En route vers la Science Ouverte - Gérer les données de sa recherche

Germain Faity, Frédéric de Lamotte, Alexandre Dehne Garcia

En route vers la Science Ouverte

Gérer les données de sa recherche

Germain Faity, Jean-François Martin, Alexandre Dehne Garcia, Frédéric de Lamotte

12 mars 2020

Ce document résume en quelques étapes la démarche d'ouverture de ses données dans une perspective de Science Ouverte. D'après la formation « En route vers la Science Ouverte : Gérer les données de sa recherche » de Montpellier SupAgro (22 janvier 2020). Formation dispensée par Jean-François Martin, Alexandre Dehne Garcia, Frédéric de Lamotte.

A propos des auteurs

Germain Faity étudie la neurophysiologie computationnelle du mouvement dans le cadre de son doctorat à l'Université de Montpellier. Sa curiosité l'a amené à se poser des questions sur les problématiques de sciences ouverte et d'intégrité scientifique. ORCID : https://orcid.org/0000-0002-6855-8938

Jean-François Martin est maître de conférences à Montpellier SupAgro. Spécialisé dans l'écologie moléculaire et la biologie évolutive il est intéressé par les questions méthodologiques de la bioinformatique et de l'analyse de données. Il est également référent intégrité scientifique de Montpellier SupAgro. ORCID : https://orcid.org/0000-0001-9176-4476

Alexandre Dehne Garcia est co-animateur de l'UAR INRAE Ingenum (Ingénierie pour le numérique) et ingénieur à la Direction pour la Science Ouverte de l'INRAE. ORCID : http://orcid.org/0000-0002-6688-3593

Frédéric de Lamotte est chercheur à INRAE. ORCID : http://orcid.org/0000-0003-4234-1172

Sommaire

Table des matières

Etape 0 – Qu'est-ce que la science ouverte ?

La science ouverte, c'est la diffusion sans entrave des publications et des données de la recherche. Elle s'appuie sur l'opportunité que représente la mutation numérique pour développer l'accès ouvert aux publications et – autant que possible – aux données de la recherche. **Son objectif** est de faire sortir la recherche financée sur fonds publics du cadre confiné des bases de données fermées. La science ouverte cherche à **augmenter l'efficacité de la recherche** en simplifiant la collecte, la création, le transfert et la réutilisation du matériel scientifique.

La science ouverte vise à construire un écosystème dans lequel la science est plus cumulative, plus fortement étayée par des données, plus transparente, plus rapide et d'accès plus universel. Elle induit une démocratisation de l'accès aux savoirs, utile à la recherche, à la formation, à l'économie, à la société. Elle favorise les avancées scientifiques ainsi que l'innovation, les progrès économiques et sociaux, en France, dans les pays développés et dans les pays en développement. Elle construit un levier pour l'intégrité scientifique et favorise la confiance des citoyens dans la science.

La France est active dans ce domaine. On pourra citer les 3 axes du **Plan National Français pour la Science Ouverte** (2018) :

- Généraliser l'accès ouvert aux publications (fonds pour la science ouverte, HAL…)
- Structurer et ouvrir les données de la recherche (diffusion ouverte des données financées sur fonds publics obligatoire)
- S'inscrire dans une dynamique durable, européenne et internationale (développer les compétences science ouverte aux seins des écoles doctorales, contribuer à la structuration européenne au sein du *European Science Cloud* et par la participation à *GO FAIR*…)

On pourra également citer la **loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, dite « loi Lemaire » :**

- Favoriser la circulation des données et du savoir à travers l'ouverture des données publiques et d'intérêt général, la création d'un service public de la donnée et le libre accès aux écrits de la recherche publique.
- Ouverture des données publiques par défaut.

L'article de Fecher & Friesike (2014) décrit les 5 courants de pensée principaux de la science ouverte. Chacun peut se sentir plus proche de l'un ou l'autre de ces courants… ou créer le sien !

1. **Infrastructure school** : l'efficacité de la recherche dépend de la disponibilité des outils et des applications. Cette école prône les outils collaboratifs et en open-access.
2. **Public school** : la Science doit être accessible au public. Cette école prône la science citoyenne.

3. **Measurement school** : les métriques utilisées actuellement pour mesurer les contributions scientifiques mènent à trop de dérives, et doivent donc changer. Elle prône une évaluation plus qualitative que quantitative, notamment avec l'abandon du facteur d'impact.
4. **Democratic school** : l'accès au savoir est inégal, et doit être équilibré notamment en le rendant gratuit et accessible à tous.
5. **Pragmatic school** : la création de savoir scientifique serait plus efficiente si les scientifiques collaboraient davantage. Elle prône une science collaborative à travers l'ouverture de ses données et de son code.

En résumé, la science ouverte repose sur **7 piliers principaux** : publications ouvertes, données ouvertes, sources et pratiques ouvertes, éducation ouverte, évaluation responsable, science citoyenne, intégrité scientifique.

Pour aller plus loin

GO FAIR : lien

Enjeux et bénéfices de la science ouverte : lien

Ministère de l'enseignement supérieur, de la recherche et de l'innovation. (2018). Plan National Français Pour La Science Ouverte : lien

Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique : lien

Fecher, B., & Friesike, S. (2014). Open Science : One Term, Five Schools of Thought. In S. Bartling & S. Friesike (Éd.), Opening Science : The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing (p. 17-47). Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_2

Etape 1 - Science Ouverte et Intégrité Scientifique

Cette partie est inspirée du MOOC « Diffuser une culture de l'intégrité scientifique » proposé par l'Université de Bordeaux sur la plateforme FUN-MOOC.

Le Code de conduite européen pour l'intégrité en recherche (édition révisée), ALLEA (All European Academies), Berlin, 2018, page 4 décrit les principes fondamentaux de l'intégrité scientifique :

> *« Les bonnes pratiques en matière de recherche reposent sur des principes fondamentaux en matière d'intégrité en recherche. Ces principes orientent les chercheurs dans leurs travaux ainsi que dans leur engagement envers les enjeux pratiques, éthiques et intellectuels inhérents à la recherche. Ces principes sont les suivants :*
>
> - ***Fiabilité**, autrement dit garantir la qualité de la recherche, qui transparaît dans la conception, la méthodologie, l'analyse et l'utilisation des ressources.*
> - ***Honnêteté**, autrement dit élaborer, entreprendre, évaluer, déclarer et faire connaître la recherche d'une manière transparente, juste, complète et objective.*
> - ***Respect** envers les collègues, les participants à la recherche, la société, les écosystèmes, l'héritage culturel et l'environnement.*
> - ***Responsabilité [partagée entre les acteurs de la recherche]** assumée pour les activités de recherche, de l'idée à la publication, leur gestion et leur organisation, pour la formation, la supervision et le mentorat, et pour les implications plus générales de la recherche. »*

Certains chercheurs ont été reconnus de manquements à l'intégrité scientifique. Les fraudes (fabrication de données, falsification de données et de résultats, plagiat et auto-plagiat) lorsqu'elles sont avérées ont souvent une couverture médiatique large (voir Andrew Wakefield, Shinichi Fujimura, Diederil Stapel, Jan Hendrik Schön, Hwang Woo-Suk…). Antoine De Daruvar[1] nous informe sur les conséquences de telles fraudes :

> *« [Ces pratiques vont] déboucher sur la publication de résultats faux, avec toutes les conséquences que cela entraîne, car une fois que ces résultats faux sont publiés, ils vont être potentiellement réutilisés par d'autres créant ainsi un phénomène de propagation. Ils peuvent être utilisés pour prendre des décisions et peuvent avoir des conséquences importantes sur la société. Donc ces fraudes sont extrêmement graves. »*

Jean Jouzel[2] ajoute que :

> *« Comme dans toute discipline, tout le monde n'est pas d'accord, ce qui est normal, mais il faut pouvoir aboutir à un état des lieux crédible.* ***Sans éthique scientifique, dans ce domaine comme dans d'autres, nous pouvons perdre la confiance du citoyen, mais aussi celle du décideur politique,***

[1,2] MOOC « Diffuser une culture de l'intégrité scientifique »

qui est importante dans ce domaine, puisque des débats ont lieu et qu'ils doivent reposer sur une éthique irréprochable, que l'on soit pour ou contre. »

Cependant, ces fraudes sont très minoritaires, devant ce qui est appelée la **« zone grise » de l'intégrité scientifique**. Il existe en effet un continuum entre les fraudes et les bonnes pratiques. Ce continuum comporte des pratiques douteuses (méconduites) plus difficiles à interpréter, reconnues comme un manque de rigueur : on trouvera la suppression de données aberrantes sans réelle raison scientifique, la non-prise en compte de biais inhérents à une étude, les pratiques statistiques douteuses telles que le p-hacking, la citation incorrecte d'auteurs, mais aussi plus largement l'utilisation souvent involontaire d'une méthode peu rigoureuse. On notera que la non-déclaration de conflits d'intérêts est en passe de devenir un manquement caractérisé à l'intégrité scientifique, c'est-à-dire une fraude. Les conséquences de ces pratiques sont la difficile reproductibilité des travaux publiés.

Une étude publiée dans Nature (Baker, 2016) conduite sur plus de 1500 chercheurs montre que :

- 70% des chercheurs interrogés ont déjà échoué à reproduire une étude faite par d'autres scientifiques ;
- Plus de la moitié des chercheurs ont déjà échoué à reproduire une de leurs propres études.

Il est rapporté que les méconduites scientifiques sont à l'origine de cette crise de la reproductibilité (favorisée par la pression à la publication). La vidéo suivante montre de façon comique les différentes raisons qui peuvent rendre une étude non reproductible, même avec toute la bonne volonté du monde : https://www.youtube.com/watch?v=N2zK3sAtr-4. On y voir par exemple le problème de la non disponibilité des données même après demande aux auteurs (perte des données, fichiers non lisibles, formats obsolètes…).

D'après Yannick Lung, l'intégrité scientifique renvoie à une façon de faire de la bonne science. Cela implique une rigueur et une honnêteté intellectuelle de l'ensemble des équipes qui participent à la recherche. En diffusant votre travail en open-data, vous permettez à des confrères d'évaluer votre travail et ainsi de vérifier des erreurs volontaires ou involontaires commises dans votre démarche. En conséquence, **s'opposer à la diffusion de ses données peut diminuer leur reproductibilité et peut ainsi être considéré comme un manquement à la rigueur scientifique.**

Si vous avez un doute sur un manquement possible à l'intégrité scientifique, privilégiez le dialogue ou demandez conseil auprès du référent à l'intégrité scientifique de votre établissement. La liste des référents est disponible à cette adresse :
https://www.hceres.fr/fr/liste-des-signataires-des-chartes-et-des-referents-integrite-scientifique

Pour aller plus loin

MOOC « Diffuser une culture de l'intégrité scientifique » proposé par l'Université de Bordeaux sur la plateforme FUN : lien

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature News, 533(7604), 452. https://doi.org/10.1038/533452a

Déclaration de Singapour (2010) : lien

Charte nationale de déontologie (2015) : lien

Rapport Corvol (2016-17) : lien

Office Français de l'intégrité scientifique (2017) : lien

Guide pour une recherche intègre et responsable (CNRS-CPU) : lien

Pannucci, C. J., & Wilkins, E. G. (2010). Identifying and Avoiding Bias in Research. Plastic and reconstructive surgery, 126(2), 619-625. https://doi.org/10.1097/PRS.0b013e3181de24bc

Harvard guide to using sources : *lien*

Etape 2 – Et au niveau juridique ?

Ce chapitre est inspiré de l'intervention du Dr. Agnès Robin lors de la formation « Science Ouverte » décrite en première page. Merci à elle pour son intervention !

Qu'est-ce qu'une donnée ? D'après l'arrêté sur l'enrichissement du vocabulaire de l'informatique (22 déc. 1981), une donnée est une « représentation d'une information sous une forme conventionnelle destinée à faciliter son traitement ». Pour les données de la recherche, il existe une définition spécifique proposée par l'OCDE : « Les « données de la recherche » sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche. »

1) Droit de la propriété intellectuelle

Cette sous-partie ne concerne pas directement les données, mais plutôt les autres produits de la recherche (bases de données, articles, etc.).

Tout le monde connaît le plagiat. Mais savez-vous qu'il existe de nombreuses formes possibles de plagiat ? Cliquez-moi pour voir une introduction au plagiat dans la bande dessinée !

Droit d'auteur : comme vous le savez, il est interdit d'utiliser un son, une image, une vidéo comme vous le souhaitez. Le droit d'auteur protège les auteurs d'un vol de leur travail jusqu'à 70 ans *post mortem auctoris*. Plus précisément, la loi protège la mise en forme des idées/données mais pas directement les idées, la récolte des données ou leur exploitation. La loi du 11 mars 1957 sur la propriété littéraire et artistique précise que le droit d'auteur est constitué :

- **Des droits moraux** : ce sont des droits irrévocables (ne peuvent pas être cédés), perpétuels et imprescriptibles (sans limite de temps) s'appliquant à la personne physique. Ils englobent le droit de paternité, le droit de divulgation, le droit au respect de l'intégrité de l'œuvre et le droit de retrait et de repentir.
- **Des droits patrimoniaux** : ils permettent à l'auteur ou à ses ayant droit d'exploiter son œuvre sous quelque forme que ce soit, contre rémunération. Ce droit peut être cédé à travers un contrat par exemple. On y trouvera le droit de représentation, le droit de reproduction, le droit d'adaptation et le droit de traduction. Attention tout de même, il est impossible de céder à l'avance des œuvres futures qui ne sont au jour de formation du contrat, non déterminée ou non déterminables (sauf dans le cas des logiciels).

En réalité, les agents soumis à autorité hiérarchique ne conservent que les droits de paternité de leur œuvre, les droits patrimoniaux et les autres droits sont exercés par l'employeur public (sauf chercheurs,

enseignant-chercheurs et catégories d'agents non soumis à autorité hiérarchique, qui conservent la plénitude de leurs droits d'auteur). Nous reviendrons sur cette exception en conclusion.

Il existe de nombreuses exceptions au droit d'auteur, mais en résumé, pour citer le travail de quelqu'un en respectant le droit d'auteur, on utilisera le droit de citation : est autorisée une courte citation, mise entre guillemets, avec sources et dans le respect de la finalité de l'auteur premier. Pour en savoir plus sur la façon de citer, « A guide to ethical writing » (Roig, 2015) rapporte 28 guidelines très intéressantes pour écrire de façon éthique.

Attention, le droit de citation n'est pas applicable à l'image. Par conséquent, il est illégal d'intégrer une image (graphique, schéma, photo…) dans une publication sans avoir au préalable obtenu l'accord écrit de l'auteur ou de son ayant-droit, ou si la licence attachée à l'image permet explicitement sa réutilisation.

Droit *sui generis* des bases de données : en plus du droit d'auteur, les bases de données sont protégées *sui generis* jusqu'à 15 ans à compter de l'achèvement de la fabrication de la base de données. Le producteur d'une base de données bénéficie d'une protection du contenu de la base lorsque la constitution, la vérification ou la présentation de celui-ci atteste d'un investissement financier, matériel ou humain substantiel. Ainsi, concernant votre travail de scientifique, ***le droit de la propriété intellectuelle n'appréhende pas les données, mais les bases de données***.

En dehors du droit d'auteur, il existe aussi le droit de brevet, le droit des marques, la certification d'obtention végétale […] et le droit des contrats :

Clause de confidentialité : un contrat peut contenir une clause de confidentialité. Il est important de savoir que cette clause ne dépend que du contrat, et pas de la loi : seules les parties du contrat s'y engagent, pas les autres. *Attention, si une partie viole une clause de confidentialité en diffusant des données au-delà de votre droit, l'autre partie peut demander réparation. Et cette réparation peut s'élever très vite si des informations précieuses du nouveau prototype de votre entreprise sont récupérées par un tiers puis rediffusées avant que vous ne les retiriez…*

Données confidentielles : est confidentiel ce que les parties du contrat définissent comme confidentiel (sauf données personnelles et sensibles confidentielles par défaut). Afin d'éviter tout risque de diffusion non voulu (hacking par exemple), les parties du contrat doivent mettre en œuvre une politique de gestion des données sécurisées à la hauteur de la criticité du risque (criticité = probabilité x gravité). *Un échange de données personnelles par mail non sécurisé qui fuite pourra se retourner contre vous…*

Mise en garde : si vous faites partie d'un travail collaboratif, ou que vous avez signé un contrat contenant une clause de confidentialité, renseignez-vous auprès de votre hiérarchie pour connaître les modalités de partage. Vous n'avez peut-être pas le droit de partager votre travail !

2) Droit des données publiques

Les données publiques (par exemple, fournies par la Région, par l'Etat grâce aux documents administratifs) sont soumises à la politique d'Open Data depuis la loi du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public, dite loi Valter, et la loi pour une République numérique du 7 octobre 2016, portée par Axelle Lemaire. Cette loi promulgue la transparence des données et permet leur réutilisation :

> *« Lorsqu'un écrit scientifique issu d'une activité de recherche* ***financée au moins*** *pour moitié par des dotations de l'Etat, des collectivités territoriales ou des établissements publics, par des subventions d'agences de financement nationales ou par des fonds de l'Union européenne est* ***publié dans un périodique*** *paraissant au moins une fois par an, son auteur dispose, même après avoir accordé des droits exclusifs à un éditeur, du* ***droit de mettre à disposition*** *gratuitement dans un format ouvert, par voie numérique, sous réserve de l'accord des éventuels coauteurs, la version finale de* ***son manuscrit acceptée pour publication****, dès lors que l'éditeur met lui-même celle-ci gratuitement à disposition par voie numérique ou, à défaut, à l'expiration d'un délai courant à compter de la date de la première publication. Ce* ***délai*** *est au maximum de six mois pour une publication dans le domaine des sciences, de la technique et de la médecine et de douze mois dans celui des sciences humaines et sociales. »*

Cas du manuscrit : si vous êtes financé au moins pour moitié par de l'argent public (incluant les salaires), la publication de vos travaux dans une revue en accès ouvert ou sur une plateforme d'archive ouverte telle que HAL est obligatoire. Un délai d'embargo imposé par les éditeurs est possible, mais au maximum égal à 6 à 12 mois suivant votre discipline.

Cas des données & droit de réutilisation : les données doivent également être partagées publiquement (nous verrons dans les prochaines étapes comment) sauf par dérogation si vous indiquez une cause de confidentialité ou d'embargo temporaire. Les 3 raisons principales sont à cause de droits de propriété intellectuelle (partenariat public-privé…), de données à caractère personnel (RGPD, données de santé…) ou bien si la divulgation des données risque de mettre en péril l'objectif principal du projet. C'est donc au chercheur que revient la décision finale de ce qui doit être rendu public, dans le respect de la loi et des conventions de recherche ou accords de consortium établis en accord avec celle-ci.

3) Droit des données personnelles

Le guide d'analyse du cadre juridique en France sur l'ouverture des données de recherche diffusée par le ministère de l'enseignement supérieur, de la recherche et de l'innovation avec le soutien du comité pour la science ouverte a rédigé une fiche juridique sur les données personnelles en page 34-36. Voici quelques extraits :

Les données à caractère personnel :

*« Les données sont donc considérées « **à caractère personnel** » dès lors qu'elles concernent des personnes physiques :*

*- **identifiées directement** : lorsque par exemple son nom apparaît dans un fichier ;*

*- **identifiables indirectement** : lorsqu'un fichier comporte des informations telles que l'adresse I.P., le numéro d'immatriculation, le numéro de téléphone, un numéro d'identification lié à un fichier où se trouvent les données à caractère personnel (dans ce cas : données dites codées ou pseudonymisées), une photographie, des éléments biométriques, etc.*

Pour déterminer si une personne est identifiable via les données traitées, il faut donc analyser les risques en fonction du contexte et des moyens à disposition des utilisateurs leur permettant d'identifier cette personne. Par exemple, un croisement de données peut permettre une identification indirecte de la personne concernée : « Certaines données peuvent donc constituer des données à caractère personnel si elles permettent d'identifier indirectement ou par recoupement d'informations une personne précise. Il peut en effet s'agir d'informations qui ne sont pas associées au nom d'une personne mais qui permettent aisément de l'identifier et de connaître ses habitudes ou ses goûts. »

Exemple : une date de naissance associée à une commune de résidence »

Principes des données à caractère personnel :

« La loi « Informatique et libertés » définit les principes à respecter lors de la collecte, du traitement et de la conservation des données personnelles.

*- **Principe de finalité** : les données à caractère personnel ne peuvent être recueillies et traitées que pour un usage légitime et déterminé correspondant aux missions de l'établissement.*

*- **Principe de proportionnalité** : seules doivent être enregistrées les informations pertinentes et nécessaires à l'égard de la finalité déclarée.*

*- **Principe de durée limitée de conservation des données** : les informations ne peuvent être conservées de façon indéfinie dans les fichiers informatiques. Une durée de conservation doit être établie en fonction de la finalité de chaque fichier. Par la suite, les données doivent être supprimées ou archivées sur un support distinct (car elles ne doivent plus être utilisées). Il y a toutefois quelques exceptions, notamment pour les données conservées « en vue d'être traitées à des fins historiques, statistiques ou scientifiques ».*

*- **Principe de sécurité et de confidentialité** : le responsable du traitement est astreint à une obligation de sécurité. Il doit prendre les mesures nécessaires pour garantir la confidentialité, l'intégrité et la sécurité des données.*

*- **Principe de transparence** : le responsable du traitement doit informer les personnes des traitements auxquels leurs données sont soumises tout en leur accordant un droit d'accès, de modification, de rectification, voire d'opposition au traitement. Le responsable du traitement doit*

avertir ces personnes dès la collecte des données et en cas de transmission de ces données à des tiers. »

Ces principes protègent les individus mais entravent la réutilisation des données dans le cadre de l'open data. Il est donc important de réfléchir aux réutilisations éventuelles des données en amont de leur collection, et inversement de respecter ces principes lors de la réutilisation des données.

Données sensibles :

« Certaines données dites « sensibles » bénéficient d'un régime spécifique. Les données sensibles sont celles qui font apparaître, directement ou indirectement :

- les origines raciales ou ethniques ;

- les opinions politiques, philosophiques ou religieuses ;

- l'appartenance syndicale des personnes ;

- ***des informations relatives à la santé*** *ou à la vie sexuelle.*

Par principe, la collecte et le traitement de ces données sont interdits*. Cependant, dans la mesure où la finalité du traitement l'exige, ne sont pas soumis à cette interdiction :*

- les traitements pour lesquels la personne concernée a donné son consentement exprès ;

- les traitements justifiés par un intérêt public après autorisation de la CNIL ou adoption d'un décret en Conseil d'État ;

[…] Le traitement de certaines autres données à risque doit respecter un formalisme particulier : données génétiques, données relatives aux infractions pénales, aux condamnations, données comportant des appréciations sur les difficultés sociales des personnes, données biométriques, données comprenant le NIR, etc. »

Traitement informatisé des données personnelles :

« Tout fichier ou traitement informatisé comportant des données personnelles doit être déclaré au correspondant informatique et libertés (CIL) qui, selon le type de données ou de finalité du traitement, l'inscrit au registre des traitements de l'établissement ou instruit avec le responsable de traitement la demande d'autorisation auprès de la CNIL. »

Droit à l'image : le droit à l'image est une protection juridique sur les données personnelles incluant photos, vidéos mais aussi enregistrements vocaux. Ainsi, pour utiliser ces données, il est nécessaire d'avoir l'accord écrit de la personne majeure (diffusion, publication, reproduction ou commercialisation).

L'accord sera écrit et éclairé : quelle sera l'utilisation de mon image ? Sur quel support ? Pour quelle durée ? L'accord devra être redemandé pour une utilisation différente de celle précisée initialement.

Il n'y a pas besoin de recueillir le droit à l'image si c'est l'image concerne un groupe ou une scène de rue (si aucune personne n'est individualisée).

Dans le cadre d'une recherche à Euromov (hors données de santé) :

- Je délimite clairement les informations dont j'aurai besoin, et je ne collecte pas plus que cela.
- Je prévois une pseudoanonymisation suffisamment rigoureuse pour empêcher la possibilité d'identifier même indirectement mes sujets.
- Je soumets mon protocole à l'Institutional Review Board (IRB) d'Euromov qui validera ou non ma démarche.
- J'informe de manière éclairée les participants à mon étude sur la modalité de traitement de ces données.
- Je collecte les informations personnelles sur papier, je ne les informatise pas.
- Je prévois la destruction des informations personnelles non nécessaires à la fin du traitement de mes données.

Dans le cadre d'une recherche nécessitant des données de santé :

- Je dois obligatoirement passer par le comité de protection des personnes (CPP) qui m'éclairera sur la démarche à suivre.

4) En résumé

Le guide d'analyse du cadre juridique en France sur l'ouverture des données de recherche diffusée par le ministère de l'enseignement supérieur, de la recherche et de l'innovation avec le soutien du comité pour la science ouverte propose un logigramme de communicabilité des données en pages 28-29. Il est grandement recommandé d'aller le consulter à l'adresse suivante : lien 1, lien 2. De plus, tous les établissements publics ont un juriste compétent en matière de règlement RGPD. Il peut être utile de le consulter en amont de sa recherche.

5) Comment partager ses données (lorsqu'on est en droit de le faire) ?

Afin d'éviter tout problème juridique lié au droit d'auteur, il est impératif **d'associer une licence à vos données** (en l'ajoutant à vos métadonnées). Afin de favoriser leur réutilisation, la loi pour une République numérique précise que les données publiques Française doivent être associées à la licence ouverte Etalab (sauf dérogation).

D'autres types de licence existent, telles que les licences Creative Commons. Il existe plusieurs types de licences CC qui permettent d'ouvrir plus ou moins ses données. Par exemple, une licence CC BY SA est apposée à ce document (figure 1). En mettant cette image ici, je déclare que cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution – Partage dans les Mêmes Conditions 4.0

Figure 1. CC BY SA

International. Cela signifie que ce travail peut être repris et partagé par d'autres gratuitement même dans le cadre d'une utilisation commerciale (CC), tant que le travail original est cité (BY) et que les propos originaux sont utilisés dans le même objectif et dans les mêmes conditions (SA).

D'autres types de licences existent (figure 2). Pour apposer une licence CC sur un de ces documents, il suffit de mettre l'image correspondante dans une partie du document. Attention, pas de retour en arrière possible ! *Une licence très ouverte (CC) ne pourra pas être refermée plus tard puisqu'un tiers aura pu la recopier entre temps…*

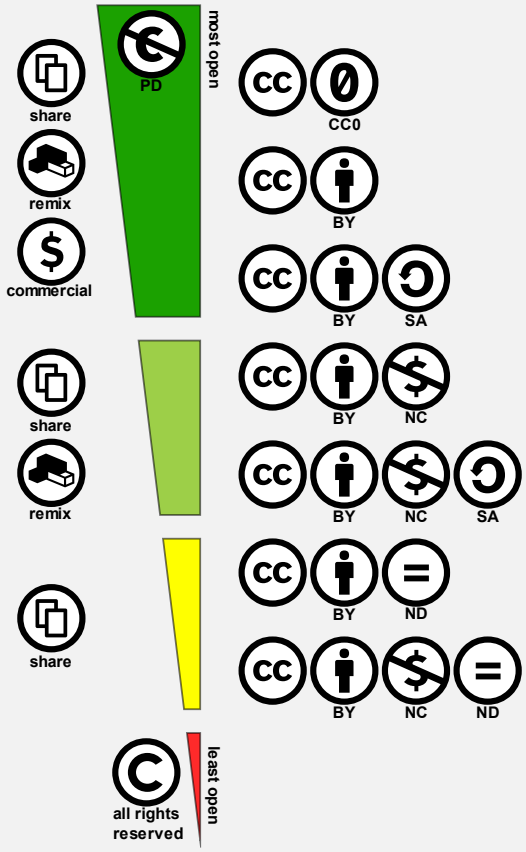

Figure 2. Licences Creative Commons.

Et pour le code, d'autres licences peuvent être plus adaptées. Il est recommandé d'utiliser une licence approuvée par l'Open Source Initiative. Le site choosealicense.com vous aide à choisir facilement la licence la plus adaptée à votre travail.

Pour aller plus loin

Définition des données de la recherche selon l'OCDE : lien

Plagiat en BD : lien

A guide to ethical writing: lien

Loi du 11 mars 1957 sur la propriété littéraire et artistique : lien

Le droit d'auteur dans nos classes : lien

Conférence/spectacle sur le droit d'auteur : lien

Aspects juridiques des droits d'auteurs : lien

Droit à l'image : lien

Loi Valter : lien

Loi Lemaire : lien

Plateforme d'archivage électronique ouverte (HAL) : lien

Ouverture des données de la recherche. Guide d'analyse du cadre juridique en France : lien

Droit de diffusion des données (résumé) : lien

Données obligatoirement diffusables (résumé) : lien

Bonnes pratiques juridiques et éthiques pour la diffusion des données de la recherche : lien

Licence ouverte Etalab : lien

Choisir une licence Creative Commons : lien

Open Source Initiative : lien

Choose a licence : lien

Etape 3 – Données, métadonnées, ID pérenne

Les principes européens ***GO FAIR*** (*FAIR Principles*) visent à lutter contre la crise de la reproductibilité évoquée plus tôt. En suivant les étapes décrites ci-après, vous faciliterez la diffusion et réutilisation des données par un processus de FAIRification de votre recherche (figure 3). Le principe général est le suivant : ***As open as possible, as closed as necessary.*** En pratique, les données doivent être :

- **Findable**
  - **F1.** (Meta)data are assigned a globally unique and persistent identifier (PID)
  - **F2.** Data are described with rich metadata (defined by R1 below)
  - **F3.** Metadata clearly and explicitly include the identifier of the data they describe
  - **F4.** (Meta)data are registered or indexed in a searchable resource
- **Accessible**: (Meta)data are retrievable by their identifier using a standardised communications protocol
  - **A1.1** The protocol is open, free, and universally implementable
  - **A1.2** The protocol allows for an authentication and authorisation procedure, where necessary
  - **A2.** Metadata are accessible, even when the data are no longer available
- **Interoperable**
  - **I1.** (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - **I2.** (Meta)data use vocabularies that follow FAIR principles
  - **I3.** (Meta)data include qualified references to other (meta)data
- **Reusable:** Meta(data) are richly described with a plurality of accurate and relevant atributes
  - **R1.1.** (Meta)data are released with a clear and accessible data usage license
  - **R1.2.** (Meta)data are associated with detailed provenance
  - **R1.3.** (Meta)data meet domain-relevant community standards

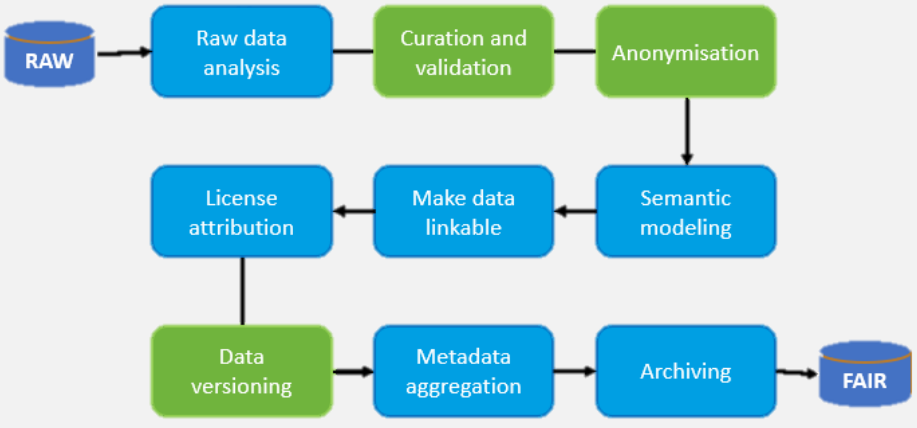

Figure 3. Le processus de FAIRification.

1) FINDABLE : Mes données doivent être trouvées facilement

Le premier príncipe FAIR, rappelé ci-dessous renvoi à 2 notions : "F1. (**Meta)data** are assigned a globally unique and **persistent identifier (PID)**".

Persistent identifier (PID) **:** C'est une référence durable à un document, un dossier, une page web ou un autre objet. *Vous avez déjà tous été confronté au problème "Erreur HTTP 404 : aucune ressource n'a été trouvée à l'adresse demandée". Cette erreur peut subvenir lorsque la page web a été supprimée, mais aussi lorsqu'elle a simplement changé d'adresse !* Pour éviter ce problème, il est possible d'assigner un identifiant pérenne à notre donnée. Par exemple les DOI ou les PMID s'appliquent aux articles.

En règle générale, un tel identifiant est non seulement pérenne, mais il peut aussi faire l'objet d'une action : vous pouvez l'utiliser dans un navigateur web et être dirigé vers la source identifiée (on parle d'URI).

Il existe un identifiant pérenne pour les chercheurs : **ORCID** (Open Researcher and Contributor ID). C'est un code alphanumérique non propriétaire pour identifier de façon unique un auteur/contributeur scientifique ou académique. ***En précisant cet identifiant sur tous vos articles et communications***, vous permettez :

- Aux lecteurs d'identifier sans conflit l'auteur (pas de problème d'homonyme).
- Aux lecteurs de vous contacter si besoin (pas de problème de changement d'adresse mail etc. : un seul compte ORCID tout au long de votre carrière).
- Aux institutions de repérer facilement vos contributions : lorsque votre ORCID est indiqué dans un papier, celui-ci est automatiquement ajouté à votre compte ORCID en quelques mois.

Aller sur le site : https://orcid.org/ puis créer un compte. Veillez à choisir un nom facilement identifiable, par exemple avec au moins 2 à 3 prénoms. Exemple : Germain Valentin FAITY (n'ayant pas de 2e prénom, avec son accord, j'ai ajouté en 2nd celui de mon frère qui ne travaille pas dans le même domaine que moi). N'oubliez pas votre mot de passe !

2) ACCESSIBLE : L'accessibilité des données doit être définie

Metadata (métadonnées) : ce sont des données qui fournissent des informations sur d'autres données. Couramment, les métadonnées décrivent une ressource afin d'en faciliter l'identification et l'utilisation. Par exemple, l'objet d'un mail, l'heure de réception, le destinataire, l'expéditeur sont des métadonnées descriptives d'un mail. Pour de la capture du mouvement, on notera la date et l'heure, le nom de l'expérience, le code du sujet, le numéro de l'essai, l'outil d'enregistrement, la fréquence d'échantillonnage, la tâche effectuée…

Les jeux de données doivent être dotés de métadonnées cohérentes. Il existe des standards établissant des métadonnées spécifiques à un type de données. Ces standards sont souvent spécifiques à un champ et peuvent être trouvés sur les sites suivant :

- http://www.dcc.ac.uk/resources/metadata-standards
- http://rd-alliance.github.io/metadata-directory/standards/

Le standard le plus simple et pouvant s'appliquer à la plupart des données est le **Dublin Core** : https://www.dublincore.org/specifications/dublin-core/dces/. Ce standard de métadonnées indique le contributeur, la couverture, le créateur, la date, la description, le format, l'identifiant, le langage, l'éditeur, les droits, la source, le sujet, le titre et le type de la ressource.

Un autre standard plus large et plus détaillé (**ISA-TAB 1.0**) permet de capturer et communiquer des métadonnées complexes nécessaires à l'interprétation des expériences utilisant des combinaisons de technologies et des fichiers de données associés : http://isatab.sourceforge.net/docs/ISA-TAB_release-candidate-1_v1.0_24nov08.pdf.

Et le nom de fichier ?

La métadonnée par excellence est le nom du fichier. Voici 5 règles émises par la plateforme de science ouverte DoRANum pour bien nommer ses fichiers :

- Donner un nom bref et explicite. *Supprimer ce qui n'est pas essentiel, CR pour compte-rendu…*
- Ne pas mettre d'espace ni de caractères spéciaux. *Séparer par majuscule ou underscore.*
- Indiquer les dates au bon format. *AAAAMMJJ.*
- Placer l'élément important en premier.
- Indiquer les versions des documents. *V1 pour version 1, VF pour version finale…*

Enfin, il est fortement recommandé de rédiger un document explicatif de la méthode utilisée pour renommer vos fichiers, surtout si vous utilisez des règles de nommages un peu complexes.

Metadata : Pour chacun de vos jeux de données, trouver le standard le plus adapté et créer des métadonnées cohérentes afférentes à ses données. Dans le cas où les données couvrent plusieurs champs différents, il est possible d'additionner les informations provenant de plusieurs standards, en précisant la logique utilisée.
Pensez à nommer vos fichiers en accord avec les 5 règles données par DoRANum !

3) INTEROPERABLE : Mes données doivent être (ré)utilisables facilement

Est-il possible de lire un disque 78 tours à l'aide d'un lecteur USB ? A moins qu'il ne soit doté d'un gramophone ou d'un tourne-disque, le format de cette donnée (l'enregistrement audio) n'est pas adapté à l'appareil de lecture. De même, comment exécuter un fichier matlab lorsque l'on n'a pas de licence matlab ?

L'évolution rapide des technologies oblige à mettre à jour régulièrement les formats des données et à privilégier des formats pérennes. Le Centre Informatique National de l'Enseignement Supérieur (CINES) met à disposition une liste des formats recommandés à cette adresse : https://facile.cines.fr/

Heureusement, si notre format préféré n'est pas dans cette liste, cela ne veut pas dire que l'on sera forcément dans l'incapacité de lire notre fichier dans 5 ans ou dans un autre laboratoire. Il sera néanmoins plus difficile de le faire…

Formats standardisés : il existe des cas plus complexes : la documentation d'un format peut devenir une norme officielle nationale ou internationale ou un standard de facto.

- PDF/A1 est une version standardisée (ISO 19005) du format PDF. Les autres versions de PDF ne sont pas standardisées.
- Les formats Libre office (ODS, ODT…) sont standardisés (ISO/IEC 26300).
- Le format XML est standardisé par une « recommandation » du W3C (équivaut à une norme).
- Le format CSV est décrit dans la RFC 4180 de l'IETF, mais n'est pas réellement standardisé (la RFC est un document indicatif), plusieurs versions existent.
- Les formats bureautique Microsoft (XLSX, DOCX…) sont standardisés (ISO/IEC 29500). Mais les logiciels semblent parfois s'écarter du standard.
- Bien que scilab soit open source, matlab est pour l'instant considéré comme le standard…

Ci-dessous quelques exemples :

	Format ouvert	**Format fermé**
	Spécifications publiques et gratuites	Spécifications non publiques
	Aucune restriction légale pour l'utiliser	Des restrictions légales s'opposent à son utilisation (droit d'auteur, copyright, brevet)
	Format indépendant du logiciel utilisé qui assure l'interopérabilité des données	Format lisible qu'avec un logiciel particulier
	Maintenu par une organisation à but non lucratif	Format propriétaire
Doc Texte	PDF, TXT, ODT	MS Word, RTF
Feuille de calcul	ODS, CSV	MS Excel, PDF, OOXML
Base de données	SQL, SIARD, DB tables (.CSV)	MS Access, dBase (.dbf), HDF5
Images	JPEG, TIFF, PNG	DICOM
Audio	BWF, MXF, Matroska (.mka), FLAC, OPUS	WAVE, MP3, AAC, AIFF, OGG
Vidéo	MXF, MKV	MPEG-4, MPEG-2, AVI, QuickTime (.mov, .qt)

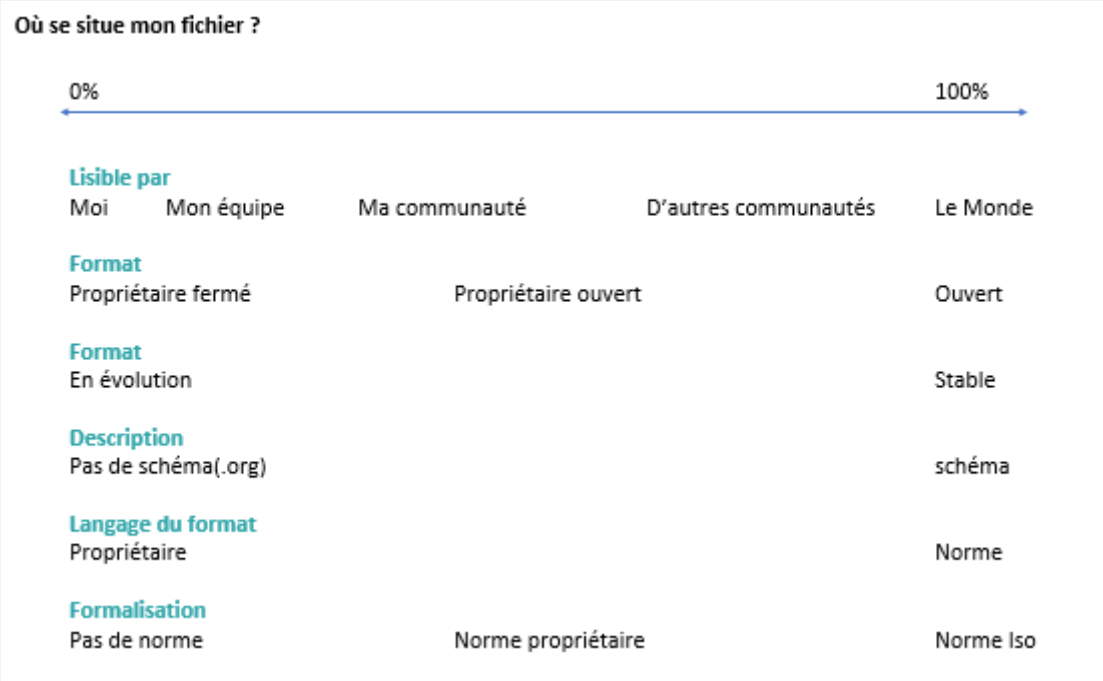

Figure 4. Essayons de nous améliorer...

Format Ouvert ou Fermé : pour chaque type de format de données, réfléchissez s'il est plutôt ouvert ou fermé (voir figure 4). Si le format est fermé, comment pouvez-vous le rendre plus interopérable ?

FAIR : pour chacun de vos jeux de données, mesurez sa conformité aux principes FAIR, par exemple en utilisant un tableur. Comme vous pouvez le constater, il est bien plus simple de rendre ses données FAIR en pensant le processus de FAIRification en amont de la création des données ! Vous le saurez pour la prochaine fois...

Pour aller plus loin

Acteurs :

- *GO FAIR : lien*
- *Science Europe : lien*
- *Research Data Alliance : lien*
- *European Open Science Cloud : lien*
- *Open AIRE : lien*
- *FAIRSFAIR : lien*

Identifiant pérenne : lien

Métadonnées, standards, formats : lien

Renommer un fichier : lien

Stockage-archivage : lien

Format Ouvert ou fermé : lien

Dot What, une base de données géante sur les formats de fichier : lien

Etape 4 – Stockage des données pendant le projet

Que se passe-t-il si votre ordinateur plante ? Si un cambrioleur vole votre ordinateur et vos 2 disques durs ? Ces histoires vraies sont malheureusement bien trop récurrentes (comme on peut voir ici, là et là…). Nous allons voir comment mettre en place une stratégie de sauvegarde efficace de vos données.

Quelle que soit la nature de vos données, le support de stockage optimal dépend de plusieurs critères :

- Fréquence d'utilisation des données / Vitesse d'accès à la donnée
- Besoins en capacité de stockage (taille) / Coût du support
- Fiabilité du support / Sécurité des données

Pour les données numériques, *DoRANum* précise les avantages et inconvénients des supports commun :

- **Ordinateur personnel** : permet un stockage temporaire facile et pratique mais est peu sécurisé (piratage, pannes, vol…). Possibilité d'améliorer la sécurité en cryptant les données sensibles.
- **Support externe** (clé USB, disque dur externe…) : permet un stockage temporaire très peu sécurisé (piratage, pannes, vol…) mais pratique et rapide pour transférer ses données vers un autre ordinateur. Il est également possible de crypter ses données.
- **Serveur institutionnel** : permet un stockage plus pérenne (fiable, durable et sécurisé). Le coût est important mais souvent déjà pris en charge par l'institution.
- **Serveur cloud** : sécurité des données fragiles mais permet le travail collaboratif facilement. On évitera d'y stocker nos données confidentielles. Cette solution est payante à partir d'une certaine limite de stockage.

Le stockage sécurisé est important pour assurer la continuité de l'exploitation des données sur du court terme. Afin de ne pas perdre ses données, il existe une règle simple de la sauvegarde : 3-2-1-0.

- **Avoir au moins 3 copies de ses données.**
- **Stocker les copies sur au moins 2 supports différents.**
- **Avoir au moins 1 sauvegarde hors site.**

⇨ Zéro inquiétude pour la sauvegarde de ses données.

Si vous êtes affiliés au laboratoire Euromov, vous disposez :

- D'un espace de travail à sécurité renforcée, accessible en- et hors- ligne : « Seafile Client ».
- D'un serveur de stockage à grande capacité accessible via cable ethernet : « Backup ».

Rapprochez-vous du support d'euromov pour la première connexion.

Comment automatiser vos sauvegardes ? #Robocopy

Afin de gagner du temps et d'éviter les erreurs humaines lors des sauvegardes, il est possible d'automatiser vos sauvegardes. Une solution efficace sur Windows est l'outil BATCH avec la commande Robocopy.

Automatiser ses sauvegardes sur Windows avec l'outil Robocopy

1. Voir cette vidéo pour mieux comprendre : https://www.youtube.com/watch?v=Xbo65eNq6fs
2. Désigner un ou plusieurs dossiers **sources** qui seront copiés.
3. Désigner un ou plusieurs dossiers **destinations** qui accueilleront les fichiers du dossier source**.**
4. Ecrire un script du type :

```
C:\Windows\System32\robocopy "C:\Users\germain.faity\Source " "Z:\ Destination "
```

et l'enregistrer dans un fichier .txt

5. Améliorer votre script en ajoutant des options en fin de ligne. Voir ici. Par exemple :

```
C:\Windows\System32\robocopy "C:\Users\germain.faity\Source " "Z:\Destination " /E /A-:SH /V /R:1 /W:0 /XJ /MT:32 /PURGE
```

6. Modifier l'extension du nom de fichier .txt par .bat (BATCH). Ce fichier (le bout de code) est maintenant un exécutable !
7. Double cliquez sur ce fichier .bat pour exécuter la copie de vos fichiers. Une fenêtre de type console doit s'ouvrir. Si ce n'est pas le cas, vérifiez votre code. Une fois terminé, vérifiez que la sauvegarde a été effective.
8. Si vous le souhaitez, vous pouvez programmer votre système à exécuter ce fichier à des dates récurrentes grâce au planificateur de tâche de Windows.

11 étapes pour une sauvegarde réussie

1. Vérifiez si votre institution dispose d'une stratégie de sauvegarde.
2. Déterminez ce que vous voulez sauvegarder.
3. Décidez du nombre et de la fréquence des sauvegardes.
4. Décidez où les sauvegardes seront stockées.
5. Déterminer la capacité de stockage nécessaire.
6. Déterminez s'il existe des outils que vous pourriez utiliser pour automatiser la sauvegarde.
7. Déterminez combien de temps les sauvegardes seront conservées et comment elles seront détruites.
8. Déterminer comment les données personnelles seront protégées.
9. Elaborer un plan de reprise après sinistre.
10. Attribuer des responsabilités.
11. Déterminer comment vérifier l'intégrité des fichiers sauvegardés.

Pour aller plus loin

Avantages et inconvénients des supports : lien

Automatiser ses sauvegardes, l'outil robocopy : lien

Tuto robocopy : lien

Etape 5 – Accès et partage des données pendant le projet

A l'ère numérique, il existe une multitude d'outils permettant le partage de nos données et la collaboration inter- équipes. Ces outils s'apparentent à des cloud (stockage de fichier en ligne) plus ou moins structurés. Les plus connues sont les outils privés tels que Google Drive, Microsoft OneDrive ou Dropbox. Les mails sont également très utilisés… Bien que ces solutions soient « pratiques », on ne sait jamais vraiment où et comment sont stockées nos informations. Attention s'il s'agit d'informations sensibles (RGPD, données de santé…) !

Des alternatives publiques existent, même si elles sont moins connues :

- PARTAGE par RENATER : environnement d'outils collaboratifs.
- AlfrescoShare : gestion électronique de documents.
- OSFHome : gestion collaborative de documents de l'Open Science Fundation (connexion via ORCID).
- European Open Science Cloud (EOSC) : fédération de solutions de cloud européens.

Enfin, il existe des solutions mises en place par vos institutions.

Renseignez-vous auprès de votre institution pour savoir si vous bénéficiez d'un accès à serveur collaboratif. Si vous êtes affiliés au laboratoire Euromov, le serveur « Seafile Client » permet le partage de données très facilement.

Et pour le code ?

Diverses solutions sont également disponibles pour enregistrer, gérer et partager son code. Ces services utilisent le logiciel Git, qui permet la gestion de versions de façon décentralisée. Il existe plusieurs clients permettant d'utiliser Git, qui ont chacun leurs spécificités. Ici, les 3 principaux :

- Publique : Git par RENATER
- Open Source : GitLab
- Privé : GitHub

Il sera possible d'utiliser Git en interface bureau avec un Git Desktop (par exemple : GitHub Desktop).

Identifier des pratiques douteuses de partage de fichiers que vous utilisez à ce jour (par exemple, envoi de données sensibles par mail, ou stockage de données confidentielles sur google drive).
Mettre en place une solution plus sécurisée de partage de fichiers.
Si vous codez : créer un compte GitLab et installer un Git Desktop. Enregistrer son premier projet et le partager avec ses collaborateurs.

Etape 6 – Diffuser et partager les données de recherche

En première partie, nous avons vu l'intérêt de diffuser ses données pour améliorer la reproductibilité de la science. ***Alors, quelles sont vos habitudes ?***

Les grands journaux (Nature…) commencent à demander des matériels supplémentaires y compris les jeux de données utilisés. D'après une étude publiée dans Nature (Cousijn *et al.*, 2018), le schéma idéal lors de la soumission de son manuscrit est d'établir un lien entre les données de sa recherche :

- **L'article manuscrit** (avec lien persistant : DOI) renvoie à une landing page grâce à un PID.
- La **landing page** (avec lien persistant : PID) renvoie à notre data set.
    - Contient des Human Readable Metadata
    - Contient des Machine Readable Metadata
- Le **data set** (avec lien persistant : PID) hébergé dans un entrepôt de données contient :
    - Données brutes et/ou pré-traitées
    - Code utilisé
    - Matériel supplémentaire si nécessaire

La landing page avec OSF : The OSF (Open Science Framework) est un projet logiciel open source qui facilite une collaboration ouverte dans la recherche scientifique. Il permet de collaborer, documenter, archiver, partager et enregistrer des projets, des matériels et données. *Par exemple, il est possible d'indiquer un identifiant OSF dans un article afin de renvoyer à des matériels supplémentaires. Un autre exemple est d'enregistrer votre méthode à une date qui fait foi en amont de sa mise en place dans votre étude, ce qui permettra d'augmenter la reproductibilité de votre recherche (registered report ou preregistration).*

> Connectez-vous sur : https://osf.io/ en utilisant votre compte ORCID. Créer un projet et remplissez le, soit en entier si votre manuscrit est prêt, soit en partie si votre projet est en cours. Vous y ajouterez un lien vers vos données.

Les entrepôts de données (repository) : si vos données dépassent la capacité de votre landing page, il est possible de les héberger dans un entrepôt de données. Un entrepôt est défini comme "a searchable and queryable interfacing entity that is able to store, manage, maintain and curate data" (Jonhston, 2017). Les entrepôts peuvent être disciplinaires (chercher sa discipline dans re3data.org ou voir ceux que Nature recommande), institutionnels ou tout venant.

Un fois les données stockées dans un entrepôt, il est possible de les chercher et potentiellement de les réutiliser grâce aux sites DataCite ou Google DataSet Search. Certains articles (data paper) servent uniquement à décrire un jeu de données en vue de leur ré-utilisation. Pour ce type d'articles, il est possible de trouver des modèles qui faciliteront l'écriture du manuscrit.

Et si je veux ré-utiliser des données en open-data ? Contactez l'auteur grâce à son ORCID ou la revue dans laquelle le data paper est publié pour vérifier la licence d'exploitation et les problèmes de droits d'auteurs et de confidentialités afférents, puis lancez-vous !

Choisissez la licence CC la plus adaptée pour chacun de vos jeux de données (voir Etape 2).
Chercher un entrepôt de données qui pourrait accueillir vos données (voir aller plus loin). Vérifiez si l'entrepôt est gratuit, et si ce n'est pas le cas, renseignez-vous pour savoir si les frais peuvent être pris en charge par votre institution comme frais de publication.

Une méthode plus robuste avec les registered reports

Le centre pour la science ouverte (Center for Open Science, https://www.cos.io/our-services/registered-reports) décrit le registered report comme le processus de revue par les pairs avant l'étape de collection des données. Il s'agit de soumettre un article « protocole » à un journal, contenant les motivations de l'expérience, les hypothèses testées et la méthode qui sera utilisée. En soumettant cette partie, vous garantissez une méthode robuste (peer-reviewed) en amont de votre expérimentation et augmentez ainsi la qualité de votre étude. Vous soumettez ensuite dans une deuxième phase les résultats de votre étude. L'article est automatiquement validé par les reviewers s'il respecte les méthodes présentées en phase 1 et ce, quels que soient les résultats.

Daniel Simon, co-éditeur de la partie Registered Replication Reports de la revue Psychological Science, ajoute que les registered repports éliminent le biais de résultats négatifs lors de la publication, puisque les résultats ne sont pas encore connus lors de la revue par les pairs.

Les journaux acceptant les registered repport sont disponibles dans l'onglet « Participating Journals » de la page du centre pour la science ouverte.

Améliorer la qualité de sa recherche avec les preprints

Un preprint est la version d'un article pas encore soumis à un éditeur mais mis à disposition du public. L'intérêt des preprints est multiple. Nous l'avons vu avec la crise du COVID-19, les preprints permettent de mettre à disposition rapidement le travail d'une équipe, ce qui permet d'accélérer la transmission et la visibilité de l'information scientifique, surtout dans les domaines qui évoluent rapidement. La visibilité est également améliorée si l'on en croît le nombre de citation plus élevé pour les articles ayant fait l'objet d'un preprint (Serghiou & Ioannidis, 2018). Par ailleurs, les feedbacks reçus sur les sites de preprints permettent d'améliorer le manuscrit prochainement soumis, et ainsi d'augmenter la qualité de sa recherche. Enfin, la date de sortie du preprint fait foi quant au crédit des auteurs. Cette pratique permet ainsi de diminuer les conflits de découverte simultanée. Il est toutefois nécessaire de faire attention à la qualité du travail des preprints puisque ces articles ne sont pas encore revus par les pairs (Bourne *et al.*, 2017).

Il existe maintenant des serveurs de preprints pour la plupart des domaines de recherche, incluant les sciences physiques, les sciences sociales, la biologie, la médecine, la science des activités physiques et sportives et d'autres domaines.

La publication d'un preprint ne constitue pas une violation de la condition de non duplication des publications demandée par la plupart des éditeurs. Lorsque l'article est accepté, les éditeurs peuvent demander à le mentionner sur le preprint, en indiquant le journal, la date de publication et le lien vers l'article publié.

Augmenter la visibilité de son travail avec l'Open Access

Publier en libre accès signifie rendre l'accès à votre article gratuit. Il existe plusieurs façons de procéder :

- La **voie dorée** (« Gold Open Access ») est le principe de publier directement dans un journal gratuit pour les lecteurs (les auteurs payent souvent les frais de publications). Une liste non exhaustive est disponible sur https://doaj.org/
- La **voie verte** (« Green Open Access ») renvoie à l'autoarchivage de l'article. Il s'agit de mettre à disposition librement son article (version postprint, non mise en page par l'éditeur) sur des plateformes d'archive ouverte (telle que HAL) une fois le délai d'embargo imposé par l'éditeur levé (maximum 6 à 12 mois selon la discipline). Les abonnés de la revue d'origine payent donc la publication initiale, mais son archivage sur les plateformes ouvertes sont aux frais de la plateforme (la plupart des établissements scientifiques français ont un entrepôt de ce type).
- La **voie hybride** (« Hybrid Open Access ») renvoie aux journaux conventionnels qui incorporent des articles en accès ouvert. Lorsque l'on souhaite publier en accès ouvert dans un journal hybride, il faut le préciser et régler des frais de publication (*e.g.* frais des revues du groupe Springer Nature ici). L'article sera accessible librement sur le site internet de la revue. Attention, selon ce modèle, l'éditeur est payé deux fois (par l'auteur et par les abonnés). Il n'est pas éligible auprès des organismes financeurs publics (ANR par exemple).

En facilitant l'accès à votre article, vous augmentez par la même occasion votre visibilité et donc le nombre de citations (Eysenbach, 2006).

Pour aller plus loin

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., Murphy, F., Polischuk, P., Taylor, S., Martone, M., & Clark, T. (2018). A data citation roadmap for scientific publishers. Scientific Data, 5(1), 1-11. https://doi.org/10.1038/sdata.2018.259

Bonnes pratiques pour faciliter l'accès aux données de la recherche : lien

Curating research data : lien

Chercher des données

- *DataCite :lien*
- *DataSet Search : lien*

Entrepôts disciplinaires

- *Chercher son entrepôt disciplinaire sur re3data.org :lien*
- *Sciences biomédicales : lien*
- *Exemple 1. Collaborative research in computational neuroscience : lien*
- *Exemple 2. Motion capture database HDM05 : lien*

Entrepôt tout venant. Dataverse : lien. Attention, en mode démo, les données sont supprimées au bout de 24h, et ça ne donne pas un véritable PID !

Lisa R. Jonhston. (2017). Curating Research Data : Volume One : Practical Strategies for Your Digital Repository. Association of College Ad Research Librairies, 294.

Data reusing : lien 1 ; lien 2

Liste de revues publiant des data papers : lien 1 ; lien 2

Registered reports sur cos.io : lien

Preprints

- *Sciences physiques : lien*
- *Sciences sociales : lien*
- *Biologie : lien*
- *Médecine : lien*
- *Science des activités physiques et sportives : lien*
- *Autres serveurs soutenus par l'open science framework : lien*

Liste non exhaustive des revues en gold open access : lien

Springer Nature open-access publication fee : lien

Etape 7 – L'archivage à long terme

Est-il facile aujourd'hui de lire une disquette ou une VHS ? Qu'en sera-t-il dans 100 ans ? Ces outils pourtant très répandus il y a une vingtaine d'années sont difficilement lisibles maintenant. Maintenant réfléchissez à vos données. ***Sera-t-il facile, voire possible de lire vos données dans 100 ans ?***

Bien que nous soyons capables de lire des textes politiques, philosophiques, scientifiques d'Egypte antique vieux de plus de 4 000 ans, nous sommes en difficulté pour assurer une pérennité d'une centaine d'années à nos données numériques. Heureusement, il existe des solutions ! En effet, l'archivage électronique vise à conserver la lisibilité de nos données numériques malgré l'obsolescence des anciens supports.

En France, un seul acteur œuvre pour l'archivage électronique à long terme : le CINES (Centre Informatique National de l'Enseignement Supérieur). Ce centre conserve vos données en assurant leur pérennité. Comme l'indique le CINES :

> *« L'archivage pérenne à 3 objectifs principaux :*
>
> - *Conserver le document : conserver l'intégrité du document sur son support de stockage.*
> - *Le rendre accessible : pouvoir retrouver le document et le lire.*
> - *En préserver l'intelligibilité : faire en sorte que le document reste compréhensible par ses utilisateurs potentiel à travers le temps.*
>
> *Il existe 4 principaux risques qui menacent inéluctablement un fichier : l'obsolescence matérielle, l'obsolescence logicielle, l'obsolescence du format de fichier, la perte de la signification du contenu. »*

Le CINES met en œuvre plusieurs types de solutions pour éviter ces risques en mettant à jour régulièrement les supports, en choisissant des formats durables voir en assurant la conversion de format lorsqu'un format durable devient obsolète. Enfin, pour préserver l'intelligibilité, la seule solution reste des métadonnées riches et complètes. *Comment aurait-on fait pour déchiffrer les hiéroglyphes sans la pierre de Rosette ?*

Vous l'aurez compris, cet archivage à long terme n'est pas une simple copie sur un serveur distant, mais bien une capsule temporelle de votre travail. Cet archivage a un coût, il est donc nécessaire de bien choisir quelle sont les données vraiment importantes ? Brutes ou lissées ? Quel format de métadonnées utiliser ?

Exemple simple pour passer d'un tableau « à risque » à un tableau assez pérenne : l'enregistrer en .csv sur cloud avec accès partagé et description des données disponibles au lieu de l'enregistrer en .xls sur disque dur externe sans métadonnées et avec un titre incompréhensible du type « Nouveau Feuille de calcul Microsoft Excel.xlsx ».

Pour chacun de vos jeux de données :

1. Choisissez quelles données seront à archiver (data curation).
2. Choisissez quel est le format le plus adapté pour ces données. A vous de choisir un compromis entre confort et efficacité !
3. Sous quelle forme allez-vous stocker les métadonnées afférentes ? Dans le même fichier ou dans un fichier à part ?
4. Quelle plateforme d'archivage sera la plus adaptée ?
5. Archivez ces données !

Pour aller plus loin

Court, moyen, long-terme : lien

Le concept d'archivage numérique pérenne par le CINES : lien

Etape finale – Le Plan de Gestion des Données (PGD)

Nous sommes bientôt arrivés à la fin ! Cette dernière étape vise à synthétiser les étapes précédentes dans un document de synthèse.

Poser vous la question : si pour une raison X ou Y vous ne continuez plus votre pratique actuelle de recherche (du jour au lendemain), qu'advient-il de votre travail ? Quelqu'un peut-il FACILEMENT reprendre vos travaux là où vous les aviez laissés ? Ou devra-t-il tout recommencer ?

Pour que la science soit cumulative, il est nécessaire de mettre en place une gestion correcte des données de leur acquisition jusqu'à leur diffusion et archivage à long terme. Nous avons vu qu'il existe des risques afférents à chaque étape de la vie des données (voir vidéo étape 1). Pour les minimiser, il est important de les prévoir, les documenter et proposer une solution permettant de les éviter.

Le **Plan de gestion des données (PGD)** ou Data Management Plan (DMP) est une approche pragmatique simple à comprendre, à mettre en place, à évaluer et à faire évoluer. Ce document initié en début de projet (avant l'acquisition des données) sera mis à jour périodiquement afin de gérer au mieux les données utilisées et générées dans le cadre de son projet de recherche.

Le *Digital Curation Centre* précise le contenu d'un plan de gestion :

- Description des données
- Standards / méthodologies utilisés pour la collecte et la gestion des données
- Préoccupations ou restrictions éthiques et de propriété intellectuelle
- Plan pour le partage des données et l'accès
- Stratégie de préservation à long-terme

Afin de rendre la science plus efficace, plus intègre et minimiser les risques de perte de données, de temps et d'argent, les financeurs souhaitent de plus en plus voir un PGD accompagner une réponse à un appel d'offre… Ce qui fait une raison supplémentaire pour prendre en main cet outil !

1. **Créez votre PGD**

Des modèles (templates) de PGD sont présent au site suivant : https://dmp.opidor.fr/public_templates
Nous vous conseillons de regarder les modèles « INRAE – General Project Template » (modèle plutôt complet, à remplir) et « Horizon 2020 FAIR DMP » (modèle plutôt simple, à étayer vous-même).

Une fois le template le plus adapté choisi (ou créé), vous devez le remplir consciencieusement avec les informations recueillies jusque-là, pour chaque ensemble de données (en général, 1 expérience = 1 ensemble de donnée). Si vous êtes passé par dmp.opidor.fr, il est possible de remplir en ligne votre PGD et de le télécharger une fois rempli !

2. **Partagez votre PGD avec vos collaborateurs**

Ce document synthétise votre stratégie de gestions de vos données, il est donc important qu'il soit connu de vos collaborateurs en cas de besoin.

Pour aller plus loin

Digital Curation Center : lien
Le cycle des métadonnées : lien
Plan de gestion des données : lien
Templates de PGD : lien

Bilan – Comment je me situe par rapport à l'Open Science ?

Recopiez le radar ci-dessous sur une feuille. Faites une double évaluation : notez-vous de 0 à 5 sur chaque critère selon votre situation d'avant ce guide, et votre situation maintenant. Admirez le résultat sur le graphique radar. Êtes-vous devenu plus OPEN ?

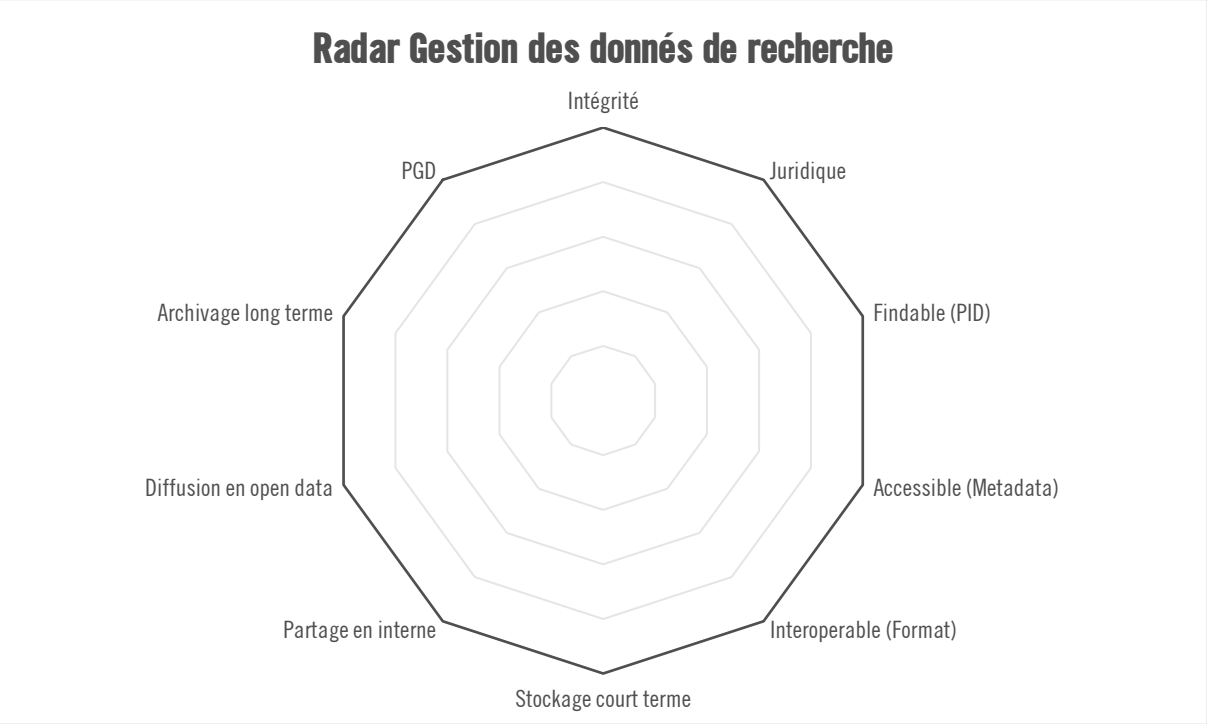

Figure 5. Radar Open Science