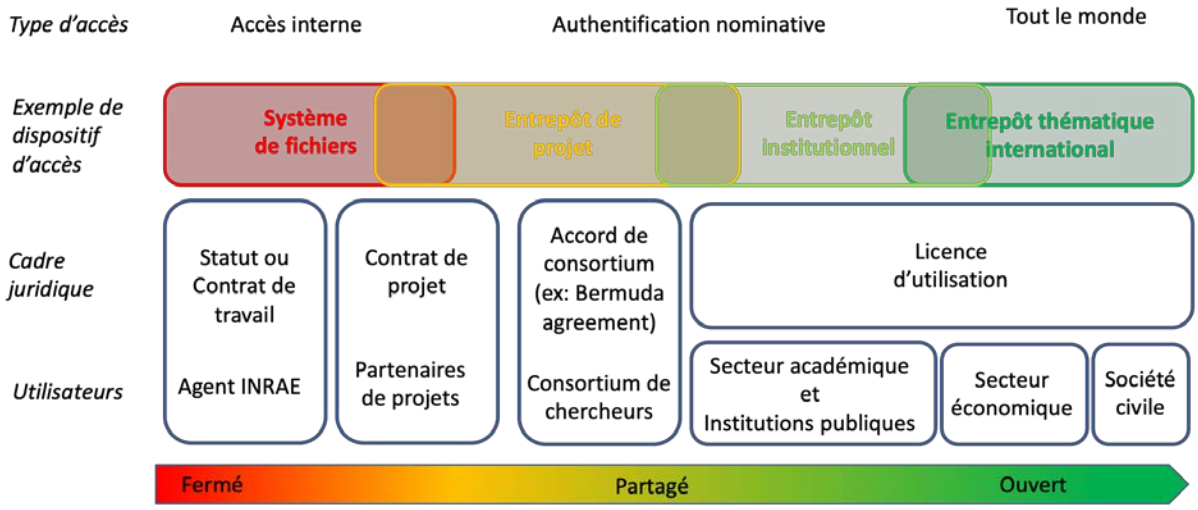

Figure 1: Spectre d'ouverture des données

Une attention particulière sera portée sur l'analyse de l'impact de cette ouverture sur les sphères scientifique, politique, économique, et sociétale. Analyser les aspects positifs et négatifs sera important pour faire évoluer le dispositif dans le sens désiré.

Rôles et responsabilités

Les acteurs de la chaine de production des données

Plusieurs processus de travail sur les données sont liés et s'alimentent : (i) la tenue d'un Plan de Gestion des Données (PGD), (ii) la démarche de FAIRisation des données et (iii) la décision sur les modalités d'ouverture des données.

(i) Le PGD, comme les questions de clauses partenariales, est pris en charge, pour l'essentiel, par le porteur du projet scientifique dès le montage de celui-ci, avec l'appui local des réseaux fonctionnels concernés (IST[7], CPI[8], DPO[9] si données personnelles, etc.). Le porteur du projet scientifique organise et s'assure de la mise en œuvre du PGD au sein de l'équipe.

(ii) La FAIRisation des données se situe également dans le domaine de responsabilité directe du porteur du projet scientifique, comme partie intégrante de la conduite du projet, voire comme la réponse à un engagement formel pris vis-à-vis du bailleur (ANR, Europe…). Le porteur du projet scientifique organise et s'assure de la réalisation des degrés de FAIRisation retenus notamment en mobilisant les acteurs des CATI[10], d'ISC ou d'IR qui apportent leurs connaissances et compétences des standards internationaux des communautés scientifiques en matière de FAIRisation des données (entrepôts de données de référence, éditions de data papers, etc.).

[7] IST : Information Scientifique et Technique

[8] CPI : Chargés de Partenariat et d'Innovation

[9] DPO : Délégué à la Protection des données

[10] CATI : Un Centre Automatisé de Traitement de l'Information est une modalité d'organisation de la production informatique et de la production de services en soutien à la production scientifique sur une thématique donnée ou relative à l'appui à la recherche, nécessitant l'intégration de compétences informatiques adaptées pour conduire l'intégralité du processus (« de la donnée brute à la valorisation »).

(iii) Les départements scientifiques INRAE proposent des recommandations pour la mise à disposition ou de réutilisation de données produites par d'autres, et leur gestion en phase avec leur stratégie scientifique.

(iv) Le processus de décision sur les modalités d'ouverture des données est la conséquence directe du PGD et en cohérence avec le degré de FAIRisation des données. Le porteur du projet scientifique décide et organise les modalités d'ouverture des données, en cohérence avec les grandes orientations définies par les départements scientifiques.

Dans le cas de projets inter-établissements, ou dans le cadre d'un partenariat européen ou international, le suivi des bonnes pratiques en matière de PGD, de FAIRisation des données et d'ouverture, est assuré par la personne représentant INRAE au niveau du consortium des partenaires, en lien avec le(s) département(s) impliqués. Il lui revient également de s'assurer de la bonne mise en œuvre des options retenues.

Certains projets, par leur ambition scientifique et les moyens engagés, nécessitent - dès le montage du projet - une réflexion et des choix en matière de propriété et d'usage des données, qui dépassent le cadre d'une unité ou des appuis fonctionnels disponibles à l'échelle d'un centre INRAE. Dans ce cas, l'avis du ou des départements impliqués est sollicité par le Directeur d'Unité et le porteur du projet scientifique. Les départements interrogés s'appuyant éventuellement sur un réseau de CPI[11], proposent alors une stratégie de valorisation.

Si le(s) département(s) considère(nt) qu'il(s) ne disposent pas de toutes les expertises nécessaires, ils sollicitent l'avis de l'Administrateur des Données Scientifiques placé auprès de la Directrice Générale Déléguée à la Science et à l'Innovation (DGDS&I).

Une Fiche Technique « Rôles et responsabilités » reprend plus en détail ces dispositions.

Cellule « Gouvernance des données » et Administrateur des données scientifiques

Les problématiques multidisciplinaires (scientifique, technique, juridique, éthique) liées à la question des données, ainsi que la nécessité d'une animation dans le temps et à l'échelle de l'établissement, justifient la création d'une Cellule Nationale « Gouvernance des données », animée par l'Administrateur des données scientifiques, et appuyé par la Direction pour la Science Ouverte (DipSO).

L'Administrateur des données scientifiques, assisté par la Cellule Nationale « Gouvernance des données », pourra rassembler autour des cas complexes qui lui seront soumis, l'ensemble des compétences INRAE concernées pour instruire les situations difficiles d'ouverture des données. Au besoin, en cas de situation particulièrement complexe techniquement ou à forte dimension politique, avec un risque juridique significatif en matière d'ouverture des données, un arbitrage de la Direction Générale pourra être demandé.

Cette Cellule Nationale de 3-4 personnes, mise en place à l'automne 2020, sera chargée de la coordination des différents acteurs internes (DipSO[12], PRADA[13], DPO[14], RSSI[15], FSD[16] … et experts scientifiques en fonction des sujets) pour :

[11] Notons que certains départements ont initié une dynamique de nomination d'une personne « correspondante données » qui peut être sollicitée dans ce cadre. Cette initiative pourrait être étendue à l'ensemble des départements en réfléchissant, le cas échéant, à des mutualisations.

[12] DipSO : Direction pour la Science Ouverte.

[13] PRADA : personne responsable de l'accès aux documents administratifs.

[14] DPO : Délégué à la Protection des données

[15] RSSI : Responsable Sécurité des Systèmes d'Information

[16] FSD : Fonctionnaire Sécurité Defense

- Mobiliser ponctuellement les acteurs compétents pour l'instruction de situations non résolues à l'échelle « locale » en vue d'un arbitrage éventuel de la Direction Générale.
- Proposer des mises à jour des règles de fonctionnement de la gouvernance des données.

Des fiches techniques pédagogiques

Des fiches pédagogiques sont accessibles sur le site de datapartage (https://www6.inrae.fr/datapartage).

Elles ont pour objectif :

- De donner des points de repères, bonnes pratiques
- De lutter contre les idées reçues
- De renvoyer vers des experts