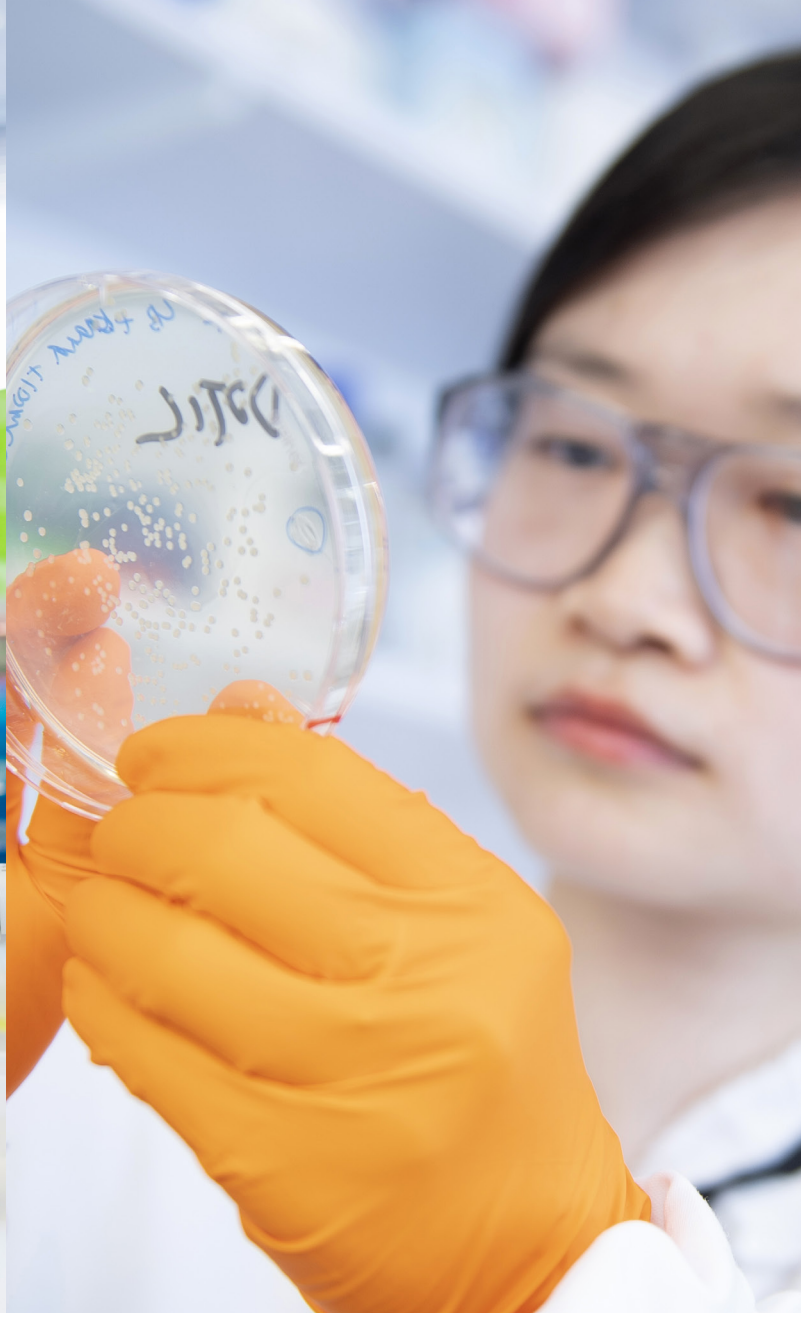
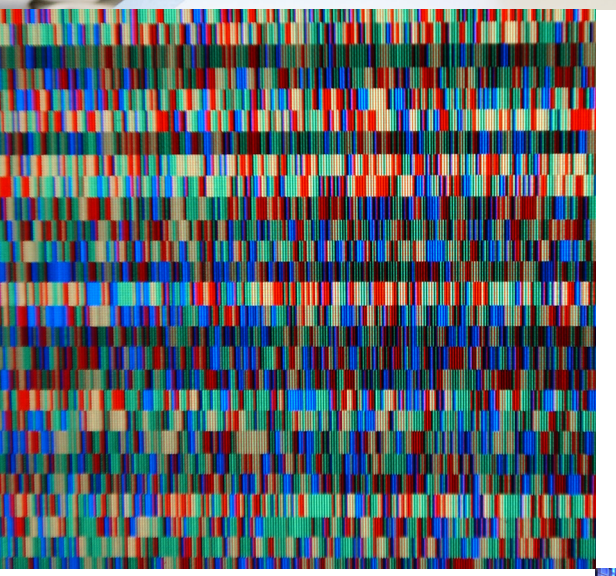


© Institut Pasteur, T. Lang, F. Gardy / AdobeStock



Politique de gestion et partage des données de la recherche et codes logiciels

Mai 2021

INSTITUT PASTEUR

Table des matières

1. Contexte 2
2. Domaine d'application 3
3. Objectifs de la politique 3
4. Rôle, droits et responsabilités du scientifique 4
4.1. Planifier, anticiper 4
4.2. Réutiliser, créer des données 4
4.3. Utiliser, traiter, stocker 5
4.4. Diffuser, publier, partager 5
4.5. Préserver, archiver 7
4.6. Gérer des logiciels, scripts, programmes 7
5. Rôle, droits et responsabilités de l'institution 8
5.1. Mettre à disposition une infrastructure 8
5.2. Accompagner et former 8
5.3. Reconnaître et valoriser les bonnes pratiques 9
6. Glossaire 10

Résumé

Cette politique fixe les **lignes directrices** de l'Institut Pasteur sur la **gestion et le partage des données de la recherche et des codes logiciels**. Elle résume les bonnes pratiques que l'Institut Pasteur demande ou recommande de mettre en œuvre tout au long du processus de recherche.
L'Institut Pasteur accompagne les scientifiques dans l'application de la présente politique. Cette dernière renvoie vers des **fiches pratiques** (liste complète en annexe) afin de donner aux scientifiques les moyens opérationnels de mettre en œuvre les bonnes pratiques.

1. Contexte

La recherche internationale se mobilise pour assurer la préservation, le partage et la réutilisation des produits de la recherche scientifique. Ce mouvement global mobilisant chercheurs, politiques et financeurs a pour objectif d'améliorer la qualité, l'intégrité et la reproductibilité de la recherche.

A l'Institut Pasteur, l'écosystème de la recherche (entités de recherche, plateformes, centres nationaux de référence…) utilise, manipule et génère une **grande quantité de données numériques** et s'appuie notamment sur le **développement de codes logiciels**[1] plus ou moins complexes (court script ou logiciel élaboré). Certaines données produites à l'Institut Pasteur sont appropriables c'est-à-dire protégées et/ou protégeables par un droit de propriété intellectuelle (inventions, logiciels, base de données) ; d'autres impliquent de recourir au secret (données à caractère personnel, savoir-faire, secret des affaires…).

En l'absence d'un corpus de règles homogènes, un état des lieux[2] des usages et pratiques en matière de gestion et partage des données de la recherche au sein de l'Institut Pasteur a été dressé en 2019. Il a permis de constater des disparités de traitement et de gestion des données et codes logiciels au sein d'un large panel de répondants.

L'Institut Pasteur travaille en **étroite collaboration avec des établissements publics** d'enseignement supérieur (universités) et des établissements publics de recherche (CNRS, INSERM, INRAE…), au travers d'unités fonctionnelles telles que les unités mixtes (UMR, USR…) mais aussi dans le cadre de partenariats public/privé. De plus, de nombreuses collaborations ont lieu avec des établissements internationaux, notamment avec les instituts du Réseau International des Instituts Pasteur. La recherche bénéficie également de **financements extérieurs de bailleurs de fonds publics et privés,** qui peuvent exiger, en contrepartie des financements accordés, la gestion et l'ouverture des données et codes logiciels générés au cours des projets, dans le respect du principe « Aussi ouvert que possible, aussi fermé que nécessaire ».

Il est donc apparu essentiel pour l'Institut Pasteur de **fixer des lignes directrices pour la gestion et le partage des données de la recherche et codes logiciels**, à la fois cohérentes avec celles des établissements partenaires de l'Institut Pasteur et alignées sur les exigences des financeurs de la recherche et éditeurs de revues. L'objectif est d'accompagner les chercheurs et de s'assurer que les données et codes logiciels soient structurés et gérés selon **les principes FAIR*** [3], et quand cela est possible « ouverts ».

La mise en place des bonnes pratiques de gestion des données de la recherche et codes logiciels est bénéfique à différents niveaux. Elle permet de :

- améliorer la **qualité**, l'**intégrité** et la **reproductibilité** de la recherche, dans la mesure où les données sont précises, complètes, authentiques et fiables ;
- renforcer la **sécurité des données** et prémunir l'Institut Pasteur contre le risque de perte ou de vol de données ;

[1] Le terme « code logiciels » employé dans la présente politique fait référence à la fois aux composants logiciels, qu'il s'agisse de scripts, programmes ou workflows, et aux logiciels plus élaborés.

[2] Les résultats de l'état des lieux sont disponibles sur l'Intranet.

[3] Les termes suivis d'un astérisque sont définis dans le glossaire en annexe.

- rendre les données **accessibles** et **compréhensibles** sur le long terme, qu'elles soient rendues publiques ou non ;
- faciliter la **réutilisation** des données, soit par l'investigateur initial, soit par des collaborateurs pasteuriens, soit par d'autres scientifiques si les données sont rendues publiques et ainsi **éviter leur duplication**, permettant un gain de temps et de ressources.

De plus, des données bien gérées peuvent être partagées et diffusées au sein de la communauté scientifique, pour ainsi :

- augmenter la **visibilité** et l'**impact** du travail du scientifique ;
- favoriser la **collaboration** scientifique et permettre des **recherches interdisciplinaires** ;
- permettre de mener des **nouvelles recherches** non envisagées par l'investigateur initial et déboucher sur des innovations.

2. Domaine d'application

La présente politique s'adresse à **tout le personnel scientifique** (chercheurs, ingénieurs, chefs de projet, doctorants, post-doctorants...) travaillant à l'Institut Pasteur de Paris (campus de la rue du docteur Roux et Institut de l'Audition, pasteuriens et OREX) et dans ses établissements qui lui sont rattachés juridiquement (Instituts Pasteur de Guyane, Guadeloupe et Nouvelle-Calédonie), quelle que soit leur entité de rattachement (unité de recherche, G5, plateforme, centre national de référence…). Dans la suite du document, le terme « Institut Pasteur » englobe donc à la fois l'Institut Pasteur de Paris et ses établissements.

La politique concerne toutes les **données de la recherche**, c'est-à-dire les enregistrements factuels (chiffres, textes, images, sons…) qui sont utilisés comme source principale pour la recherche scientifique et sont nécessaires pour valider les résultats de recherche[4]. Elle s'applique à toutes les données générées ou collectées dans le cadre de projets de recherche conduits à l'Institut Pasteur et financés par des fonds externes ou internes. Elle s'applique également à tous les **composants logiciels (scripts, programmes ou workflows)** et aux **logiciels plus élaborés**, développés à l'Institut Pasteur.

3. Objectifs de la politique

Cette politique fournit un **document de référence** clef en main précisant les **lignes directrices** et les **bonnes pratiques** que l'Institut Pasteur recommande de suivre pour la gestion et le partage des données de la recherche et codes logiciels.

Elle spécifie les **rôles, droits et responsabilités** de chacun :

- rôle, droits et responsabilités des scientifiques dans le traitement et la gestion des données de la recherche et codes logiciels ;
- rôle, droits et responsabilités de l'institution dans l'adoption et la mise en place de la politique.

[4] Source de la définition : OCDE, 2007. « Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics ». Disponible sur : http://www.oecd.org/fr/science/inno/38500823.pdf

Cette politique a été élaborée afin de sensibiliser les scientifiques aux **bonnes pratiques** en les impliquant et en leur offrant les **moyens opérationnels** de les mettre en œuvre. Elle est associée à des fiches pratiques dont la liste est disponible en annexe.

4. Rôle, droits et responsabilités du scientifique

4.1. PLANIFIER, ANTICIPER

L'Institut Pasteur demande la mise en place d'un **Plan de Gestion de Données*** (PGD ou Data Management Plan, DMP) au début de chaque projet de recherche financé par un bailleur externe ou par l'Institut Pasteur et dont le coordinateur est membre de l'Institut Pasteur. Le PGD doit être initié le plus en amont possible d'un projet et une première version doit être déposée dans l'**espace dédié sur les serveurs de l'Institut Pasteur** dans les 6 premiers mois du projet. Le PGD est ensuite mis à jour régulièrement jusqu'à la fin du projet. Les scientifiques devront décider en fin de projet si la **version finale du PGD** est confidentielle, si elle peut être diffusée en interne ou si elle peut être rendue accessible publiquement sur le site research.pasteur.fr.

L'Institut Pasteur demande à chaque entité de recherche de se doter d'un **plan général de gestion des données**. Ce PGD d'entité, plus succinct qu'un PGD de projet, relève de la responsabilité du chef d'entité. Il permet d'harmoniser les pratiques au sein de l'entité, de guider les nouveaux arrivants dans la gestion et le partage de leurs données de recherche et de faciliter la mise en place de PGD de projet.

4.2. RÉUTILISER, CRÉER DES DONNÉES

L'Institut Pasteur encourage les scientifiques à **réutiliser des données existantes** avant de créer leurs propres données, et ce afin d'éviter la duplication de données. Les scientifiques peuvent consulter de nombreuses sources afin de trouver des données adaptées à leur besoin.

La réutilisation des données doit se faire dans le **respect du cadre juridique**, qu'il s'agisse des prescriptions réglementaires, légales, ou d'exigences contractuelles. De plus, **les jeux de données réutilisés doivent être cités** dans les publications, en suivant si possible le modèle proposé par l'Institut Pasteur. En effet, l'Institut Pasteur considère que les jeux de données sont des produits de la recherche légitimes et citables, en accord avec les principes décrits dans la *Joint Declaration of Data Citation Principles*[5].

La **création ou la collecte de nouvelles données** doit se faire dans le respect du cadre juridique. Les scientifiques se doivent de vérifier si des contraintes juridiques s'appliquent à leur projet de recherche avant toute création ou collecte de données. En particulier, l'Institut Pasteur tient à souligner que dans le cas d'une **recherche impliquant la personne humaine et/ou d'une recherche portant sur des données à caractère personnel** (soit des informations relatives à une personne déterminée ou déterminable) **ou générant de**

[5] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk

nouvelles données à partir d'échantillons préexistants, les scientifiques doivent informer le Délégué à la protection des Données (DPO) en amont du projet et peuvent se faire accompagner par la Coordination Clinique du Centre de Recherche Translationnelle (CRT-CC) pour cette démarche.

Pour faciliter l'utilisation et la réutilisation des données, l'Institut Pasteur encourage à recourir aux **formats ouverts**, **standards et interopérables*** et incite à la **création de métadonnées*** selon les **standards*** du domaine dès la génération des premières données.

4.3. UTILISER, TRAITER, STOCKER

L'Institut Pasteur recommande fortement d'utiliser **les infrastructures fournies par la DSI** pour le traitement et le stockage des données de recherche et des logiciels scientifiques, quel que soit le niveau de sensibilité des données. La directive de classification de l'information permet de déterminer le **niveau de sensibilité** des données et **les règles de manipulation applicables.**

Les données de la recherche avant publication doivent être **stockées et diffusées via des outils adaptés à leur sensibilité** et n'être accessibles qu'aux personnes autorisées. Les **droits d'accès** aux données doivent être définis dès le début du projet. L'Institut Pasteur recommande de s'assurer que les données de la recherche sont sauvegardées régulièrement sur des espaces de stockage maîtrisés et contrôlés. Leur manipulation, transfert ou échange doit également être réalisé via des outils adaptés à leur sensibilité.

Chaque entité dispose d'un espace de stockage sur les serveurs de la DSI pour les données de l'entité et peut demander la création d'un autre espace pour chaque projet de recherche, avec des accès limités aux personnes intervenant sur le projet. De façon à ce que chacun puisse facilement retrouver les données, l'Institut Pasteur recommande d'organiser ces espaces de stockage à l'aide d'un **plan de classement***. Pour faciliter le repérage et l'identification des données, il est également recommandé d'adopter des **règles de nommage*** précises et communes.

L'Institut Pasteur recommande de mettre en place dans les bases de données des **contrôles de qualité et de cohérence** des données adaptés et de **tracer et documenter** les actions effectuées sur les données. Les **métadonnées** décrivant les traitements effectués sur les données doivent être créées et associées aux données dès leur création et tout au long du traitement des données. En particulier, les données concernant la personne humaine, si possible gérées avec l'eCRF de l'Institut Pasteur, doivent être tracées et gérées selon des bonnes pratiques de recherche clinique et dans le respect de la réglementation en vigueur.

4.4. DIFFUSER, PUBLIER, PARTAGER

L'Institut Pasteur encourage la **diffusion, la publication et le partage** des données de la recherche et codes logiciels en libre accès, selon le principe **« aussi ouvert que possible, aussi fermé que nécessaire »**.

Certaines données ne peuvent pas être partagées ou seulement sous certaines conditions. La **publication ou le partage des données** doit se faire dans le respect du cadre juridique qui

s'y applique le cas échéant. En particulier, avant de rendre publiques ou de partager des données, **l'Institut Pasteur demande aux scientifiques de vérifier** :

- Que les données ne sont pas sensibles ou confidentielles ;
- Que les données ne permettent en aucun cas l'identification directe ou indirecte des personnes se prêtant à la recherche ;
- Que les données ne peuvent pas faire l'objet d'une protection par un titre de propriété industrielle (demande de brevet par exemple) ;
- Que les données ne peuvent pas être valorisées économiquement ;
- Que le partage des données est bien compatible avec les termes et conditions du contrat de recherche.

De plus, il convient de vérifier que le moyen de partage ou de mise à disposition est bien adapté à la sensibilité des données.

Pour toutes les données qui peuvent être publiées ou partagées car elles ne sont pas soumises aux contraintes citées ci-dessus :

L'Institut Pasteur reconnaît que toutes les données ne peuvent pas être partagées. Dans tous les cas, que la recherche soit financée ou non sur fonds publics, il est recommandé aux scientifiques de publier au minimum **les données qui sous-tendent les publications***, au moment de la publication des articles (données brutes - si non confidentielles - et données analysées). Dans la mesure où les scientifiques ont procédé aux vérifications énoncées ci-dessus, ils demeurent libres de publier ou partager toute autre donnée. Ils sont notamment **encouragés à rendre publics les résultats négatifs** (données qui ne donneront pas lieu à un article scientifique).

L'Institut Pasteur demande que les jeux de données soient **partagés à un large public** en utilisant les moyens de partage adéquats : bases de données, entrepôts de données*. Partager ses données ne signifie pas que les données seront accessibles à tous sans restriction. Certains entrepôts permettent de contrôler l'accès aux données (accès sur demande ou sous réserve de l'approbation d'un comité scientifique par exemple). Si les données doivent rester en accès restreint, l'Institut Pasteur demande que les métadonnées soient accessibles afin de signaler l'existence des données et d'indiquer les conditions d'accès.

Les scientifiques devront publier des jeux de données **de qualité** et **réutilisables** en étant correctement décrits, documentés et contextualisés. Pour être conforme aux exigences des financeurs, la mise à disposition des données doit suivre les **principes FAIR***. Les jeux de données mis à disposition du public doivent toujours être associés à une **licence de diffusion/exploitation*.**

L'Institut Pasteur encourage la publication de **data papers***, publications revues par les pairs qui permettent de décrire et valoriser un jeu de données* et sont citables au même titre que toute autre publication.

Cas particulier des urgences de santé publique :
Selon les principes édictés par GloPID-R[6] (*Global Research Collaboration for Infectious Disease Preparedness*), l'Institut Pasteur exige que les données liées à une urgence de santé

[6] GLoPID-R (2018). Principles of Data Sharing in Public Health Emergencies. https://www.glopid-r.org/wp-content/uploads/2018/06/glopid-r-principles-of-data-sharing-in-public-health-emergencies.pdf

publique soient **partagées et mises à disposition le plus rapidement possible**, avec le moins de restrictions d'accès possible. Il est toutefois essentiel de **s'assurer de la qualité des données avant leur mise à disposition**. Un juste équilibre doit être trouvé entre l'exigence de rapidité du partage et la conformité avec les principes FAIR*.

4.5. PRÉSERVER, ARCHIVER

A l'issue de chaque projet de recherche et **avant chaque fermeture d'entité**, l'Institut Pasteur demande aux scientifiques de conserver les données produites dans un espace de conservation spécifique pour assurer leur **pérennité**. Pour que les données soient compréhensibles et réutilisables, les données déversées dans cet espace doivent être **correctement décrites et documentées** et être enregistrées dans un **format durable***.

L'Institut Pasteur demande aux scientifiques de vérifier la **réglementation** et les **contraintes légales ou contractuelles de conservation** associées à leurs données. En particulier, dans le cas de données à caractère personnel (soit de données relatives à une personne déterminée ou déterminable), la loi prescrit le tri des données pour ne conserver que les données "adéquates, pertinentes, et non excessives" au regard des finalités pour lesquelles elles ont été collectées ou traitées.

Dans tous les cas, avant de déverser leurs données dans l'espace de conservation, l'Institut Pasteur demande aux scientifiques d'**effectuer un tri** parmi les données pour ne conserver que les données dont la conservation est obligatoire ou ayant un intérêt sur le long terme :

- Données soumises à des **obligations légales ou contractuelles de conservation** ;
- Données ayant valeur de **preuve** (preuve d'antériorité par exemple) ;
- Données pouvant servir à la **reproductibilité** de travaux scientifiques ;
- Données **uniques**, non reproductibles ou difficilement reproductibles.

4.6. GÉRER DES LOGICIELS, SCRIPTS, PROGRAMMES

L'Institut Pasteur encourage les scientifiques à **réutiliser des logiciels ou composants logiciels existants** avant tout développement, en vérifiant au préalable les contraintes liées aux licences qui leur sont associées. Des répertoires existent et permettent aux scientifiques de trouver les ressources dont ils ont besoin.

Lorsque les scientifiques réutilisent des logiciels ou des composants logiciels d'un tiers, ils doivent **citer le composant logiciel réutilisé** dans leurs publications, en suivant si possible le modèle proposé par l'Institut Pasteur. En effet, l'Institut Pasteur considère que les logiciels sont des produits de recherche légitimes et citables, en accord avec les *Software Citation Principles*[7].

L'Institut Pasteur encourage tous les scientifiques développant des composants logiciels, qu'il s'agisse de scripts, programmes ou workflows, à suivre les **bonnes pratiques de développement**, notamment de les versionner, documenter et tester, et dans la mesure du possible de faire appel aux standards du domaine. Il est recommandé de déposer les logiciels

[7] Smith AM, Katz DS, Niemeyer KE, FORCE11 Software Citation Working Group. (2016) Software Citation Principles. *PeerJ Computer Science* 2:e86. https://doi.org/10.7717/peerj-cs.86

créés à l'Institut Pasteur dans la **forge logicielle mise à disposition par l'Institut Pasteur**, depuis leur création et jusqu'à leur diffusion éventuelle.

Pour éviter d'introduire des vulnérabilités, le développement doit se faire en suivant les bonnes pratiques indiquées dans le guide de développement sécurisé.

Avant de diffuser un logiciel, l'Institut Pasteur demande aux scientifiques de vérifier s'il ne peut pas faire l'objet d'une **application industrielle** (en accord avec les contraintes imposées par les financeurs). Dans ce cas, les scientifiques s'engagent à en informer, sans délai, et en priorité, le Service des Brevets et Inventions de l'Institut Pasteur et à établir auprès d'eux une déclaration d'invention.

Si le logiciel peut être diffusé, l'Institut Pasteur recommande de rendre le code source public, de fournir des packages et d'enregistrer le logiciel sur un catalogue de logiciels. Un logiciel publié doit toujours être associé à une **licence de diffusion/exploitation***. Il est par ailleurs rappelé que le choix de la licence peut avoir des conséquences sur l'exploitation future du logiciel (commerciale, académique, gratuite, payante…).

Un logiciel diffusé, indépendamment de sa licence, peut faire l'objet d'une **valorisation**, sous différentes formes : création de start-up, support et formation, développement de fonctionnalités spécifiques, valorisation du savoir-faire… Dans tous les cas, si un scientifique est contacté par un industriel à propos d'un de ses logiciels publiés, il doit contacter le service Transfert de Technologie.

Afin de permettre à d'autres de retrouver et citer la bonne version du code logiciel (par exemple, celle utilisée et citée dans une publication), l'Institut Pasteur recommande de **déposer cette version sur HAL-Pasteur**. Un transfert vers **Software Heritage** sera automatiquement effectué afin d'assurer la pérennisation du logiciel et de le rendre citable. En effet, Software Heritage attribue un identifiant unique et pérenne* à chaque composant logiciel.

Pour planifier toutes les étapes du cycle de vie du logiciel (depuis l'idée jusqu'à la diffusion éventuelle), l'Institut Pasteur recommande la mise en place d'un **plan de gestion de logiciel*** (PGL ou Software Management Plan, SMP) dès le début de la conception d'un logiciel de recherche. Le PGL devra être déposé sur l'**espace dédié sur les serveurs de l'Institut Pasteur**.

5. Rôle, droits et responsabilités de l'institution

5.1. METTRE À DISPOSITION UNE INFRASTRUCTURE

L'Institut Pasteur met à disposition des scientifiques des **infrastructures** adéquates, robustes et économiquement viables pour la gestion et le stockage des données de la recherche et codes logiciels. La Direction des Systèmes d'Information accompagne les scientifiques et propose des outils facilitant l'utilisation de ces infrastructures.

5.2. ACCOMPAGNER ET FORMER

L'Institut Pasteur accompagne les scientifiques dans l'application de la présente politique. Chaque **fiche pratique** (liste complète en annexe) indique les services à contacter, les outils

disponibles et les formations existantes. En particulier, l'Institut Pasteur s'engage à développer des formations concernant la gestion des données.

Par ailleurs, l'Institut Pasteur apporte un **soutien opérationnel** aux équipes de recherche sur les différents aspects décrits dans la présente politique (voir les contacts dans chaque fiche pratique).

5.3. RECONNAÎTRE ET VALORISER LES BONNES PRATIQUES

L'Institut Pasteur considère que les jeux de données et les logiciels de recherche sont des produits de recherche légitimes et éligibles à une évaluation. De plus, l'Institut Pasteur reconnaît que l'activité de gestion des données et codes logiciels fait partie intégrante du processus de recherche et contribue à la qualité de la recherche.

A ce titre, à partir de 2021, le **COMESP** (comité d'évaluation des activités scientifiques des personnels) sera sensible aux diverses initiatives pour gérer et rendre accessible les jeux de données et logiciels de recherche, notamment la mise en place de plans de gestion des données et de plans de gestion des logiciels pour les projets de recherche menés à l'Institut Pasteur.

De plus, le plan de gestion des données de l'entité ainsi que toute autre initiative mise en place pour la gestion et la diffusion des données et codes logiciels seront pris en compte dans l'**évaluation de l'entité**.

Par ailleurs, l'Institut Pasteur reconnaît les logiciels comme des produits de recherche valorisables pouvant donner lieu à une rémunération selon l'accord d'entreprise en vigueur.

6. Glossaire

Data paper (ou Data Article)
Publication scientifique examinée par les pairs dont le but principal est de décrire un ou plusieurs jeux de données. Les données décrites doivent être accessibles, soit sous forme de fichiers annexés, soit par un lien pérenne vers l'entrepôt où elles sont déposées. Le data paper peut être publié dans un data journal (une revue contenant exclusivement des data papers) ou dans une revue scientifique classique.

Données qui sous-tendent les publications
Données nécessaires à la validation des résultats présentés dans les publications scientifiques.

Entrepôt de données
Un entrepôt permet de stocker des données de la recherche mais aussi de les retrouver et de les réutiliser grâce à une description par des métadonnées. Il existe de nombreux entrepôts de données répartis en plusieurs types : disciplinaires, multidisciplinaires, propres à un éditeur, institutionnels...

Format durable
Un format peut être considéré comme « durable » s'il est ouvert (format dont les spécifications internes sont librement accessibles), largement utilisé et normalisé (si possible).

Formats ouverts, **standards et interopérables**
Un format ouvert est un format dont les spécifications fonctionnelles et techniques sont publiques et disponibles gratuitement (ou à faible coût). Un format ouvert et standardisé (par exemple, le format PDF/A normalisé ISO 19005) est interopérable : une donnée enregistrée dans ce type de format est indépendante du logiciel utilisé pour la créer ; elle pourra être lue et modifiée par tous les logiciels destinés à traiter le type du fichier (image, texte, audio, etc.).

Identifiant unique et pérenne (ou PID pour Persistent IDentifier)
Un identifiant unique et pérenne permet l'identification unique d'une ressource numérique et sa citation. Il garantit un lien stable à la ressource en ligne en faisant correspondre en permanence l'identité de la ressource à sa localisation sur le web.

Jeu de données (ou Dataset)
Un jeu de données est un groupement de données similaires ou connexes, rassemblées pour former un ensemble cohérent. Un jeu de données doit toujours être accompagné de métadonnées pour faciliter sa compréhension et sa réutilisation.

Licence de diffusion/exploitation
Une licence de diffusion est un instrument juridique, complémentaire au droit d'auteur. Elle permet au titulaire des droits sur une œuvre d'accorder à l'avance aux utilisateurs certains droits d'utilisation de cette œuvre.

Métadonnées
Les métadonnées sont des informations structurées servant à décrire une donnée ou une ressource. Les métadonnées accompagnent les données pour permettre à la communauté scientifique d'y accéder, de les comprendre et de les réutiliser.

Plan de classement
Le plan de classement est l'arborescence de répertoires et de dossiers, avec des intitulés intelligibles par tous, commune à un projet ou à une entité. Le plan de classement permet d'aider chacun à retrouver plus facilement les données et de faciliter la transmission des informations dans le contexte de turn-over régulier des personnels (doctorants, post-doctorants, CDD...).

Plan de Gestion des Données (ou Data Management Plan)
Le Plan de Gestion de Données est un document synthétique et évolutif qui aide à organiser et anticiper toutes les étapes du cycle de vie de la donnée. Il est rédigé au début d'un projet de recherche puis régulièrement mis à jour et explique la façon dont les données de recherche recueillies ou générées seront gérées durant le projet et après son terme.

Plan de Gestion de Logiciel (ou Software Management Plan)
Un Plan de Gestion de Logiciel est un document évolutif décrivant les informations concernant un logiciel de la recherche, la façon dont il est conçu et développé, ses objectifs, à qui il s'adresse, les résultats attendus et obtenus, son éventuelle diffusion, les informations de propriété intellectuelle... tout au long du cycle de vie de ce logiciel.

Principes FAIR
Les principes FAIR (Findable, Accessible, Interoperable, Reusable) publiés en 2016, correspondent à des lignes directrices et des bonnes pratiques dont l'objectif est d'améliorer la réutilisation des données de la recherche[8]. Ces principes peuvent également s'appliquer aux logiciels de recherche en les adaptant légèrement[9].

Règles de nommage
Adopter des règles de nommage précises et communes est nécessaire pour repérer et identifier les données, éviter les problèmes lors des transferts et permettre leur conservation à moyen et long terme. Quelques règles à respecter : donner un nom bref et explicite, ne pas mettre d'espace ni de caractères spéciaux, indiquer les versions…

Standards de métadonnées
Les standards de métadonnées sont des modèles qui précisent toutes les métadonnées nécessaires pour décrire une ressource. Par exemple, le standard *Minimum Information about a Genotyping Experiment (MIGen)* décrit toutes les métadonnées à fournir pour pouvoir comprendre et réutiliser des données issues d'une expérience de génotypage.

[8] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* (2016). https://doi.org/10.1038/sdata.2016.18

[9] Lamprecht, A.-L., Garcia, L., Kuzak, M. *et al.* Towards FAIR principles for research software. Data Science (Pre-press, 2019). https://doi.org/10.3233/DS-190026

Annexe de la politique de gestion et partage des données de la recherche et codes logiciels – Fiches pratiques –

Les fiches pratiques sont disponibles sur l'intranet de l'Institut Pasteur (accès limité).

TITRE DE LA FICHE PRATIQUE	BREF RESUMÉ
Le plan de gestion des données de projet : un outil pour vous aider à planifier la gestion de vos données de recherche	Un Plan de Gestion des Données (PGD) ou Data Management Plan (DMP) est un document qui définit comment sont gérées les données d'un projet. Cette fiche pratique répond à différentes questions : - Quelles sont les informations demandées dans le PGD ? - Comment utiliser la trame proposée par l'Institut Pasteur ? - Quels outils utiliser ? - Quand doit être rédigé le PGD et où déposer les différentes versions ?
Mettre en place un plan de gestion des données dans votre entité	Un plan de gestion des données d'entité (PGD d'entité) est un document qui définit comment sont gérées les données d'une entité (unité de recherche, CNR, plateforme…). Cette fiche pratique vous explique à quoi sert un PGD d'entité et comment utiliser le template proposé par la bibliothèque du CeRIS.
Sources pour trouver des données et composants logiciels à réutiliser	Avant de générer des données ou de développer un logiciel, il est recommandé de vérifier s'il n'est pas possible de réutiliser des jeux de données ou des logiciels produits par d'autres scientifiques. Cette fiche pratique vous aide à trouver les ressources dont vous avez besoin : - les différentes sources que vous pouvez consulter - les étapes pour une recherche efficace.
Questions juridiques liées à la réutilisation des données	Cette fiche pratique liste les questions que chaque scientifique doit se poser avant de réutiliser un jeu de données, de façon à vérifier que la réutilisation se fait bien dans le respect du cadre juridique en vigueur.
Citer un jeu de données ou un logiciel de recherche	Avant de générer des données ou de développer un logiciel, il est recommandé de vérifier s'il n'est pas possible de réutiliser des jeux de données ou des logiciels produits par d'autres scientifiques. Cette fiche pratique vous propose un modèle pour citer, dans vos publications, les jeux de données ou logiciels réutilisés.
Collecter des données dans le cadre d'une recherche en santé impliquant des données à caractère personnel	Cette fiche précise la notion de données à caractère personnel et présente succinctement les points de vigilance avant leur collecte.

Décrire ses données de recherche : métadonnées et documentation	Cette fiche pratique vous aide à comprendre les différentes façons de décrire des données scientifiques : par de la documentation et des métadonnées. La fiche aborde les sujets suivants : - Quelles métadonnées associer aux données ? - Pourquoi associer des métadonnées en plus de la documentation ?
Nature et format des données dans le domaine biomédical	Pour pouvoir gérer correctement ses données, il est important de les définir précisément au préalable : quel type de données et quel format ? Dans cette fiche pratique, nous vous proposons une liste non exhaustive de types de données et de formats dans le domaine biomédical. Cette fiche précise également la différence entre formats ouverts et formats fermés.
Directive de classification de l'information	La classification de l'information a pour but d'identifier les informations sensibles et d'harmoniser les pratiques au sein de l'Institut Pasteur en matière de gestion et de protection de l'information. Elle permet de déterminer objectivement le niveau de sensibilité de l'information grâce à différentes échelles et définit des règles de manipulation associées.
Tools and infrastructures for the management, storage and sharing of data and software codes at the Institut Pasteur	This practical sheet summarizes the different tools and infrastructures available at the Institut Pasteur for the management, storage and sharing of data and software codes. The tools are classified into different families: scientific information system, high performance computing cluster, scientific software development, scientific software catalogue, web communication, collaborative work, container hosting, machine/application hosting.
Organiser ses espaces de stockage pour assurer le repérage et la pérennité des données	Chaque entité dispose d'espaces de stockage sur les serveurs de la DSI. Cette fiche pratique vous donne des conseils sur la façon d'organiser ces espaces de stockage : - en mettant en place un plan de classement - en adoptant des règles de nommage précises et communes - en créant un espace de conservation et en y transférant régulièrement les données finalisées.
Bonnes pratiques de gestion des données dans REDCap	Cette fiche résume les bonnes pratiques de gestion des données, en particulier lors de l'utilisation de REDCap® : les questions à se poser en amont, comment améliorer la qualité des données dans vos projets et surtout, mettre « en production » vos projets REDCap® avant toute collecte de données réelles pour garantir leur intégrité.
Questions juridiques liées à la diffusion des données	Cette fiche pratique liste les questions que chaque scientifique doit se poser avant de publier ou partager un jeu de données, de façon à vérifier que la diffusion se fait bien dans le respect du cadre juridique en vigueur.
Transfert d'information depuis le SI Pasteur	Ce document a pour objectif de synthétiser les informations concernant les moyens de transfert de données depuis et à destination du système d'Information de l'Institut Pasteur.

Partager ses données à un large public via les entrepôts de données	Pour partager ses données à un large public, la meilleure solution est de les déposer dans un entrepôt de données. Cette fiche pratique vous explique pourquoi et comment partager vos données dans un entrepôt, interne ou externe.
Comment rendre ses données FAIR et choisir une licence de diffusion ?	Les principes FAIR (Findable, Accessible, Interoperable, Reusable) correspondent à des lignes directrices dont l'objectif est d'améliorer la réutilisation des données de la recherche. Après une explication détaillée des principes FAIR, cette fiche pratique répond aux questions suivantes : - Comment faire en pratique pour rendre ses données FAIR ? - Comment choisir sa licence de diffusion ?
Le data paper : une publication pour décrire et valoriser des données	Un data paper (ou data article) est une publication scientifique examinée par les pairs dont le but est de décrire un ou plusieurs jeux de données. Cette fiche pratique vous indique pourquoi et comment publier un data paper.
Archivage des données issues de recherche sur la personne humaine	Cette fiche pratique précise les durées et modalités de conservation des données issues de la recherche clinique et de la recherche sur la personne.
Bonnes pratiques de développement logiciel	Les "bonnes pratiques" sont un ensemble de règles informelles à appliquer dans le développement d'un logiciel. Cette fiche vous explique quelles sont ces règles, pourquoi il est important de les mettre en place et quels sont les outils à votre disposition pour faciliter leur mise en œuvre.
Développement, gestion et valorisation des codes et logiciels de recherche : ressources et expertises à l'Institut Pasteur	Cette fiche pratique précise les personnes à contacter et les outils disponibles à l'Institut Pasteur sur questions de développement, gestion et valorisation des codes et logiciels de recherche.
Guide de développement sécurisé	Ce guide a pour objectif de sensibiliser et former les développeurs par rapport à la sécurité informatique. Il est composé de plusieurs sous-parties : - principes du développement sécurisé - top 10 des vulnérabilités - règles pour l'écriture de code sécurisé - bibliothèque de fonctionnalités et tests de sécurisé
Protéger et valoriser vos inventions	Cette FAQ apporte des réponses aux questions fréquemment posées par les chercheurs de l'Institut Pasteur sur divers aspects relatifs à la protection et à la valorisation des résultats de leur recherche.

Diffuser un logiciel : bonnes pratiques et choix de la licence	Diffuser un logiciel, c'est le rendre facile à trouver et accessible pour les utilisateurs, les développeurs et les organismes. Cette fiche pratique répond aux questions suivantes : - comment mettre à disposition et faire connaître son logiciel ? - comment choisir une licence pour son logiciel ?
Valoriser un logiciel	Le potentiel développement commercial de votre logiciel n'est pas seulement lié au code source que vous avez développé, mais également à votre savoir-faire et à votre capacité à développer de nouvelles fonctionnalités innovantes. Cette fiche précise les cas dans lesquels vous pouvez contacter l'équipe de Business Development ainsi que les points d'attention si vous envisagez une valorisation commerciale de votre logiciel, notamment concernant le choix de la licence.
Assurer la citation, la visibilité et la pérennisation de son code source avec HAL et Software Heritage	Afin de permettre à d'autres de retrouver et citer la bonne version de son code logiciel, il est recommandé de déposer cette version sur HAL-Pasteur. Un transfert vers Software Heritage sera automatiquement effectué afin d'assurer la pérennisation du logiciel et de le rendre citable. Cette fiche pratique vous explique pourquoi et comment faire en pratique.
Pourquoi et comment rédiger un plan de gestion de logiciel ?	Un Plan de Gestion de Logiciel (PGL) ou Software Management Plan est un document de référence qui décrit les (bonnes) pratiques mises en place lors du développement d'un logiciel. Cette fiche répond à différentes questions : - A quoi sert un PGL ? - Quel modèle est recommandé ? - Quand doit être rédigé le PGL et où déposer les différentes versions ?

Institut Pasteur
25-28, rue du Docteur Roux
75724 Paris Cedex 15, France
www.pasteur.fr

INSTITUT PASTEUR