LIVRE BLANC

Utilisation secondaire des données d'imagerie à des fins de recherche

2026

Sous la coordination de **TOURDIAS Thomas**, PUPH radiologie, (1) Service de neuroimagerie diagnostique et thérapeutique, CHU de Bordeaux; (2) Neurocentre Magendie, INSERM U1215, Université de Bordeaux

Liste alphabétique des contributeurs

- **AUBE Christophe**, PUPH radiologie, (1) Service de Radiologie, CHU d'Angers, (2) Laboratoire HIFIH-SFR ICAT, Université d'Angers
- **BEAUMONT Marine**, Ingénieure de recherche, (1) CIC-IT de Nancy, CHRU Nancy, Inserm CIC-IT 1433, Université de Lorraine; (2) Laboratoire IADI, Inserm U1254, Université de Lorraine
- **BENZAKOUN Joseph**, PUPH radiologie, (1) Université Paris Cité, Institute of Psychiatry and Neuroscience of Paris (IPNP), INSERM U1266, IMA-BRAIN, 75014 Paris; (2) GHU-Paris Psychiatrie et Neurosciences, Hôpital Sainte Anne, Neuroradiology Department, F-75014 Paris.
- **BOULOUIS Grégoire**, PUPH Radiologie, (1) Diagnostic and Interventional Neuroradiology, Tours University Hospital, Tours ; (2) Centre d'Investigation Clinique - Innovation Technologique (CIC-IT) 1415, Tours University Hospital Tours ; (3) Université de Tours, INSERM, Imaging Brain & Neuropsychiatry iBraiN U1253, 37032, Tours.
- **BOUSSEL Loïc**, PUPH radiologie, (1) Service d'imagerie cardiovasculaire et thoracique, Hospices civils de Lyon, Bron; (2) CREATIS UMR 5220, INSA-Lyon, Université Claude Bernard Lyon 1, Villeurbanne, Lyon.
- **CANETE Sophie**, Déléguée à la protection des données, (1) Recherche et innovation, Direction des Risques Cyber, Direction Générale, CHU de Bordeaux
- **CROISILLE Pierre**, PUPH radiologie, (1) Service de Radiologie, Hôpital Nord, CHU de Saint Etienne ; (2) CREATIS UMR CNRS 5220, INSERM U1294.
- **DOJAT Michel**, Directeur de recherche INSERM, (1) Grenoble Institut Neurosciences ; (2) Centre de recherche Inria de l'Université Grenoble Alpes.
- **ENGLENDER Olivier**, Responsable IA et DATA, (1) Hôpitaux Universitaire de Strasbourg, DSIIA, 1 place de l'hôpital, BP 426, 67091 Strasbourg Cedex.
- **FOURNIER Laure**, PUPH radiologie, (1) Service de Radiologie, Hôpital Européen Georges Pompidou, 75015 Paris ; (2) In vivo Imaging Research lab, UMR-S970, PARCC, Université Paris Cité, 75015 Paris.
- **Philippe GARTEISER**, Chargé de recherche INSERM, (1) Centre de Recherche sur l'Inflammation, Inserm U1149, Université Paris Cité.
- **JOLIOT Marc**, Directeur de recherche, (1) CEA Groupe d'imagerie neurofonctionelle, IMN, UMR5293, Université de Bordeaux.
- **KAIN Michael**, Ingénieur de recherche INRIA, (1) Centre de recherche Inria de Luniversité de Rennes.

- **KUCHCINSKI Grégory**, PUPH radiologie, (1) Service de Neuroradiologie, Pôle Imagerie et Explorations fonctionnelles, Hôpital Salengro, CHU de Lille ; (2) UMR-S 1172 LilNCog, Université de Lille.

- **Nathalie LASSAU**, PUPH radiologie, (1) Service de Radiologie. Institut Gustave Roussy, (2) BIOMAPS UMR1281 INSERM, CNRS, CEA, Université Paris Saclay

- **LION Stéphanie**, cheffe de projet de la plateforme Colybri - Core Lab virtuel des Hospices Civiles de Lyon, (1) Hospices Civils de Lyon, Direction de l'innovation, 69007 Lyon.

- **LOPES Renaud**, MCUPH médecine nucléaire, (1) Service de Médecine Nucléaire, Pôle Imagerie et Explorations fonctionnelles, Hôpital Salengro, CHU de Lille ; (2) UMR-S 1172 LilNCog, Université de Lille.

- **LUCIANI Alain**, PUPH radiologie, (1) Service de radiologie médicale, Hôpital Henri-Mondor, Paris ; (2) Université Paris Est Creteil.

- **MAIRE Florent**, Attaché de Recherche Clinique, (1) Pôle Imagerie Médicale, CHU Bordeaux.

- **MANSUY Adeline**, Coordinatrice cellule recherche imagerie, (1) Cellule Recherche en Imagerie (CRI), Direction des plateaux médico-techniques, Hospices Civils de Lyon.

- **MULLE Sébastien**, PUPH radiologie, (1) Service de radiologie médicale, Hôpital Henri-Mondor, Paris ; (2) Université Paris-Est Créteil.

- **NORMANT Sébastien**, Coordinateur cellule recherche imagerie, (1) Pôle Imagerie Médicale, CHU ROUEN NORMANDIE, 76031 Rouen Cedex.

- **POIRION Emilie**, Ingénieur de recherche, (1) Imaging department, Foundation A. de Rothschild Hospital, Paris.

- **ROMERO William**, Data scientist, (1) CREATIS UMR 5220, INSERM1294, INSA-Lyon, Université Jean-Monnet, Saint-Etienne.

- **RONOT Maxime**, PUPH radiologie, (1) Service de radiologie, Université Paris Cité, FHU MOSAIC2, Hôpital Beaujon APHP.Nord, Clichy.

- **SALESSES Fabien**, Ingénieur Coordonnateur des activités de Recherche en Imagerie, (1) Pôle Imagerie Médicale, CHU Bordeaux.

- **TEILLAC Achille**, Chef de projet, ingénieur de recherche, (1) Hospices Civils de Lyon, Direction des Plateaux Médico-Techniques, 69003 Lyon.

- **Régine TREBOSSEN**, Directrice de recherche CEA, (1) Institut des sciences du vivant Frédéric Joliot.

- **VIALLON Magalie**, MR Physicist, (1) Service de Radiologie, Hôpital Nord, CHU de Saint Etienne ; (2) CREATIS UMR 5220, INSERM1294, INSA-Lyon, Université Jean-Monnet, Saint-Etienne.

TABLE des MATIERES

TABLE ABREVIATIONS 3
PREAMBULE 5
INTRODUCTION 6
1. Panorama des données d'imagerie médicale disponibles pour la recherche 6
2. Place et spécificités des données d'imagerie dans les bases de données françaises 9
SPECIFICITES ET RECOMMANDATIONS 13
1. Nature et structuration des données d'imagerie 13
1.1. Fichiers de données brutes 13
1.2. Fichiers DICOM 14
- Définition de la norme DICOM 14
- Anatomie d'un fichier d'imagerie DICOM 14
- Hétérogénéité des implémentations de la norme DICOM 16
- Protocole de transfert DICOM 17
- Recommandations en lien avec les fichiers DICOM 19
- Spécificités des données d'échographie 19
1.3. Formats et organisations de fichiers adaptés à la recherche 20
- Formats NIfTI 21
- Organisation des fichiers (BIDS) 21
2. Enjeux liés à la pseudonymisation et à la confidentialité 22
2.1. Principes généraux de pseudonymisation des données d'imagerie 22
2.2. Recommandations pour pseudonymiser les données DICOM 24
- Pseudonymisation de l'en-tête DICOM 25
- Contrôle du contenu pixel (Pixel burning) 26
2.3. Recommandations pour les autres formats d'image 27
- Le format NIfTI 27
- Autres formats spécifiques 28
2.4. Cas particulier de la neuroimagerie (défacialisation) 28
2.5. Données "non image" associées 29
2.6. Risques spécifiques liés à l'intelligence artificielle générative 30
2.7. Bonnes pratiques et traçabilité 32

3. Spécificités liées à l'hébergement et aux échanges de données d'imagerie 33
3.1. Nécessité d'un hébergement adapté à la recherche 33
3.2. Modèle d'architecture globale 34
3.3. Sécurité, conformité et traçabilité 37
3.4. Conservation, compression et durée de stockage des données d'imagerie pour la recherche 38
3.5. Articulation entre l'échelle locale, régionale, nationale et Européenne 40
4. Réintégration de métadonnées et des données annotées 42
4.1. Principe de la ré-utilisation tertiaire des données 42
4.2. Type et format des métadonnées et données annotées 42
- Données tabulaires 42
- Données structurées 43
- Segmentations et images dérivées 44
- Autres données 44
4.3. Méthodologie de description 45
- Définir les éléments pertinents 45
- Décrire les données 45
- Décrire les producteurs de données 46
4.4. Traçabilité et conservation des résultats de recherche en imagerie 46
5. Rémunération/Valorisation 48
5.1. De la donnée au savoir : principes de valorisation 48
5.2. Cadre éthique et juridique : la notion de dépositaire 49
5.3. Cartographie des contributions et gouvernance scientifique 49
5.4. Reconnaissance académique 50
5.5. Reconnaissance financière et rémunération : modèle opérationnel 51
SYNTHESE des RECOMMANDATIONS 54
1. Recommandations en lien avec la nature et la structuration des données d'imagerie 54
2. Recommandations liées à la pseudonymisation et à la confidentialité des données d'imagerie 55
3. Recommandations liées à l'hébergement et aux échanges des données d'imagerie 57
4. Recommandations relatives à la réintégration des métadonnées et des données annotées 59
5. Recommandations relatives à la valorisation, à la gouvernance scientifique et à la rémunération 60

ABREVIATIONS

AI : Artificial Intelligence (Intelligence artificielle)

ANSSI : Agence Nationale de la Sécurité des Systèmes d'Information

API : Application Programming Interface

BIDS : Brain Imaging Data Structure

CEA : Commissariat à l'Énergie Atomique et aux Énergies Alternatives

CESREES : Comité Éthique et Scientifique pour les Recherches, les Études et les Évaluations dans le domaine de la Santé

CERF : Collège des Enseignants en Radiologie de France

CNIL : Commission Nationale de l'Informatique et des Libertés

CSV : Comma-Separated Values

DICOM : Digital Imaging and Communications in Medicine

DICOM SR : Digital Imaging and Communications in Medicine – Structured Reporting

DPI : Dossier Patient Informatisé

DPO : Délégué à la Protection des Données

EDS : Entrepôt(s) de Données de Santé

EHDS / eHDS : European Health Data Space (Espace Européen des Données de Santé)

FLI : France Life Imaging

FHIR : Fast Healthcare Interoperability Resources

GAN : Generative Adversarial Network

HDH : Health Data Hub

HDS : Hébergeur de Données de Santé

HL7 : Health Level Seven

ICMJE : International Committee of Medical Journal Editors

I**HE** : Integrating the Healthcare Enterprise

IRM : Imagerie par Résonance Magnétique

ISO : International Organization for Standardization

JSON : JavaScript Object Notation

LLM : Large Language Model

MERRI : Missions d'Enseignement, de Recherche, de Référence et d'Innovation
MR-004 : Méthodologie de Référence n°004 de la CNIL
NIfTI : Neuroimaging Informatics Technology Initiative
NLP : Natural Language Processing (Traitement automatique du langage)
OCR : Optical Character Recognition
PACS : Picture Archiving and Communication System
PDF : Portable Document Format
RGPD : Règlement Général sur la Protection des Données
RIS : Radiology Information System
RIPH : Recherches Impliquant la Personne Humaine
RSSI : Responsable de la Sécurité des Systèmes d'Information
RT-STRUCT : Radiotherapy Structure Set (DICOM)
SFR : Société Française de Radiologie
SIGAPS : Système d'Interrogation, de Gestion et d'Analyse des Publications Scientifiques
SIGREC : Système d'Information et de Gestion de la Recherche Clinique
SNDS : Système National des Données de Santé
SSO : Single Sign-On
TEP : Tomographie par Émission de Positons
UID : Unique Identifier
VNA : Vendor Neutral Archive
WADO : Web Access to DICOM Objects
XML : eXtensible Markup Language

PREAMBULE

Les données médicales peuvent être collectées dans des projets de recherche spécifiques mais sont très majoritairement produites à des fins de prise en charge clinique des patients. Leur réutilisation à d'autres fins, telles que des questions de recherche ou le développement d'algorithmes d'intelligence artificielle (après des étapes d'anonymisation ou de pseudonymisation et de structuration), offre un potentiel majeur pour faire progresser la science dans une logique d'ouverture et de partage. Cela impose une structuration des données, des infrastructures adaptées et le respect d'un cadre réglementaire pour pouvoir réutiliser ces données de façon optimale et respectueuse des droits des patients.

De nombreux efforts sont faits en ce sens au niveau local (création d'entrepôts de données de santé) et national (Health Data Hub, rapport interministériel de l'Inspection Générale des Affaires Sociales, IGAS, pour fédérer les acteurs et libérer l'utilisation des données, https://igas.gouv.fr/federer-les-acteurs). Néanmoins, ces efforts se concentrent essentiellement sur les données du dossier médical informatisé (données cliniques et biologiques) mais **ne prennent pas ou peu en compte des données plus complexes telles que les données d'imagerie médicale**. Il existe en effet des spécificités propres aux données d'imagerie par rapport à des données textes du dossier clinique ou de biologie qui nécessitent une considération particulière.

Ce document propose ***(i)*** une description des spécificités des données d'imagerie ainsi que ***(ii)*** plusieurs recommandations en vue de leur réutilisation à des fins de recherche. Le document vise ainsi à favoriser l'harmonisation des pratiques à l'échelle nationale et à informer les instances de pilotage stratégique.

INTRODUCTION

1.Panorama des données d'imagerie médicale disponibles pour la recherche

Les données d'imagerie médicale sont des données de santé au sens du règlement général sur la protection des données (RGPD), article 4, point 15 (« données à caractère personnel relatives à la santé physique ou mentale d'une personne physique, y compris la prestation de services de soins de santé, qui révèlent des informations sur l'état de santé de cette personne »). Ces données d'imagerie médicale constituent aujourd'hui une source majeure d'information pour la recherche médicale. Les données d'imagerie utilisables à des fins de recherche peuvent être séparées en deux catégories :

1.1. Les données spécifiquement acquises pour la recherche qui ne constituent en réalité qu'une fraction faible du volume total des images produites quotidiennement dans les établissements de santé. Il s'agit notamment des études prospectives incluant de l'imagerie et entrant dans les cadres réglementaires français et européens : loi Jardé (RIPH 1, 2 et 3), règlement européen UE 536/2014 relatif aux essais cliniques de médicaments et règlement européen UE 745/2017 relatif aux dispositifs médicaux. On peut aussi citer les **cohortes** qui définissent des groupes de sujets qui sont suivis individuellement, de manière longitudinale, selon un protocole de recherche préétabli. Les **registres** sont quant à eux définis comme un recueil continu et exhaustif de données pseudonymisées ou non pseudonymisées, intéressant un ou plusieurs événements de santé dans une population définie, le plus souvent selon un critère géographique, à des fins de recherche et de santé publique. Dans ce cadre, les sujets ont donné leur consentement éclairé à l'utilisation, voir à la réutilisation, de leurs données. Les données sont acquises avec un protocole bien défini à l'avance pour assurer un recueil de qualité contrôlée et uniforme dans un cadre mono ou multicentrique.

1.2 Les données acquises à des fins de prise en charge clinique des patients qui constituent le cadre de leur utilisation primaire. L'utilisation primaire fait ainsi référence à l'usage pour la prise en charge médicale des patients. Ces données constituent la très grande majorité des données produites quotidiennement dans les établissements de santé. Une fois pseudonymisées et structurées, elles présentent un potentiel considérable de réutilisation secondaire. Ainsi, l'agence du numérique en santé a défini **l'utilisation secondaire ou réutilisation des données de santé** comme le « traitement ultérieur de données de santé électroniques collectées initialement à d'autres fins pour : des statistiques publiques dans le secteur de la santé ; des activités d'intérêt public dans le domaine de la santé, telles que la protection contre les menaces transfrontalières ou la surveillance de la santé publique ; la **recherche scientifique** ayant trait aux secteurs de la santé ; des activités de développement et d'innovation de produits ou services contribuant à la santé publique ou à la sécurité des soins de santé, des médicaments ou des dispositifs médicaux ou des activités d'éducation dans le secteur de la santé [1]». L'utilisation secondaire des données de santé est donc extrêmement riche ; par exemple pour la surveillance sanitaire (exemple du Mediator), pour piloter certaines crises (remontée des données pendant la pandémie Covid), pour améliorer les soins à travers des analyses de recherche dédiées, pour permettre le développement d'algorithmes d'intelligence artificielle dont les débouchés sont multiples y compris la création de bras « virtuels » dans des essais thérapeutiques ou la création de jumeaux numériques. Le recueil étant non uniformisé, une étape de contrôle qualité et de curation est néanmoins indispensable pour cette utilisation secondaire et l'agrégation de différentes sources.

L'utilisation secondaire des données issues du soin repose sur un cadre réglementaire désormais bien établi. Toute réutilisation doit respecter les principes d'information et de **non-opposition des patients**, et s'inscrire dans le cadre d'une **méthodologie de référence** (MR-004 de la CNIL). Les projets doivent être référencés sur le **portail de transparence du responsable du traitement**

[1] https://participez.esante.gouv.fr/project/chapitre-4-utilisation-secondaire-des-donnees-de-santeelectroniques/presentation/presentation.

(généralement le centre dans lequel les soins ont été dispensés), permettant d'informer les patients sur les finalités des traitements de données et les modalités d'exercice de leurs droits. Enfin, le responsable de traitement doit garantir un **accès sécurisé**, une **traçabilité** et un **suivi des utilisateurs dans le respect du RGPD**. Cette utilisation étant encadrée par une méthodologie de référence de la CNIL, il n'est pas nécessaire de solliciter le comité éthique national CESREES (Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé). A l'échelle locale, des établissements peuvent néanmoins instaurer une instruction des demandes par un comité éthique local afin de s'assurer que la demande est bien conforme à la MR-004 et d'en évaluer la pertinence éthique. Par ailleurs, de plus en plus de revues scientifiques exigent l'avis d'un tel comité (équivalent d'un IRB, Institutional Review Board) avant publication des résultats. Il s'agit donc d'une bonne pratique même si elle n'est pas requise réglementairement par la CNIL.

Dans cette perspective, les **entrepôts de données de santé** (EDS) se sont progressivement déployés au sein des centres hospitaliers. Ces plateformes visent à regrouper, pseudonymiser et structurer les informations issues des dossiers médicaux électroniques, afin de permettre des requêtes pour la réutilisation secondaire de ces données du soin à des fins de recherche tout en préservant la confidentialité. Il se développe également des réseaux d'entrepôts à l'échelle régionale ou interrégionale (voire nationale dans certains domaines comme la cancérologie) afin de mutualiser les expertises et d'harmoniser les données.

La création d'un Health Data Hub (HDH) en 2019 est un autre élément majeur de structuration visant, à terme, à héberger et mettre à disposition les données du système national des données de santé : SNDS (données médico-administratives de l'assurance maladie et les données de facturation des établissements de santé) et à centraliser un **catalogue** référençant les bases de données à l'échelle nationale (registre de toutes études sous MR-004 : https://www.health-data-hub.fr/depot). Dans la future réglementation Européenne, il sera le point d'entrée Français pour le partage des données à l'échelle Européenne à travers le projet HealthData@EU. Il assurera la constitution d'un catalogue au vrai sens du terme, qui servira de référencement des bases, notamment pour leur permettre de gagner en visibilité, avec une description

claire du contenu et de la structuration de leurs données, mais sans transfert systématique de copies de ces bases sur la plateforme du HDH. Il pourra aussi, à terme, alimenter un portail de transparence centralisé pour porter à la connaissance du public les travaux réutilisant des données tel que défini par les articles 13 et 14 du RGPD.

Dans ce contexte, et malgré leur intérêt majeur, les données d'imagerie restent encore imparfaitement intégrées dans les EDS. Actuellement, les EDS intègrent souvent les comptes rendus radiologiques qui sont déjà une source riche d'information sur le type d'examen réalisé, les conditions techniques (par exemple, avec ou sans injection de produit de contraste) et les résultats via l'interprétation du médecin radiologue ou du médecin nucléaire. Néanmoins, de nombreuses analyses de recherche nécessitent de revenir aux données images pour extraire des informations non présentes dans le compte-rendu (par exemple des contours lésionnels) ou pour entrainer directement des algorithmes d'apprentissage machine (deep learning). L'accès aux images du soin à des fins de recherche est encore très restreint dans les EDS ce qui constitue un enjeu pour les années à venir.

2. Place et spécificités des données d'imagerie dans les bases de données françaises

Les données d'imagerie se distinguent des données cliniques ou biologiques par des spécificités propres. Les données d'imagerie ne sont bien sûr pas intégrées dans le SNDS qui ne contient que des données médico-administratives de l'assurance maladie et les données de facturation des établissements de santé. Étant donné qu'il s'agit de données complexes et plus volumineuses que les données issues des autres systèmes d'information, elles ne sont encore que très rarement intégrées dans les EDS. Leur stockage, leur partage et leur utilisation sont des défis dans les cohortes et les registres nécessitant des flux spécifiques et distincts, mais interconnectés avec les données cliniques. Les spécificités des données d'imagerie (Figure 1) portent ainsi sur :

- Le format propre des images (format répondant à la norme internationale DICOM) et la nécessité de conversion vers des formats secondaires pour des opérations de post-traitement tout en suivant des standards et des normes d'interopérabilité.

- Les procédures nécessaires pour empêcher l'identification directe des patients (pseudonymisation) selon les critères de la CNIL et le RGPD. La pseudonymisation des images nécessite de remplacer les données directement identifiantes dans les champs textes associés aux images (entête DICOM) mais aussi parfois directement au sein de l'image sans en dégrader la qualité. Certains cas d'export et de partage plus large peuvent même nécessiter des procédures plus complexes telles que l'effacement du visage sur les imageries cérébrales.

- La manière dont les différents types d'images doivent être structurés en considérant que chaque examen est constitué de multiples séries / séquences qui doivent suivre une nomenclature spécifique pour être requêtables et interopérables. Il faut aussi considérer la genèse de nombreuses métadonnées issues des calculs sur l'image constituant des étapes intermédiaires ou des données annotées, corrigées et enrichies qu'il faut savoir réassocier aux images natives pour enrichir de façon incrémentale la base de données sans avoir à répéter systématiquement des analyses identiques d'une étude à l'autre.

- Les espaces nécessaires pour leur stockage inhérents au poids des données d'imagerie.

- Les modes d'échanges à mettre en place entre un entrepôt de stockage et des espaces sécurisés pour conduire des calculs et des post-traitements nécessitant l'accès à différents outils, langages informatiques et ressources computationnelles lourdes dans le cas d'entrainement de réseaux de neurones profonds.

- La nécessité d'un modèle de valorisation et de tarification lisible proposant une grille de tarif unique pour les coûts d'extraction et de mise à disposition des données d'imagerie, les coûts liés au processus de pseudonymisation des données d'imagerie, les coûts d'accès à distance à un environnement sécurisé, ainsi que les coûts en lien avec l'annotation des images. Cela implique également l'élaboration de contrats-types spécifiques pour l'accès aux données d'imagerie (à l'instar de la convention unique en matière de recherche clinique),

- La nécessité d'une définition de lignes directrices visant à garantir la reconnaissance scientifique du travail des équipes de radiologie qui ont œuvré à la constitution des données d'imagerie.

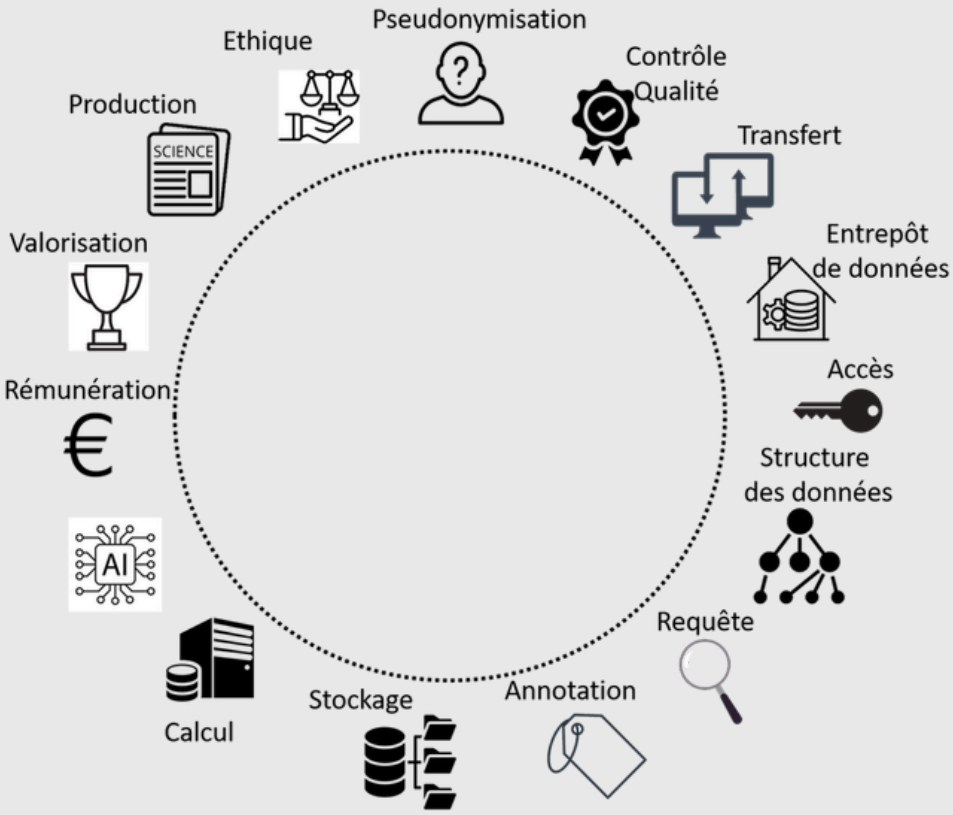


Figure 1 : Représentation schématique des différentes spécificités à prendre en compte dans le cadre de la réutilisation des données d'imagerie du soin pour la recherche.

Nous avons constitué un groupe de travail national sous l'égide du Collège des Enseignants en Radiologie de France (CERF) regroupant des représentants de la Société Française de Radiologie (SFR, groupe recherche), de France Life Imaging, et de façon plus large des centres hospitaliers et plateformes recherches sur le territoire Français. Le groupe s'est réuni régulièrement, a fait circuler une enquête sur les pratiques actuelles et propose dans ce document de détailler ces spécificités et

d'élaborer des recommandations. L'enjeu est de diffuser ce document qui pourra accompagner les centres et les initiatives nationales s'inscrivant dans une démarche de mise à disposition des données d'imagerie déjà collectées pour la recherche afin de promouvoir une harmonisation des pratiques sur le territoire.

SPECIFICITES ET RECOMMANDATIONS

1. Nature et structuration des données d'imagerie

1.1.Fichiers de données brutes

Les fichiers constructeurs, également appelés fichiers de données brutes (raw data), sont des fichiers spécifiques aux équipements d'imagerie médicale. Ils contiennent des informations non traitées, directement issues des capteurs des machines d'imagerie. Ces fichiers sont souvent propriétaires et spécifiques à chaque constructeur d'équipements médicaux. Chaque constructeur utilise ses propres formats propriétaires pour stocker ces données brutes (ex.: .dat pour Siemens, .sdat ou par/rec pour Philips, P-File pour General Electric, RunXXXX pour Canon, Jcamm pour Brucker). Des alternatives (ISMRM Raw Data Format) visent aussi à harmoniser les différents formats de données brutes. Ces fichiers contiennent des informations détaillées collectées par les capteurs, avant conversion en images standardisées. Cela peut inclure des signaux bruts, des données de calibration, et d'autres métadonnées spécifiques à l'équipement.
En dehors de situations spécifiques (travaux de recherche portant sur la reconstruction d'images), les fichiers de données brutes ne sont pas stockés de manière pérenne du fait de leur important volume et de la faible valeur ajoutée pour la majorité des projets de recherche. Pour certains types de données (ex. l'IRM, où la volumétrie des données même brute est typiquement plus restreinte), le stockage de ces données brutes peut parfois être possible voire nécessaire pour une utilisation avancée ultérieure.

Recommandation : Ne pas stocker systématiquement les données brutes. Si on anticipe qu'une certaine catégorie d'images collectées dans le soin pourrait nécessiter l'utilisation de données brutes en vue d'une analyse, il faudra stocker uniquement et spécifiquement ces données brutes. Ces situations sont peu fréquentes.

1.2. Fichiers DICOM

- **Définition de la norme DICOM**

La norme DICOM (Digital Imaging and Communications in Medicine) est un standard international utilisé pour la gestion, le stockage, l'impression et la transmission des informations d'imagerie médicale. Cette norme permet d'assurer l'interopérabilité entre les différents systèmes d'imagerie médicale, notamment entre les modalités d'imagerie (IRM, scanner, radiographie, échographie) et les sites de stockage des images (Picture Archiving and Communication System, PACS et Vendor Neutral Archives, VNA). Les images d'échographie peuvent être au format DICOM mais peuvent présenter des spécificités (cf. au-dessous).

- **Anatomie d'un fichier d'imagerie DICOM**

Un fichier conforme à la norme DICOM contient à la fois les données proprement dites (images ou autres objets) et l'ensemble des métadonnées qui les décrivent. Il est structuré autour de plusieurs composants essentiels :

- **Un en-tête (header)**

L'en-tête précise les informations générales sur le format du fichier et les conditions techniques de son encodage. Il contient un ensemble structuré d'attributs, appelés éléments DICOM (ou tags), qui regroupent les métadonnées associées au fichier. Ces éléments peuvent inclure des informations relatives à l'identité du patient (nom, identifiant, date de naissance), aux conditions de l'examen (date et heure de réalisation, modalité, protocole, région anatomique, lieu de réalisation), ainsi qu'aux paramètres techniques d'acquisition et d'affichage (type d'équipement, réglages utilisés, paramètres de reconstruction, paramètres physiologiques, paramètres d'affichage).
La norme DICOM définit certains de ces éléments comme obligatoires et d'autres comme optionnels, afin d'assurer à la fois l'interopérabilité minimale entre systèmes et la flexibilité nécessaire à la diversité des modalités et des usages.

- **Des identifiants uniques (Unique Identifiers, UIDs)**

Les fichiers DICOM utilisent des identifiants uniques pour classer et identifier sans ambiguïté les différents objets qu'ils encodent, qu'il s'agisse d'images, de séries, d'études ou d'objets non-image. Ces identifiants prennent la forme de suites numériques séparées par des points et suivent une hiérarchie normalisée, débutant par un identifiant racine attribué par une autorité de référence. Les identifiants dérivés créés selon cette norme garantissent l'unicité des objets et évitent toute collision entre systèmes, établissements ou institutions. Les groupes sont structurés de manière logique, avec par exemple les identifiants commençant par 0010 dédiés aux informations d'identification propres au patient.

- **Le contenu binaire du fichier DICOM**

Le contenu binaire dépend du type d'objet DICOM représenté. La norme DICOM ne se limite pas au stockage d'images et permet également d'encoder de nombreux types de données associées à un examen d'imagerie :

  - **Objets image**

Lorsque le fichier encode une image, les données correspondent à la matrice de pixels (en 2D) ou de voxels (en 3D ou plus). Ces données peuvent être stockées sans compression ou avec une compression avec ou sans perte, par exemple JPEG, JPEG2000 ou RLE pour les images fixes, et MPEG-2 ou MPEG-4 pour les séquences vidéo. Ces objets constituent les données sources de l'imagerie clinique et de la majorité des projets de recherche.

  - **Objets de rapport et de données structurées (DICOM SR)**

La norme DICOM permet également de stocker des informations non-image sous forme d'objets de Structured Reporting (DICOM SR). Ces fichiers contiennent des données textuelles et numériques organisées de manière hiérarchique, pouvant référencer explicitement des images ou des séries associées. Ils sont utilisés notamment pour stocker des comptes rendus d'imagerie, des informations de dosimétrie ou des résultats de mesures quantitatives. Ces objets contiennent fréquemment des informations directement ou indirectement identifiantes et nécessitent donc

une attention particulière lors des procédures de pseudonymisation.

- **Objets de segmentation (DICOM SEG)**

La norme DICOM permet aussi de stocker des objets de segmentation, qui référencent les objets sources.

- **Objets de radiothérapie et de géométrie (DICOM RT)**

Les modules DICOM dédiés à la radiothérapie permettent de stocker des informations géométriques complexes, telles que des contours d'organes ou de lésions (DICOM RT-STRUCT), des plans de traitement ou des distributions de dose. Bien que développés initialement pour la radiothérapie, ces objets sont fréquemment réutilisés en recherche pour stocker des segmentations ou des annotations issues de logiciels de post-traitement avancé. Ils constituent des données dérivées étroitement liées aux images sources.

- **Objets de signaux physiologiques (Waveform)**

Certaines données physiologiques associées à l'imagerie (par exemple électrocardiogrammes, signaux respiratoires, électroencéphalogramme...) peuvent être encodées sous forme d'objets DICOM Waveform. Ces objets représentent des signaux échantillonnés, avec leurs fréquences, unités, canaux et métadonnées associées, et peuvent être liés à des examens d'imagerie synchronisés.

- **<u>Hétérogénéité des implémentations de la norme DICOM</u>**

L'existence de la norme DICOM ne garantit pas une utilisation homogène de celle-ci par l'ensemble des équipements et des logiciels qui génèrent ou manipulent des fichiers DICOM. La norme définit un cadre commun, mais laisse une latitude importante aux fabricants dans le choix des types d'objets utilisés et des éléments DICOM effectivement renseignés. L'utilisation de balises supplémentaires, dites "privées", est prévue dans la norme, et est parfois utilisée par les

différents constructeurs pour stocker des métadonnées de manière non standardisée et pouvant varier y compris entre différente générations de machines d'un même constructeur.

Un exemple courant concerne aussi les rapports cliniques, qui peuvent être encodés soit sous forme d'objets structurés (DICOM SR), soit sous forme d'images. Dans ce dernier cas, le rapport textuel est converti en image et peut être visualisé à l'aide des mêmes outils que ceux utilisés pour l'affichage des images médicales. Certains fabricants font ce choix afin d'éviter la dépendance à des mécanismes de décodage supplémentaires (par exemple pour l'affichage de documents PDF) et de garantir une compatibilité maximale avec les stations de lecture existantes. Pour des raisons historiques liées à l'évolution progressive du standard DICOM, à l'intégration successive de nouveaux types de contenus dans l'industrie biomédicale, et à la nécessité de maintenir une rétrocompatibilité avec des équipements déjà déployés, différentes combinaisons d'éléments DICOM, de types d'objets et de contenus binaires coexistent aujourd'hui au sein des systèmes d'imagerie.

Cette hétérogénéité implique, dans un contexte de réutilisation secondaire des données à des fins de recherche, de bien connaître les variantes possibles de structuration des fichiers DICOM. Elle est particulièrement critique pour garantir une pseudonymisation correcte et complète, notamment pour l'identification des objets non-image, des rapports encodés sous forme d'images et des champs susceptibles de contenir des informations directement ou indirectement identifiantes.

- **Protocole de transfert DICOM**

La norme DICOM a été conçue pour assurer l'échange, la traçabilité et la cohérence des données d'imagerie médicale au sein de systèmes distribués. Chaque fichier DICOM intègre des métadonnées normalisées permettant l'identification fiable du patient, de l'étude, de la série et de l'instance, facilitant ainsi le routage automatique des images vers les systèmes cibles appropriés, y compris lors de transferts entre dispositifs, services ou établissements.

L'architecture DICOM repose sur un modèle client-serveur. Les modalités d'acquisition (scanner, IRM, TEP, échographes, etc.) agissent généralement comme clients, tandis que les systèmes d'archivage (PACS, VNA) assurent les rôles de serveurs, en centralisant le stockage, l'indexation et la gestion des images et des objets associés. Ce modèle permet l'interconnexion de multiples dispositifs à un ou plusieurs systèmes de stockage, garantissant l'interopérabilité et l'accès sécurisé aux images au sein d'un réseau hospitalier ou dans des architectures multi-établissements.

Les échanges DICOM s'appuient sur des services réseau standardisés, historiquement basés sur TCP/IP, tels que C-STORE pour le stockage, C-MOVE et C-GET pour le transfert, et plus récemment sur des interfaces web (DICOMweb : WADO-RS, QIDO-RS, STOW-RS), facilitant l'intégration avec des architectures modernes orientées API.

Dans une telle architecture distribuée, plusieurs nœuds peuvent héberger des copies complètes ou partielles des données (réplications locales, caches, archives de secours, serveurs intermédiaires de transfert ou d'extraction). La sécurisation de l'ensemble de ces nœuds constitue une exigence essentielle pour la protection de la confidentialité des données. En effet, la coexistence de différentes copies ou versions dérivées d'un même objet DICOM peut introduire un risque de corrélation involontaire entre jeux de données provenant de sources distinctes.

À titre d'exemple, un chercheur peut recevoir un ensemble d'images DICOM pseudonymisées dans le cadre d'un projet de recherche. Les champs directement identifiants du patient et de l'établissement ont été modifiés ou supprimés, mais les identifiants uniques des objets (par exemple les SOP Instance UID, Series Instance UID ou Study Instance UID) ont été conservés afin de préserver la structure interne des données. Si ce même chercheur dispose par ailleurs d'un autre jeu de données contenant ces mêmes identifiants uniques associés à des identités réelles, une réidentification indirecte devient possible par simple correspondance des UIDs.

Cette situation illustre l'importance, dans un contexte de réutilisation secondaire des données, de maîtriser finement les mécanismes de transfert et de réplication des fichiers DICOM, et de contrôler le traitement des identifiants uniques. Lors des procédures de pseudonymisation,

il est indispensable d'évaluer la nécessité de conserver ou de régénérer les UIDs, afin de préserver la cohérence interne des données tout en empêchant toute corrélation entre jeux de données issus de systèmes ou de contextes distincts.

- **Recommandations en lien avec les fichiers DICOM**

Les fichiers DICOM représentent le format original d'imagerie pour le soin et pour l'interprétation radiologique. Ce format est également massivement utilisé dans les projets de recherche et considéré comme les données images « sources » non modifiées qui sortent des machines et qui sont à conserver.

Recommandation : Lors de la constitution d'une base de données primaire, **s'assurer que les fichiers DICOM soient conservés sur les systèmes d'information clinique dans leur forme originale sans altération, afin de garantir l'exactitude et la fiabilité des informations qu'ils contiennent.** Nous suggérons également, quand cela est possible, **d'éviter le stockage d'images compressées via des algorithmes « avec perte »** (type JPEG). Les procédures de pseudonymisation (cf. section 2) devront respecter cette forme originale et établir une copie altérée de ces fichiers sources lors de l'export vers des environnements recherche dédiés (cf. section 3) pour réutilisation secondaire.

- **Spécificités des données d'échographie**

L'échographie constitue un cas particulier en matière de format d'image. Selon les constructeurs et les paramètres d'export, les acquisitions peuvent être enregistrées sous forme d'images fixes ou de boucles dans un format conforme à la norme DICOM, mais également sous forme de fichiers vidéo non encapsulés (AVI, MPEG-4…) ou de formats propriétaires. Lorsque l'export DICOM est disponible, il s'agit du format à privilégier, car il fournit des métadonnées structurées permettant une pseudonymisation complète. En revanche, les vidéos non-DICOM comportent généralement peu de métadonnées et incluent fréquemment des informations sur l'identité du patient incrustées directement dans l'image (pixel-burning), ce qui nécessite un contrôle manuel

renforcé (Cf. section 2). Par ailleurs, l'échographie génère souvent des données fortement compressées, susceptibles d'altérer la qualité de certains traitements de recherche.

Recommandation : Pour les images d'échographie, privilégier l'export DICOM lorsque cela est possible, ou **encapsuler les vidéos dans un conteneur DICOM** afin de garantir la présence de métadonnées exploitables et de faciliter la pseudonymisation. Si seules des vidéos non DICOM sont disponibles, nous recommandons une conversion vers un **format standardisé** (par exemple MPEG-4), associée à une **documentation détaillée** des conditions d'acquisition et à une procédure rigoureuse de **contrôle des images suivi du masquage des zones identifiantes** (cf. section 2). Dans tous les cas, les données pour la recherche doivent être sans compression ou utilisant une compression non destructrice.

1.3. Formats et organisations de fichiers adaptés à la recherche

Les fichiers DICOM, bien que constituant le format standard le plus largement utilisé en environnement clinique, présentent plusieurs limites pour leur utilisation dans un contexte de recherche : la complexité des métadonnées qu'ils contiennent et les difficultés à les pseudonymiser, la multiplicité des fichiers, et leur volume. De plus, de nombreux outils d'analyse d'images utilisés en recherche n'utilisent pas le DICOM en format d'entrée. Afin de répondre à ces différentes limites, des formats de fichiers plus compacts ont été développés, permettant de répondre aux contraintes différentes d'un environnement de recherche. Pour les données en 2 dimensions comme les radiographies, des formats de fichiers standards (TIFF, PNG, JPG ou plutôt LZW pour une compression sans perte) peuvent être utilisés. Pour les données de plus grandes dimensions (3, 4 ou 5 dimensions), le format NifTI (Neuroimaging Informatics Technology Initiative), initialement développé pour la recherche en neuroimagerie, s'est imposé progressivement comme un format image secondaire, particulièrement utile pour mener des analyses dédiées à la recherche, alors même qu'une copie au format DICOM reste la référence pour le stockage des données sources.

- **Formats NIfTI**

Un fichier NifTI contient généralement l'ensemble d'une série radiologique, que celle-ci soit en 2 dimensions (par exemple, radiographie), ou le plus souvent en 3 dimensions (par exemple, acquisition scanner), 4 dimensions (par exemple, imagerie 3D dynamique) voire plus (par exemple, imagerie de diffusion en IRM). Le format de fichier NifTI-1 est composé de deux parties :

- Un en-tête (header) contenant des informations sur le format de stockage des données, le nombre de dimensions du fichier et la taille de chacune de ces dimensions, la taille des voxels, la position et l'orientation du volume dans l'espace.
- Des données d'image, stockées de manière contiguë selon les spécifications définies dans l'en-tête du fichier.

Le format NifTI stocke donc principalement le volume image et seulement quelques métadonnées minimales nécessaires à l'analyse scientifique (informations géométriques).

- **Organisation des fichiers (BIDS)**

Contrairement aux fichiers DICOM où toutes les métadonnées, sont stockées dans le header du fichier DICOM, le format NifTI-1 ne contient que des informations dimensionnelles et géométriques sans métadonnées supplémentaires. Des extensions de ce format de fichier (NifTI-2) permettent de stocker des métadonnées plus complexes. Cependant, cette stratégie est peu utilisée en recherche car les métadonnées sont le plus souvent séparées des données brutes afin de faciliter leur organisation et d'éviter leur redondance. La stratégie la plus largement utilisée est de séparer physiquement les métadonnées des données images. Afin de structurer au mieux les imageries et leurs métadonnées, un standard permettant leur organisation a été conçu.

Le standard BIDS (Brain Imaging Data Structure, https://bids.neuroimaging.io/) définit une structure de répertoires et de fichiers pour stocker et partager des données d'imagerie de manière cohérente et organisée. Cette structuration permet d'organiser les différents sujets

d'une étude, les sessions d'acquisition, les types d'imagerie réalisées, et leurs métadonnées. Cette structuration permet de simplifier la collaboration et l'échange, et de conserver également les données dérivées de l'analyse des images (cf. plus bas, gestion des métadonnées).

Recommandation : Pour un projet de recherche donné, nous recommandons de prévoir, au sein de l'environnement recherche, une copie DICOM pseudonymisée + une éventuelle copie convertie dans un **format open-source adapté aux analyses pour la recherche de type NifTI** si cela est requis pour le projet considéré. Nous recommandons également de **séparer les données images des métadonnées pertinentes à la recherche en utilisant des méthodes d'organisation (de type BIDS) documentées par le responsable de l'export des données.**

2. Enjeux liés à la pseudonymisation et à la confidentialité.

2.1. Principes généraux de pseudonymisation des données d'imagerie

La pseudonymisation constitue une étape essentielle dans le processus de réutilisation des données d'imagerie issues du soin à des fins de recherche. Elle vise à garantir la confidentialité des personnes tout en permettant la traçabilité et la réutilisation scientifique des données. Conformément au règlement général sur la protection des données (RGPD), les données de santé appartiennent à la catégorie des données sensibles et leur traitement n'est autorisé qu'à des conditions strictes de sécurité et de proportionnalité.

Contrairement à l'anonymisation, la pseudonymisation ne supprime pas tout lien potentiel avec l'identité d'une personne : elle consiste à remplacer les identifiants directs par des codes pseudonymes, tout en conservant la possibilité d'une ré-identification encadrée et justifiée, par exemple dans le cadre d'un audit ou d'une actualisation des données.

Les définitions données par la CNIL[1] sont les suivantes (Tableau 1) :

PROCESSUS	PSEUDONYMISATION	ANONYMISATION
STATUT DES DONNÉES	Personnelles (restent indirectement identifiantes et donc soumises au RGPD et à la loi Informatique et Libertés)	Anonymes
RÉUTILISATION DES DONNÉES	Sous conditions	Sans restriction
UTILITÉ DES DONNÉES	Préservée car pas d'altération du niveau de détail des données	Plus ou moins altérée en fonction des objectifs poursuivis et des méthodes appliquées
MÉTHODES À METTRE EN ŒUVRE	Compteur, générateur de nombres aléatoires, décalage des dates d'examens, fonction de hachage, chiffrement à clé secrète, etc.	Dépend des objectifs poursuivis : confidentialité différentielle, randomisation, k-anonymat, l-diversité, t-proximité, etc.
COMPLEXITÉ DE LA MISE EN ŒUVRE	Simple à moyenne	Dépend des objectifs poursuivis : simple dans certains cas comme l'agrégation ou le comptage et complexe dans d'autres

Tableau 1 : Principe de l'anonymisation et de la pseudonymisation

- **Anonymisation** : processus de traitement des données irréversible impliquant la suppression d'éléments suffisants afin qu'une personne physique ne puisse plus être identifiée en cas d'utilisation de « tous les moyens raisonnablement susceptibles d'être utilisés », prenant en compte les coûts, le temps et la disponibilité des technologies. Le processus d'anonymisation élimine donc toute possibilité de ré-identification des individus que ce soit par individualisation, corrélation ou inférence.

- **Dé-identification** (version américaine de la suppression des données identifiantes qui se rapproche de la définition RGPD de la pseudonymisation) : contrairement à l'anonymisation, la dé-identification permet la ré-identification sous certaines conditions, notamment lorsque l'entité responsable attribue un code unique aux données dé-identifiées permettant de les réassocier à une personne.

[2] https://www.cnil.fr/fr/recherche-scientifique-hors-sante-enjeux-et-avantages-de-lanonymisation-et-de-la-pseudonymisation

- **Pseudonymisation** : processus où les données personnelles ne peuvent plus être attribuées à un individu sans informations supplémentaires. C'est la solution recommandée pour les données d'imagerie utilisées à des fins de recherche.

Le processus de pseudonymisation doit être mis en œuvre selon un protocole défini et validé par le responsable de traitement après avis du délégué à la protection des données (DPO) et, le cas échéant, par le comité d'éthique. Une analyse des risques de réidentification des patients peut également être menée, conformément aux recommandations de la CNIL.
La pseudonymisation des données d'imagerie est complexe et ne peut pas se limiter à une simple suppression d'informations. Elle nécessite une compréhension fine des formats d'image (DICOM, NIfTI ; cf. au-dessus), et une adaptation des outils utilisés à l'objectif scientifique poursuivi. Enfin, il est essentiel d'assurer une traçabilité complète des opérations (scripts utilisés, versions logicielles, date d'exécution) afin de garantir la transparence et la reproductibilité du traitement.

Au-delà du strict respect des obligations légales, la pseudonymisation et la gestion rigoureuse de la confidentialité représentent un enjeu éthique et de confiance : elles conditionnent l'acceptabilité sociale de la réutilisation des données de santé et la légitimité des projets de recherche qui en dépendent.

2.2 Recommandations pour pseudonymiser les données DICOM

Les fichiers DICOM contiennent à la fois les données d'image et un grand nombre de métadonnées administratives et techniques susceptibles d'identifier le patient, le lieu voire le personnel de soin. Leur pseudonymisation repose donc sur deux niveaux d'intervention complémentaires : le traitement de l'en-tête DICOM pour les métadonnées, et le contrôle du contenu pixel pour les images.

- **Pseudonymisation de l'en-tête DICOM**

L'en-tête DICOM regroupe plusieurs centaines de champs, dont certains contiennent des informations nominatives ou indirectement identifiantes.

Recommandation : Modifier ou remplacer les champs DICOM, plutôt que de les supprimer, afin de ne pas altérer la compatibilité technique des fichiers avec les outils d'analyse.

Les **champs à traiter impérativement** incluent :

- PatientName (0010,0010) : à supprimer ou remplacer selon la pratique ;
- PatientID (0010,00120) : à remplacer par un identifiant pseudonyme ou à remplacer par une chaine vide;
- PatientBirthDate (0010,0030) : à modifier (par troncature ou décalage temporel);
- AccessionNumber (0008,0050) : à modifier ou supprimer ;
- StudyID (0020,0010) : à modifier ou supprimer ;
- StudyInstanceUID (0020,000D), SeriesInstanceUID (0020,000E) et SOPInstanceUID (0008,0018) : à régénérer automatiquement pour rompre tout lien avec les séries sources ;
- FrameOfReferenceUID (0020,0052) : à adapter avec prudence ; il peut être remplacé pour rompre tout lien avec les données sources, mais doit rester cohérent entre les séries appartenant à un même espace de référence afin de préserver l'intégrité structurelle des volumes.
- Les informations identifiant l'établissement ou les professionnels (InstitutionName - 0008,0080 ; InstitutionAddress - 0008,0081 ; ReferringPhysicianName - 0008,0090 ; PerformingPhysicianName -0008,1050 ; OperatorsName - 0008,1070; StationName -0008,1010; DeviceSerialNumber - 0018,1000) sont également hautement identifiantes dans le RGPD : à modifier ou supprimer.

Il convient également de traiter les champs contenant les dates d'acquisition (StudyDate - 0008,0020 ; SeriesDate - 0008,0021 ; AcquisitionDate - 0008,0022 ; ContentDate - 0008,0023). Ces dates doivent être modifiées mais de façon relative c'est-à-dire en conservant les intervalles temporels mais en ne gardant pas la date réelle pour éviter la ré-identification directe.

Une attention particulière doit être portée aux champs privés ou private tags (commençant par un nombre impair, par exemple 0009,xxxx), souvent utilisés par les constructeurs pour stocker des informations internes, dont certaines peuvent contenir des données identifiantes. Leur vérification est donc indispensable. Ils peuvent contenir des **informations importantes pour les analyses d'images** et nous recommandons donc de ne pas les supprimer de façon systématique.

De même, certains champs techniques doivent être conservés de façon impérative pour garantir la reproductibilité des analyses ou le contrôle des biais: les paramètres d'acquisition (constructeur, protocole, séquence, temps d'acquisition, orientation) ou les informations de calibration ne doivent pas être altérés, sous peine de compromettre la validité scientifique des résultats.

Cette étape doit donc assurer l'équilibre entre la protection des données personnelles et la préservation de la valeur scientifique.

Enfin, le processus de pseudonymisation doit être entièrement documenté : la liste des champs modifiés ou remplacés, la méthode de génération des identifiants pseudonymes et les outils utilisés (par exemple : pydicom, dicom-anonymizer, DICOMCleaner) doivent être archivés et rattachés au protocole de recherche.

- **Contrôle du contenu pixel (Pixel burning)**

Certaines modalités d'imagerie, notamment l'échographie, les captures vidéo, les planches avec des images clés, les captures d'écran, peuvent comporter des informations textuelles incrustées directement dans les pixels de l'image (noms, numéros de patient, dates d'examen). Ce phénomène, appelé pixel burning, constitue un risque majeur pour la confidentialité. Le champ DICOM BurnedInAnnotation (0028,0301) peut donner des orientations.

Recommandation : Contrôle le contenu pixel en combinant :

- une détection automatique par reconnaissance de texte (OCR : Optical Character Recognition) ;
- un contrôle visuel sélectif sur les séries identifiées comme à risque ;
- un effacement ou masquage précis des zones concernées, suivi d'une vérification visuelle.

Cette étape doit être intégrée au protocole de pseudonymisation, avec traçabilité des outils et validation par le responsable scientifique du projet.

2.3. Recommandations pour les autres formats d'image

Le format NIfTI

Le format NIfTI, très utilisé pour l'analyse en recherche, ne conserve qu'un sous-ensemble limité des métadonnées issues du DICOM. La conversion de DICOM en NifTI constitue souvent une étape de pseudonymisation effective, car les identifiants du patient ne sont pas exportés. Le processus est d'autant plus sûr quand une copie NifTI est créée dans un second temps à partir des DICOM pseudonymisés.

Cependant, des précautions demeurent nécessaires :

- vérifier que les champs de l'en-tête NIfTI (descrip, aux_file, etc.) ne contiennent pas d'informations identifiantes ;
- contrôler la structure du nom des fichiers et des dossiers, qui peut véhiculer des identifiants implicites. Cette précaution est, bien sûr, également vraie pour le nom des fichiers/dossiers contenant des images DICOM.

Plusieurs outils open source permettent la conversion des données DICOM issues du soin vers le format NIfTI, utilisé pour les analyses de recherche. Parmi eux, dcm2niix s'impose comme la référence internationale que nous recommandons, offrant une conversion fiable, rapide et compatible avec les principaux constructeurs d'équipements. Des outils complémentaires, tels que HeuDiConv pour la structuration BIDS ou dicom2nifti pour les

pipelines Python, facilitent l'intégration de cette étape dans les workflows de pseudonymisation et d'analyse.

- **Autres formats spécifiques**

Pour les données vidéo ou doppler, le risque de fuite d'informations identifiantes via le signal image ou sonore nécessite un contrôle manuel renforcé. L'utilisation de scripts automatisés (OpenCV, ffmpeg, OCR intégré) peut faciliter la détection des zones à masquer, mais une revue humaine finale reste obligatoire avant tout transfert.

Recommandation : Lors de l'utilisation de fichiers non-DICOM, s'assurer que les métadonnées sont non-identifiantes et contrôler la structure du nom des fichiers et des dossiers, ainsi que les images qui peuvent véhiculer des identifiants implicites. Nous recommandons aussi une revue humaine de ces cas spécifiques.

2.4 Cas particulier de la neuroimagerie (défacialisation)

Dans le domaine de la neuroimagerie, les acquisitions scanner ou IRM anatomiques 3D isotropiques permettent une reconstruction du visage du sujet, rendant possible une ré-identification. Il existe des outils qui suppriment les structures faciales tout en préservant les zones cérébrales utiles à l'analyse. Ces outils sont relativement performants pour les acquisitions IRM 3D T1 (mri_deface (FreeSurfer), fsl_deface (FSL) ou pydeface) mais ne sont pas adaptés aux données scanner, pour lesquelles les méthodes de défacialisation restent encore peu matures.

Plusieurs éléments sont importants à préciser. Les outils actuels ont été conçus pour manipuler des volumes tridimensionnels continus, ce qui est le cas du format NifTI. Par contre les outils actuels ne savent pas gérer les en-têtes DICOM ni la structuration des images en DICOM (multiples coupes 2D) et la défacialisation n'est donc pas mature sur des images DICOM. Certains outils se développent mais ne sont pas encore considérés comme stables ou validés.

Si on considère que la version des images dédiées à la réutilisation pour la recherche doit contenir une copie DICOM pseudonymisée servant de données sources, en plus des images NifTI servant aux analyses d'images, il n'est pas possible de procéder à une défacialisation complète dans un environnement secondaire. La conservation de ces données anatomiques possiblement identifiantes n'est pas prohibée si l'environnement est contrôlé et les risques maitrisés au sens du RGPD (en France, certification Hébergeur de Données de Santé - HDS - avec journalisation, cloisonnement des accès, clé d'appariement sécurisée).

De plus, la défacilaisation est une opération lourde et destructrice d'information anatomique ce qui peut altérer la qualité de certaines analyses (recalage, segmentation anatomique fine...).

Recommandation : A ce jour, nous ne recommandons pas de défacialisation pour une utilisation secondaire d'images cérébrales dans un EDS. La défacialisation sera par contre nécessaire si on envisage une base publique ou semi-publique ce qui nécessitera une anonymisation (et non pas seulement une pseudonymisation) et ce qui pourra se faire uniquement via des images NifTI.

2.5 Données "non image" associées

Comptes rendus et fichiers texte (XML, CSV, JSON, HL7)

Les données textuelles issues des systèmes d'information radiologique (RIS), des dosimétries ou des comptes rendus doivent également faire l'objet d'une pseudonymisation rigoureuse.

Deux approches complémentaires peuvent être utilisées pour la pseudonymisation des données textuelles ou tabulaires :

- L'approche dite **"blacklist"** consiste à **supprimer ou modifier les champs identifiants connus**, tout en conservant l'ensemble des autres informations.

- L'approche dite "whitelist" adopte une logique inverse : elle ne conserve que les champs utiles à la recherche (par exemple la taille de la lésion dans un compte rendu), tous les autres étant supprimés par défaut.

Le choix entre ces deux stratégies dépend du type de données et du niveau de risque. Dans la plupart des cas, la méthode blacklist est privilégiée pour les métadonnées des images DICOM en modifiant/remplaçant les champs sensibles (cf. au-dessus).

Recommandation : Pour les données textuelles associées aux images (comptes rendus, tableaux), nous recommandons l'utilisation d'une whitelist afin de limiter au mieux les risques d'inclusion d'informations nominatives.

Dans le cas de textes libres, l'emploi d'outils de **traitement automatique du langage (NLP)** peut faciliter l'extraction des éléments d'intérêt et la détection d'identifiants personnels (Named-entity recognition). En complément des approches classiques de traitement automatique du langage (NLP), les modèles de grande taille (LLM) peuvent faciliter la détection d'informations identifiantes dans les textes libres, en particulier pour des formulations ambiguës ou non structurées. Toutefois, leur utilisation doit respecter des exigences strictes : les traitements doivent être réalisés **exclusivement dans un environnement local et sécurisé**, sans transfert vers un service externe, et validé par le DPO. En raison du caractère probabiliste des LLM et de l'absence de garantie d'exhaustivité, une **revue humaine** demeure indispensable avant diffusion. Les LLM ne remplacent donc pas les outils déterministes classiques (regex, NER spécialisés), mais peuvent les compléter pour améliorer la détection ou la normalisation des données textuelles.

2.6. Risques spécifiques liés à l'intelligence artificielle générative

La pseudonymisation supprime les éléments directement identifiants (nom, numéro patient, dates…), mais elle ne modifie pas le contenu de l'image elle-même, en dehors de la défacialisation abordée en 2.4. Or, une image médicale contient souvent des caractéristiques propres à la personne : anatomie particulière (y compris forme du crâne ou épaisseur des méninges sur une imagerie cérébrale défacialisée), pathologie rare ou motif unique (lésion, chirurgie, malformation), prothèse identifiable, etc.

Ces éléments constituent une empreinte biométrique pouvant permettre une ré-identification indirecte. En effet, si un modèle reproduit une image trop proche du cas réel, ce cas peut être « reconnaissable » pour quelqu'un qui connait le patient ou possède l'examen source.

Les modèles d'intelligence artificielle générative (modèles de diffusion, GAN…) apprennent à reproduire la structure complète des images sur lesquelles ils sont entraînés. Contrairement aux algorithmes classiques, ils peuvent parfois **mémoriser des exemples individuels** plutôt que seulement des tendances générales. Dans ce cas, un modèle peut :

- générer une image très proche d'un examen réel, même si celui-ci a été pseudonymisé ;
- révéler indirectement des informations sensibles à travers une image synthétique qui ressemble fortement à un patient précis ;
- être vulnérable à des attaques permettant de reconstituer ou reconnaître une image ayant servi à l'entraînement.

Ainsi, même si les données d'entraînement sont pseudonymisées, un modèle génératif peut devenir un vecteur de fuite d'information biométrique, si aucune mesure de protection n'est appliquée.

Recommandation : Procéder à un réexamen spécifique du risque de ré-identification pour tout projet faisant intervenir des modèles génératifs, incluant :
(i) la vérification des procédures de pseudonymisation en amont ; (ii) la limitation stricte des accès aux environnements d'entraînement et aux modèles; (iii) un contrôle de l'absence de reproduction d'images réelles ou d'éléments identifiants lors de tests en sortie de modèle.
Ces mesures complètent les dispositifs classiques de pseudonymisation et de sécurité, et sont indispensables dans un contexte où les capacités de génération et de mémorisation des modèles d'IA évoluent rapidement.

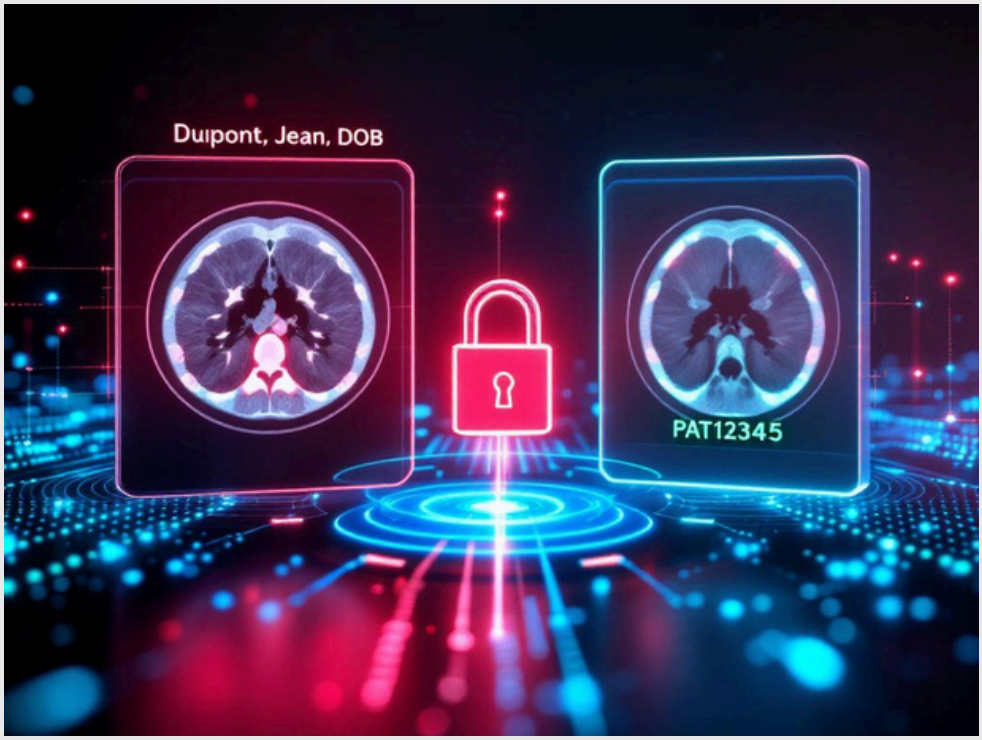
2.7. Bonnes pratiques et traçabilité

La pseudonymisation des données d'imagerie ne constitue pas un acte isolé, mais un **processus gouverné** qui s'inscrit dans une politique institutionnelle de protection des données.

Recommandation : Les bonnes pratiques institutionnelles que nous recommandons sont :

- la documentation détaillée des scripts et des versions logicielles utilisées ;
- l'association en amont du DPO et du RSSI pour la validation de la conformité des procédures de pseudonymisation, et, le cas échéant, l'examen par un comité d'éthique ou scientifique local afin d'évaluer la proportionnalité et la pertinence éthique des traitements ;
- la conservation sécurisée de la clé d'appariement dans un environnement séparé ;
- la réalisation d'audits réguliers de conformité ;

et l'alignement sur les référentiels reconnus : ISO 27001, HDS et guides de la CNIL relatifs à la recherche en santé.



3. Spécificités liées à l'hébergement et aux échanges de données d'imagerie

L'hébergement des données d'imagerie pour la recherche reste marqué par une forte hétérogénéité locale, dépendante des ressources techniques et humaines propres à chaque établissement. Nous recommandons un effort d'harmonisation des structures d'hébergement dans les différents centres tel que décrit ci-dessous.

3.1 Nécessité d'un hébergement adapté à la recherche

Les systèmes cliniques, tels que les **PACS** (Picture Archiving and Communication System), répondent aux besoins du diagnostic et du suivi des patients via l'analyse des images produites dans le soin.

La réutilisation secondaire de ces images à des fins de recherche impose un hébergement distinct, spécifique, et dédié qu'on peut appeler « hébergement secondaire » ou « infrastructure de recherche ». **L'hébergement secondaire** destiné à la recherche en imagerie doit satisfaire à deux exigences complémentaires :

(1) assurer la conformité réglementaire et la sécurité des données (confidentialité, traçabilité, intégrité, gouvernance des accès), conformément aux référentiels ISO 27001 et HDS (norme hébergeur de données de santé). Bien que la recherche ne soit pas encore formellement incluse dans le périmètre de la certification HDS, cette évolution est anticipée avec l'entrée en vigueur du **règlement européen sur l'Espace Européen des Données de Santé** (**EHDS**), qui harmonisera ces exigences à l'échelle européenne.

(2) offrir des capacités opérationnelles adaptées à la recherche (stockage et calcul performants, interopérabilité avec les systèmes hospitaliers, outils de visualisation et de post-traitement des images).

Recommandation : Prévoir une solution d'hébergement secondaire pour la recherche qui offre une articulation équilibrée, garantissant à la fois la protection des données des personnes - en anticipant dès à présent l'application des exigences HDS aux infrastructures de recherche accueillant des données d'imagerie - et l'efficacité pour l'exécution des projets scientifiques.

3.2 Modèle d'architecture globale

L'hébergement des données d'imagerie doit reposer sur une architecture à plusieurs niveaux si on veut inclure une dimension recherche visant à réutiliser les images du soin (Figure 2) :

- L'hébergement primaire correspond à l'environnement clinique (PACS, RIS, DPI) où les images sont produites et archivées pour le soin.

- L'hébergement secondaire correspond à l'environnement de recherche, recevant une copie pseudonymisée des données pour leur exploitation scientifique.
  - D'un point de vue matériel, cet hébergement peut correspondre à des serveurs physiques hébergés au sein des établissements de santé, certifiés HDS conformément aux exigences réglementaires. Il peut également s'agir de solutions externalisées (cloud), à condition qu'elles soient elles aussi certifiées HDS. Certaines plateformes cloud disposent en complément d'une certification SecNumCloud, délivrée par l'ANSSI (Agence Nationale de la Sécurité des Systèmes d'Information), qui renforce les garanties de cybersécurité et de souveraineté et peut être privilégiée pour des usages collaboratifs, nationaux ou européens.

  - D'un point de vue logiciel, il est nécessaire d'avoir des solutions permettant de manipuler les données au sein de cet environnement via des solutions logicielles qui peuvent être académiques type Xnat, Orthanc, Shanoir ou ArchiMed, ou privées type entrepôts de données spécialisés imagerie.

  - Cet hébergeur doit s'intégrer à l'environnement de travail imposant des solutions d'interfaçage (i) d'une part avec le PACS pour assurer les requêtes et (ii) d'autre part avec

des ressources pour faire les analyses de recherche via des API permettant de se connecter et de faire transiter de façon temporaire des données (hébergement temporaire) vers des clusters de calculs, des plateformes clouds ou des solutions de post- traitement en local.

Recommandation : Penser les infrastructures visant à héberger les données d'imagerie avec une architecture générale offrant le plus de flexibilité entre l'hébergement primaire (soin) et l'hébergement secondaire (recherche).

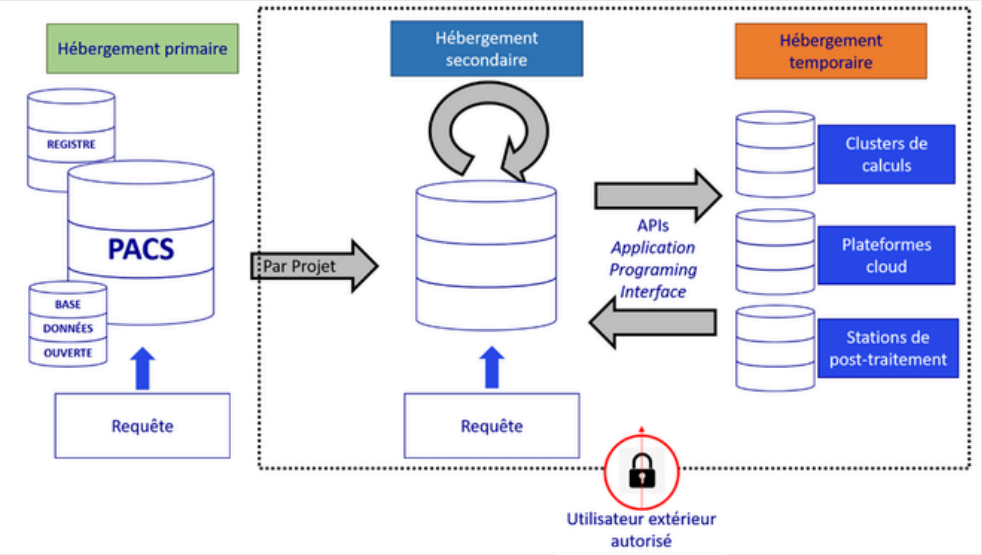


Figure 2 : Représentation schématique de l'architecture générale des hébergements des données d'imagerie.

Plusieurs éléments nécessitent d'être précisés :

- Typologie de données transférées vers l'hébergement secondaire : les échanges entre l'hébergement primaire et cet hébergement secondaire reposent sur des moteurs d'intégration (tels que Mirth Connect, InterSystems Ensemble) pour relier le PACS à la plateforme d'hébergement secondaire de recherche, assurant le transfert et la transformation des flux selon

des formats et des protocoles standardisés (de type DICOMweb, HL7, FHIR, IHE-XDS). C'est à ce stade que les étapes de pseudonymisation décrites ci-dessus sont appliquées.

L'export pourrait théoriquement consister à dupliquer l'intégralité des données du PACS vers l'hébergement secondaire au sein duquel on pourra ensuite disposer des éléments permettant la requête (type API RESTful permettant aux chercheurs autorisés d'interroger, de sélectionner et de récupérer un jeu de données spécifique). L'alternative est une extraction par projet via une requête appliquée au sein de l'hébergement primaire pour n'extraire que les données nécessaires à une étude donnée.

Recommandation : Favoriser une approche par projet qui est la plus conforme au principe de minimisation du RGPD et qui limite les coûts (écologiques et économiques) tout en évitant de saturer l'hébergement secondaire. La constitution systématique de copies non compressées de toutes les images du PACS n'est ni réaliste ni souhaitable.

Ce second modèle peut toutefois être difficile à mettre en place car il nécessite une bonne interopérabilité et performance du système source ce qui n'est pas le cas de tous les PACS hospitaliers. En cas de difficulté, nous recommandons en alternative, une réplication partielle régulière.

- Interopérabilité une fois les images hébergées dans l'environnement secondaire. Les analyses d'images pour la recherche nécessitent souvent des pipelines complexes avec de la puissance de calcul et des ressources logiciels.

Recommandation : Prévoir une infrastructure qui puisse exposer une couche d'API (Application Programming Interface) standardisée permettant de relier de manière sécurisée l'hébergement secondaire à des ressources de calcul externes - telles que des clusters haute performance, des plateformes d'intelligence artificielle ou des outils de post-traitement spécialisés.

Ces interfaces garantissent un accès transitoire et contrôlé aux données, limité à la durée et aux besoins du traitement, tout en préservant leur intégrité et leur traçabilité.
Elles constituent ainsi un levier majeur pour favoriser la mutualisation des capacités de calcul et l'intégration fluide des outils analytiques, sans compromettre la sécurité ni la conformité réglementaire de l'hébergement.

3.3 Sécurité, conformité et traçabilité

La sécurité des données constitue une exigence centrale pour tout hébergement de données d'imagerie. Les mesures techniques et organisationnelles recommandées doivent garantir :

- **La confidentialité des données** : chiffrement systématique des données en transit et au repos.
- **La sécurité d'accès** : authentification forte (SSO, MFA), gestion centralisée des habilitations, cloisonnement des environnements.
- **La traçabilité et l'intégrité** : journalisation complète des accès, des modifications et des exports ; conservation des logs d'audit.
- **La conformité et la gouvernance documentaire** : existence de procédures validées (plan de gestion de la sécurité, charte d'accès aux données, registre des traitements, cahier des charges HDS).
- **L'encadrement des droits d'accès et des responsabilités** : attribution des rôles (administrateur technique, administrateur de données, chercheur) et application du principe de moindre privilège.

Les environnements de recherche doivent offrir des espaces de travail étanches, contenant uniquement les données nécessaires à une étude donnée, accessibles uniquement aux personnes habilitées.

Les exports et extractions doivent être réalisés par des administrateurs désignés ("data managers"), conformément à la délibération CNIL n°2021-118 du 7 octobre 2021.

Recommandation : Mettre en place un comité de pilotage pour chacune des infrastructures pour coordonner l'accès aux données. Ce comité doit impliquer les équipes médicales et notamment les radiologues en charge de la production et de la gestion des images, la direction de la recherche clinique, la direction du numérique, le responsable sécurité des systèmes d'information (RSSI), le délégué à la protection des données (DPO) et le département juridique.

Ce comité doit éditer une gouvernance des données qui gère notamment les accès qu'il faut rendre possible pour les chercheurs en imagerie extérieurs à l'établissement de soins et travaillant au sein d'un EPST (Univ, INSERM, CNRS, Inria), d'un établissement public à caractère industriel et commercial (par ex, CEA), ou d'une structure privée tout en maintenant les principes de sécurité.

3.4 Conservation, compression et durée de stockage des données d'imagerie pour la recherche

La conservation des données d'imagerie dans le cadre de leur réutilisation secondaire soulève des enjeux spécifiques qui ne se posent pas avec la même intensité pour les données cliniques ou biologiques. Les données d'imagerie se distinguent par leur **volume particulièrement important, leur complexité technique**, la diversité des formats selon les modalités (IRM, scanner, échographie, médecine nucléaire), ainsi que par l'existence de politiques de **compression automatique** au sein des systèmes d'archivage cliniques (PACS). Ces particularités ont un impact direct sur la capacité des établissements à préserver des données exploitables à long terme pour la recherche.

Les systèmes PACS hospitaliers appliquent fréquemment des mécanismes de **compression avec ou sans perte** pour optimiser l'espace de stockage, parfois au bout d'un certain délai après l'acquisition. Si ces stratégies sont adaptées aux usages cliniques, elles peuvent altérer des paramètres essentiels pour certaines analyses de recherche, notamment en imagerie quantitative, radiomique ou en apprentissage profond. De même, les politiques d'archivage à long terme, qui déplacent les données vers des stockages moins performants, peuvent limiter l'accès rapide nécessaire à la constitution de bases de recherche.

En parallèle, la conservation prolongée de grandes quantités de données d'imagerie au sein d'hébergements secondaires soulève des **contraintes réglementaires**. Selon le RGPD il faut respecter le principe de **minimisation** (ne traiter que les données nécessaires à la finalité du projet) et le principe de **limitation de durée** (durées de conservation proportionnées à la finalité déclarée). Elle pose également des **contraintes environnementales et économiques**, liées à l'empreinte carbone du stockage numérique et au coût croissant des infrastructures matérielles et logicielles.

Recommandation : Ne conserver les données d'imagerie extraites et pseudonymisées au sein de l'hébergement secondaire que pour la **durée strictement nécessaire à la réalisation du projet**, conformément au RGPD, durée qui doit être définie dans la gouvernance du projet et cohérente avec les périodes de conservation prévues dans le protocole et, le cas échéant, dans la méthodologie de référence (MR) applicable.

En pratique, et sous réserve de compatibilité avec ces cadres, une durée de conservation de **l'ordre de 5 à 10 ans après publication** est fréquemment retenue afin de permettre la vérification scientifique post-publication et de répondre aux attentes des revues et des institutions.

Il est néanmoins possible de prévoir un espace dédié au sein de l'hébergement secondaire pour stocker des examens de façon prolongée en qualité maximale lorsque la compression automatique PACS risquerait de compromettre leur valeur scientifique. Cette stratégie doit rester **sélective**, conforme au principe de minimisation, reposant sur des critères objectifs (programme scientifique spécifique, pathologie rare) et être documentée dans la gouvernance.

A noter que des espaces de stockage de données images anonymisées peuvent constituer des banques de référence uniques, référençables, utilisables par la communauté scientifique et valorisant le travail d'acquisition, de curation et de traitement des chercheurs via la citation des articles correspondants (par exemple d'ADNI[1], PPMI[2], OFSEP[3], ...).

[1] https://adni.loni.usc.edu/

[2] https://www.ppmi-info.org/

[3] https://www.ofsep.org/fr/

3.5. Articulation entre l'échelle locale, régionale, nationale et Européenne

Au niveau local, les centres hospitaliers gèrent l'hébergement primaire des images (PACS) mais aussi secondaire (EDS avec infrastructure dédiée pour l'imagerie). Les plateformes d'imagerie non-hospitalières détiennent également le plus souvent des solutions de stockage plus ou moins structurées.

Au niveau interrégional ou thématique, les données d'imagerie peuvent être fédérées via des plateformes partagées.

Au niveau national, le Heath Data Hub (HDH) n'assure pas de stockage systématique des images (pas de copie centralisée) mais agit comme une couche de coordination, de visibilité et de gouvernance au-dessus des infrastructures locales. Il constitue ainsi un catalogue national en recensant les bases déclarées sous MR-004.

Avec la montée en puissance du règlement de l'espace Européen des données de santésqui est en court de finalisation (eHDS, European Health Data Space), le HDH servira de point d'entrée pour la France pour le partage des données au niveau Européen. Il assurera donc la coordination et facilitera les projets multicentriques nationaux et transnationaux mais pas la centralisation, signifiant que chaque centre doit organiser son stockage en suivant des bonnes pratiques communes.

Recommandation : Travailler à la mise en place des solutions d'hébergement secondaire des données dans le cadre de la politique d'hébergement institutionnelle locale de son EDS en y adossant des espaces de travail pour les images. Il est recommandé d'appliquer les normes ISO 27001 et HDS à toutes ces infrastructures manipulant des données d'imagerie et de garantir l'interopérabilité des systèmes via des API documentées et des protocoles standardisés (type DICOMweb, FHIR). Les besoins croissants d'interopérabilité imposent d'aller vers ces architectures harmonisées qui pourront à terme être fédérées au niveau régional, national ou Européen, s'appuyant sur des standards communs, des infrastructures certifiées et des interfaces compatibles. Des exemples nationaux comme le CATI[1], SHANOIR[2] ou Archimed[3] ont initié la gestion de certaines données au niveau national.

[1] https://joliot.cea.fr/drf/joliot/Pages/Institut/Themes_de_recherche/imagerie_medicale/cati.aspx
[2] https://project.inria.fr/shanoir/
[3] https://www.cic-it-nancy.fr/fr/plateforme-archimed/

4. Réintégration de métadonnées et des données annotées

4.1.Principe de la ré-utilisation tertiaire des données

Chaque projet de recherche basé sur la réutilisation secondaire d'un jeu de données peut générer à son tour un certain nombre de nouvelles données dérivées. Ces données peuvent regrouper, de manière non exhaustive, des données numériques (ex: volume de lésions segmentées), textuelles (ex: catégorisation d'anomalies), géométriques (ex: segmentation de lésions). Afin de centraliser les efforts et d'alimenter la base de données originale, il est souhaitable de définir un mécanisme permettant de réintégrer ces données en vue de projets de recherche ultérieurs

Recommandation : Prévoir et implémenter un mécanisme de réintégration des productions scientifiques au sein de l'hébergement secondaire pour leur ré-utilisation tertiaire.

4.2. Type et format des métadonnées et données annotées

- **Données tabulaires**

L'une des méthodes couramment employées pour stocker des métadonnées consiste à utiliser des fichiers permettant une structuration tabulaire des données, tels que les fichiers CSV ou Excel. Ces fichiers sont souvent utilisés pour conserver des données numériques ou textuelles issues d'un projet de recherche. Cependant, cette approche présente des limitations significatives dans la perspective d'une réutilisation tertiaire de ces données. En effet, la structuration tabulaire, où une ligne correspond généralement à un examen ou à un sujet, ne permet pas de couvrir toutes les complexités et les imbrications d'un dossier d'imagerie. Par exemple, elle ne permet pas facilement de gérer plusieurs examens par patient ou plusieurs séquences par examen, surtout lorsque le nombre de ces séquences n'est pas connu à l'avance. De plus, la réutilisation tertiaire d'un fichier tabulaire nécessite souvent l'extraction de certaines lignes dans de nouveaux fichiers tabulaires, ce qui peut entraîner une multiplication des fragments de la base de données et rendre leur gestion complexe.

Recommandation : **Eviter autant que possible l'utilisation des fichiers tabulaires** pour **le stockage des métadonnées**, afin de favoriser des méthodes plus flexibles et adaptées à la complexité des données d'imagerie.

Une alternative permettant également de stocker des données textuelles ou numériques est l'utilisation de bases de données relationnelles, telles que celles de type SQL (par exemple REDCap qui repose sur des bases de données relationnelles de type SQL, mais en proposent une interface applicative qui masque la complexité du modèle relationnel). Bien que ces bases de données offrent une plus grande flexibilité pour gérer des structures complexes et effectuer des requêtes avancées, elles nécessitent un effort de développement supplémentaire. En particulier, il est crucial d'anticiper à l'avance les types de données à intégrer. Une base de données relationnelle doit établir un lien direct entre les métadonnées ré-intégrées et les sujets, les examens, et/ou les séquences d'étude. Cela permet de faciliter les requêtes lors d'une utilisation tertiaire des données.

- **Données structurées**

Une autre stratégie pour réintégrer les données numériques et textuelles issues d'une production scientifique est l'utilisation de formats de fichiers structurés comme JSON ou XML. Cette solution est particulièrement adaptée à la problématique de l'étude de l'imagerie car elle est hautement flexible et le contenu de ces fichiers structurés peut être personnalisé par projet, permettant alors de maximiser la quantité d'information pouvant être réintégrée, au prix d'un travail de spécification de format éventuellement plus important. De plus, le standard BIDS (cf. section 1.3) intègre l'utilisation de fichiers JSON pour stocker des métadonnées. Les fichiers JSON, en particulier, sont largement utilisés dans ce contexte, en raison de leur lisibilité sans autre outil qu'un éditeur de texte, et de leur facilité d'intégration avec divers outils et langages de programmation. Cette approche permet non seulement une meilleure gestion des données complexes et imbriquées, mais aussi une interopérabilité accrue entre différents systèmes et plateformes de recherche.

Recommandation : Utiliser un format de fichier structuré de type JSON (ou l'utilisation des spécifications BIDS lorsque cela s'applique) pour la réintégration des métadonnées textuelles et numériques générées au décours d'une production scientifique en lien avec l'imagerie.

- **Segmentations et images dérivées**

Les données produites au décours d'un projet scientifique sont de forme variable, et notamment en imagerie les productions sont souvent des segmentations (manuelles ou automatiques) ou des images dérivées (ex. cartographies calculées, imageries génératives). Ces types de données peuvent être stockées sous forme de fichiers d'image standard (DICOM ou NIfTI) et intégrées dans un espace structuré selon le standard BIDS (cf. section 1.3) au sein de dossiers dits "derivatives". De plus, ces fichiers peuvent être associés à des métadonnées décrivant le type de processus ayant permis leur création, ainsi qu'un lien vers les données sources ayant permis leur production.

Recommandation : Utiliser les standards NifTI/BIDS pour le stockage des segmentations et images dérivées quand cela est applicable.

- **Autres données**

Enfin, certaines données et métadonnées ne sont ni textuelles, ni numériques, ni de type image. Il peut s'agir par exemple de données brutes, de fichiers géométriques de rendu surfacique, de matrice de recalage, de fichiers dans des formats propriétaires, etc. Celles-ci peuvent également être stockées sous forme de fichiers binaires au sein d'une structuration de données de type BIDS. Cependant, le stockage de fichiers en format propriétaire expose à un risque de perte de données à long terme, si le logiciel utilisé au moment de la création de la donnée n'est plus accessible ou disponible au moment de la réutilisation tertiaire des données.

Recommandation : Utiliser la structuration BIDS pour stocker les autres types de données, en favorisant au maximum **l'utilisation de formats de fichier non propriétaires** (open source ou équivalents).

4.3 Méthodologie de description

- **Définir les éléments pertinents**

L'intérêt de la réintégration réside principalement dans l'augmentation de la valeur de la base de données originale, ainsi enrichie par les données réintégrées. Cependant, tous les éléments générés au cours d'un projet de recherche ne sont pas destinés à être réintégrés à long terme dans une base de données. Il est donc essentiel que l'utilisateur et le producteur de la base de données s'accordent sur le type de données pouvant être réintégrées. L'intérêt des données à intégrer doit être évalué en fonction de leur pertinence scientifique actuelle et potentielle, ainsi que de la charge supplémentaire que leur réintégration imposera sur l'infrastructure de stockage. Typiquement, les données textuelles et numériques présentent généralement une charge faible sur l'infrastructure en raison de leur petite taille. En revanche, les données les plus intéressantes scientifiquement pour une réutilisation tertiaire sont souvent des segmentations d'image, qui ont un impact plus important sur le stockage.

Recommandation : **L'utilisateur et le producteur de la base de données doivent s'accorder en amont sur les données dérivées pertinentes pouvant être sujettes à réintégration ultérieure. Ces éléments doivent être discutés au sein de structures de gouvernance de la base de données.**

- **Décrire les données**

Un élément crucial des données réintégrées est leur description. En effet, lors d'une réutilisation tertiaire des données réintégrées, il sera indispensable à l'utilisateur final de disposer des informations nécessaires à la description des données et métadonnées réintégrées : méthode de création, outils utilisés, structure...

Recommandation : Utiliser un **système de documentation** permettant de stocker les informations et articles scientifiques pertinents à la description des données réintégrées.

Enfin, chaque donnée dérivée est créée à partir d'un ensemble de données (une ou plusieurs séquences, d'autres données dérivées...). Il est souhaitable de conserver une trace directe entre le fichier dérivé et le ou les fichiers sources ayant permis leur production (dénommé provenance). Ces fichiers sources peuvent constituer une base de données de référence permettant

(voir au-dessus) de reproduire les données dérivées ou de les étendre avec de nouveaux outils ou données.

- **Décrire les producteurs de données**

Dans les éléments de documentation des données et métadonnées réintégrées, il est fortement souhaitable d'inclure des informations sur les centres, les responsables, les experts, les utilisateurs et les annotateurs impliqués dans la production de ces données dérivées. De plus, il est important de mentionner les financements ayant soutenu cette production. En effet, le travail fourni, destiné à être réutilisé, doit être dûment documenté pour valoriser les producteurs de données en cas de réutilisation des données qu'ils ont générées (cf. section 5). Cette documentation permettra d'attribuer correctement le travail des individus impliqués dans une étude ultérieure ayant recours à une utilisation tertiaire de données, et d'assurer la traçabilité des données, ce qui est crucial pour la validation et la vérification des résultats de la recherche.

Recommandation : Adjoindre une **description exhaustive des producteurs de la donnée secondaire** à la documentation de celle-ci.

4.4. Traçabilité et conservation des résultats de recherche en imagerie

Au-delà des données d'imagerie elles-mêmes, les projets de recherche génèrent un ensemble de résultats intermédiaires ou finaux (modèles entraînés, scripts, paramètres de traitement, journaux d'exécution, tableaux de performance) indispensables à la reproductibilité scientifique. Il y a une exigence de traçabilité de ces résultats qui concernent l'ensemble des recherches biomédicales, mais qui revêt une importance particulièrement marquée dans le domaine de l'imagerie. En effet, la taille des données, la diversité des formats, la complexité des pipelines analytiques et la dépendance à des environnements logiciels spécifiques rendent difficile l'archivage systématique de l'ensemble des éléments nécessaires à la ré-analyse ou à la vérification des résultats. Ces éléments doivent être conservés dans des conditions garantissant leur pérennité et leur auditabilité. Il s'agit d'une obligation scientifique et réglementaire. Les équipes sont confrontées à l'absence d'espace institutionnel pour archiver ces résultats et à la fragmentation des résultats entre des postes personnels ou des serveurs locaux.

Recommandation : nous recommandons que chaque projet intégrant des données d'imagerie définisse, dès sa conception, un **dispositif de conservation des résultats** (scripts, modèles, journaux, environnements d'exécution), distinct du stockage des données, permettant d'assurer la traçabilité et la reproductibilité sur la durée réglementaire dans un environnement sécurisé et documenté.

En pratique, et sous réserve de compatibilité avec la durée prévue dans le protocole et dans la méthodologie de référence (MR) applicable, une conservation de **5 à 10 ans après publication** est généralement adoptée.

5.Rémunération/Valorisation

5.1. De la donnée au savoir : principes de valorisation

La réutilisation des données d'imagerie issues du soin dans le cadre d'un travail de recherche va créer de la valeur qui peut prendre la forme de nouvelles connaissances scientifiques, de publications, de communications, de brevets, de dépôts de solution logiciels, ou de résultats amenant à obtenir de nouvelles subventions dans le cadre d'appels d'offre.

Cet aboutissement implique une mobilisation de moyens et compétences multiples (**chaîne de valeur complexe**) : extraction des examens depuis le PACS, anonymisation des fichiers DICOM, curation et structuration dans un hébergement secondaire, stockage sécurisé, développement et mise en œuvre d'outils d'analyse, analyses automatiques et/ou contrôle manuel, création d'annotations ou de masques, analyses de résultats, statistiques, diffusion des résultats scientifiques. Il est primordial que chacune de ces étapes soient valorisées.

La valeur ne réside **ni dans la donnée brute, ni dans son volume**, mais dans le **travail collectif de transformation** qui la rend exploitable et qui permet de produire de la connaissance scientifique.

Ainsi, la valorisation doit être comprise comme la **reconnaissance** - **(i) scientifique/académique et (ii) financière** - **de la chaîne de travail complexe depuis la curation jusqu'à l'analyse**, et non comme une transaction portant sur la donnée elle-même.

Recommandation : Affirmer le principe que la donnée d'imagerie n'a pas de valeur marchande propre de façon isolée ; sa valeur découle du travail collectif et qualifié nécessaire à sa réutilisation scientifique.
Il peut être recommandé de publier dans un journal adapté un article décrivant la base de données à des fins de valorisation et citation lors de son utilisation.

5.2. Cadre éthique et juridique : la notion de dépositaire

Les données d'imagerie produites dans le cadre du soin **n'appartiennent à personne:** ni au patient, ni au radiologue / médecin nucléaire, ni à l'établissement. Si on prend l'exemple d'une IRM réalisée dans le soin du patient, chaque acteur reçoit une contribution dans la chaîne du soin : le patient bénéficie de l'examen pour son soin, le professionnel radiologue / médecin nucléaire est rémunéré pour son acte, et l'établissement de santé perçoit le forfait technique correspondant. Ainsi, la valeur de la donnée d'imagerie a déjà été attribuée à chacun et ne peut plus être réattribuée sans un nouveau travail collectif et nécessaire pour la réutiliser en recherche. Ainsi, personne n'est propriétaire de la donnée au sens juridique, mais toute réutilisation des images dans un projet de recherche confère à l'équipe en charge le rôle de **dépositaire des données**.

Ce rôle implique la **responsabilité** devant la loi de leur pseudonymisation, de leur sécurité, de leur utilisation conforme au cadre éthique et juridique, de la traçabilité et de la rigueur des traitements effectués pour la recherche.

L'approche dépositaire est particulièrement pertinente pour l'imagerie, dont la sensibilité technique et la complexité des métadonnées exigent une supervision experte, souvent assurée par les radiologues et les équipes de techniciens et d'ingénieurs. Cette logique s'oppose à toute revendication de propriété et replace la gouvernance des données dans un cadre de **responsabilité partagée et d'obligation de moyens.**

Recommandation : Remplacer toute référence à la « propriété » des données par la notion de responsabilité du dépositaire, afin de clarifier le cadre éthique et juridique des projets de recherche.

5.3. Cartographie des contributions et gouvernance scientifique

L'exploitation des données d'imagerie repose sur une chaîne de production impliquant de nombreux acteurs : ingénieurs pour l'extraction et la qualité, radiologues, médecins nucléaires, data scientists, chercheurs en imagerie, cliniciens non radiologues, biostatisticiens, responsables

de la sécurité des systèmes d'information, délégué à la protection des données, coordinateurs de projet, attachés de recherche clinique...

Pour assurer une reconnaissance équitable, il est recommandé d'instaurer un **registre « vivant » des contributions**, mis à jour tout au long du projet. Ce registre recense les actions menées à chaque étape - de la pseudonymisation à la publication ou tout autre livrable - et sert de référence pour l'attribution des rôles et la planification des publications.

Cet outil permet :

- de **rendre traçables les apports individuels et collectifs** ;
- de **prévenir les conflits de liste d'auteurs** pour rétribuer la valeur scientifique;
- de **quantifier les contributions** pour une rétribution financière
- et de **favoriser la transparence dans la gouvernance des projets multicentriques.**

Recommandation : Rendre obligatoire la mise en place d'un registre des contributions et d'une charte de gouvernance des publications dans tout projet collectif ou multicentrique impliquant des données d'imagerie.

5.4. Reconnaissance académique

L'implication dans les différentes étapes conduisant à la valorisation de la recherche doit faire l'objet d'une reconnaissance qui peut être académique. On entend par reconnaissance académique l'affiliation des acteurs de la chaîne dans les publications scientifiques, les communications ou les brevets issus de la recherche. Etant donné la présence d'acteurs multiples aux différentes étapes, il faut s'appuyer sur des règles d'auteurs suivant les critères du **Comité International des Rédacteurs de Revues Médicales (ICMJE)**, qui définissent quatre conditions cumulatives : participation significative à la conception, à l'acquisition et/ou à l'analyse, rédaction ou révision intellectuelle, approbation de la version finale, et responsabilité de l'intégrité du travail. Ces critères doivent être appliqués de façon stricte en veillant à prévenir les dérives courantes (dit gift authorship, ghost authorship, white bull effect) tel que formalisé, par exemple, par la Société Française de Neuroradiologie pour les travaux impliquant l'imagerie médicale[1]).

[1] https://www.sfnr.net/recherche-innovation/regles-authorship-recherche

Pour ces projets sur données d'imagerie avec acteurs multiples, nous recommandons qu'une gouvernance des publications soit anticipée dès la phase de conception, avec :

- la création d'un **comité de publication**, garant de l'équité et de la rotation des positions d'auteur ;
- la formalisation d'une **charte de liste d'auteurs** précisant les critères d'éligibilité, les règles d'ordre des auteurs;
- la diffusion systématique d'une section « Author Contributions » explicitant les tâches réalisées par chacun.
- L'anticipation des contraintes de publication (limite du nombre des auteurs) et la gestion des remerciements (section acknowledgments)

Chacun doit être valorisé à la hauteur de ses contributions. Si le travail a utilisé des annotations, des contours ou toutes autres métadonnées créés par un contributeur lors d'un travail antérieur et réintégrées dans la plateforme d'hébergement comme recommandé plus haut (cf. section 4 ; métadonnées pour utilisation tertiaire), ce contributeur doit être également valorisé à la hauteur du travail accompli.

Recommandation : Inscrire l'adoption d'une charte de liste d'auteurs et d'un comité de publication comme exigence de bonne pratique dans toute étude multicentrique en imagerie.

5.5 Reconnaissance financière et rémunération : modèle opérationnel

La réutilisation des données d'imagerie mobilise des ressources techniques et humaines qui peuvent être considérables. La reconnaissance académique participe à la valorisation institutionnelle via les indicateurs bibliométriques (points SIGAPS et financements MERRI) mais ceci est insuffisant pour reconnaitre la contribution de tous les acteurs. La réumunération doit donc aussi **couvrir le coût réel des opérations nécessaires à la mise à disposition et à la qualité des données**, sans constituer une vente des données.

Recommandation : Mettre en place une grille nationale harmonisée sur la base des éléments proposés par le HDH et le comité stratégique des données de santé.

Une telle grille pourra être appliquée dans chaque centre en toute transparence selon le modèle de la grille unique des coûts/surcoûts des actes d'imagerie pour la recherche éditée il y a quelques années à l'initiative du Collège des Enseignants en Radiologie de France (CERF), et ayant servi de base au volet Imagerie de la convention unique industrielle notamment.
Les postes typiques doivent couvrir les ressources humaines et les ressources matérielles (frais de stockage en €/To/an et frais d'analyses en €/GPU-heure ou €/heure de machine virtuelle accélérée) pour couvrir les éléments suivants :

- frais administratifs et juridiques de contractualisation ;
- extraction et anonymisation DICOM ;
- curation, structuration et contrôle qualité ;
- stockage sécurisé et gestion de projet ;
- analyses avec contrôle qualité radiologique ;
- accompagnement méthodologique et statistique.

Le processus s'appuie sur un **flux de travail en quatre étapes** :

1. Instruction scientifique et technique en réponse à une demande sur formulaire standardisé pour appréhender les éléments demandés et la faisabilité.
2. Pré-design et chiffrage. Cette étape consiste en une reformulation technique du cahier des charges, un chiffrage budgétaire et un cadrage réglementaire via l'utilisation d'une convention définissant les usages autorisés, les conditions de partage, et les retombées scientifiques. Nous recommandons le recours à des conventions cadres afin de réduire les délais de réponse.
3. Discussion avec les partenaires, validation et signature du devis,
4. Mise en œuvre opérationnelle associant l'export et l'ouverture des accès.

Cette rémunération doit viser la **soutenabilité des infrastructures**.

Des indicateurs de suivi seront intégrés : délais de production, complétude des données, conformité aux chartes qualité, et retombées scientifiques et financières (MERRI : SIGAPS/SIGREC).

Recommandation : **Adopter un modèle national de rémunération par postes** de travail, garantissant la transparence, la soutenabilité et l'équité entre établissements. Mettre en place un circuit de réception des demandes, de validation et de mise en place des espaces projets validés.

Au total, la reconnaissance de la chaîne de travail ayant conduit à valoriser des données d'imagerie du soin à des fins de recherche doit repos et sur un cadre éthique clair (notion de dépositaire des données) et une gouvernance scientifique dans laquelle les contributions de chacun sont listées et suivies. Ces éléments permettent de définir de façon transparente la reconnaissance académique à travers la liste d'auteurs dans les publications scientifiques. La standardisation d'une grille de coûts et de modèles de convention de mise à disposition des données d'imagerie doit permettre la soutenabilité financière et la reconnaissance des efforts collectifs.

SYNTHESE des RECOMMANDATIONS

1.Recommandations en lien avec la nature et la structuration des données d'imagerie

1.1. Recommandation sur l'utilisation des fichiers de données brutes

- **Recommandation 1.1**

Ne pas stocker systématiquement les données brutes. Si on anticipe qu'une certaine catégorie d'images collectées dans le soin pourrait nécessiter l'utilisation de données brutes en vue d'une analyse, il faudra stocker uniquement et spécifiquement ces données brutes. Ces situations sont peu fréquentes.

1.2. Recommandation en lien avec les fichiers DICOM utilisés comme données sources

- **Recommandation 1.2**

Lors de la constitution d'une base de données primaire, s'assurer que les fichiers DICOM soient conservés sur les systèmes d'information clinique dans leur forme originale sans altération, afin de garantir l'exactitude et la fiabilité des informations qu'ils contiennent. Nous suggérons également, quand cela est possible, d'éviter le stockage d'images compressées via des algorithmes « avec perte » (type JPEG). Les procédures de pseudonymisation devront respecter cette forme originale et établir une copie altérée de ces fichiers sources lors de l'export vers des environnements recherche dédiés pour réutilisation secondaire.

1.3. Recommandation spécifique aux données d'échographie

- **Recommandation 1.3**

Pour les images d'échographie, privilégier l'export DICOM lorsque cela est possible, ou encapsuler les vidéos dans un conteneur DICOM afin de garantir la présence de métadonnées exploitables et de faciliter la pseudonymisation. Si seules des vidéos non DICOM sont disponibles, nous recommandons une conversion vers un format standardisé (par exemple MPEG-4), associée à une documentation détaillée des conditions d'acquisition

et à une procédure rigoureuse de contrôle des images suivie du masquage des zones identifiantes. Dans tous les cas, les données pour la recherche doivent être sans compression ou utiliser une compression non destructrice.

1.4. Recommandation sur les formats et l'organisation des fichiers pour la recherche

- **Recommandation 1.4**

Pour un projet de recherche donné, nous recommandons de prévoir, au sein de l'environnement recherche, une copie DICOM pseudonymisée ainsi qu'une éventuelle copie convertie dans un format open-source adapté aux analyses pour la recherche (de type NIfTI) si cela est requis pour le projet considéré. Nous recommandons également de séparer les données images des métadonnées pertinentes à la recherche en utilisant des méthodes d'organisation de type BIDS documentées par le responsable de l'export des données.

2. Recommandations liées à la pseudonymisation et à la confidentialité des données d'imagerie

2.1. Recommandation sur la pseudonymisation de l'en-tête DICOM

- **Recommandation 2.1**

Modifier ou remplacer les champs DICOM, plutôt que de les supprimer, afin de ne pas altérer la compatibilité technique des fichiers avec les outils d'analyse.

2.2. Recommandation sur le contrôle du contenu pixel (pixel burning)

- **Recommandation 2.2**

Contrôler le contenu pixel en combinant :
- une détection automatique par reconnaissance de texte (OCR) ;
- un contrôle visuel sélectif sur les séries identifiées comme à risque ;
- un effacement ou masquage précis des zones concernées, suivi d'une vérification visuelle.

Cette étape doit être intégrée au protocole de pseudonymisation, avec traçabilité des outils et validation par le responsable scientifique du projet.

2.3. Recommandation pour les formats d'image non-DICOM

- **Recommandation 2.3**

Lors de l'utilisation de fichiers non-DICOM, s'assurer que les métadonnées sont non-identifiantes et contrôler la structure du nom des fichiers et des dossiers, ainsi que les images qui peuvent véhiculer des identifiants implicites. Nous recommandons une revue humaine de ces cas spécifiques.

2.4. Recommandation spécifique à la neuroimagerie (défacialisation)

- **Recommandation 2.4**

À ce jour, nous ne recommandons pas de défacialisation pour une utilisation secondaire d'images cérébrales dans un EDS. La défacialisation sera en revanche nécessaire si on envisage une base publique ou semi-publique, ce qui nécessitera une anonymisation (et non pas seulement une pseudonymisation) et ce qui pourra se faire uniquement via des images NIfTI.

2.5. Recommandation pour les données non-images associées aux examens d'imagerie

- **Recommandation 2.5**

Pour les données textuelles associées aux images (comptes rendus, tableaux), nous recommandons l'utilisation d'une whitelist afin de limiter au mieux les risques d'inclusion d'informations nominatives.

2.6. Recommandation relative aux risques liés à l'intelligence artificielle générative

- **Recommandation 2.6**

Procéder à un réexamen spécifique du risque de ré-identification pour tout projet faisant intervenir des modèles génératifs, incluant :
(i)la vérification des procédures de pseudonymisation en amont ;
(ii)la limitation stricte des accès aux environnements d'entraînement et aux modèles ;

(i)un contrôle de l'absence de reproduction d'images réelles ou d'éléments identifiants lors de tests en sortie de modèle.

2.7. Recommandation sur les bonnes pratiques institutionnelles et la traçabilité

- **Recommandation 2.7**

Les bonnes pratiques institutionnelles que nous recommandons sont :
– la documentation détaillée des scripts et des versions logicielles utilisées ;
– l'association en amont du DPO et du RSSI pour la validation de la conformité des procédures de pseudonymisation, et, le cas échéant, l'examen par un comité d'éthique ou scientifique local afin d'évaluer la proportionnalité et la pertinence éthique des traitements ;
– la conservation sécurisée de la clé d'appariement dans un environnement séparé ;
– la réalisation d'audits réguliers de conformité ;
– l'alignement sur les référentiels reconnus : ISO 27001, HDS et guides de la CNIL relatifs à la recherche en santé.

3. Recommandations liées à l'hébergement et aux échanges des données d'imagerie

3.1. Recommandation sur l'hébergement secondaire dédié à la recherche

- **Recommandation 3.1**

Prévoir une solution d'hébergement secondaire offrant une articulation équilibrée, garantissant à la fois la protection des personnes – en anticipant dès à présent l'application des exigences HDS aux infrastructures de recherche accueillant des données d'imagerie – et l'efficacité pour l'exécution des projets scientifiques.

3.2. Recommandations sur l'architecture des infrastructures d'imagerie pour la recherche

- **Recommandation 3.2**

Penser les infrastructures visant à héberger les données d'imagerie avec une architecture générale offrant le plus de flexibilité entre l'hébergement primaire (soin) et l'hébergement secondaire (recherche).

- **Recommandation 3.3**

Favoriser une approche par projet, conforme au principe de minimisation du RGPD, limitant les coûts et évitant de saturer l'hébergement secondaire.

- **Recommandation 3.4**

Prévoir une infrastructure exposant une couche d'API standardisée permettant de relier de manière sécurisée l'hébergement secondaire à des ressources de calcul externes.

3.3. Recommandation sur la gouvernance, la sécurité et la conformité des infrastructures

- **Recommandation 3.5**

Mettre en place un comité de pilotage pour chacune des infrastructures pour coordonner l'accès aux données. Ce comité doit impliquer les équipes médicales et notamment les radiologues en charge de la production et de la gestion des images, la direction de la recherche clinique, la direction du numérique, le responsable sécurité des systèmes d'information (RSSI), le délégué à la protection des données (DPO) et le département juridique.

3.4. Recommandation sur la conservation, la compression et la durée de stockage

- **Recommandation 3.6**

Ne conserver les données d'imagerie extraites et pseudonymisées au sein de l'hébergement secondaire que pour la durée strictement nécessaire à la réalisation du projet, conformément au RGPD. En pratique, et sous réserve de compatibilité avec la durée prévue dans le protocole et dans la méthodologie de référence (MR) applicable, une conservation de 5 à 10 ans après publication est généralement adoptée.

3.5. Recommandation sur l'articulation entre échelles locale, nationale et européenne

- **Recommandation 3.7**

Travailler à la mise en place de solutions d'hébergement secondaire dans le cadre de la politique institutionnelle locale des EDS, en appliquant les normes ISO 27001 et HDS et en garantissant l'interopérabilité via des API et protocoles standardisés afin de pouvoir garantir une fédération régionale et nationale.

4. Recommandations relatives à la réintégration des métadonnées et des données annotées

4.1. Recommandation sur la réutilisation tertiaire des données

- **Recommandation 4.1**

Prévoir et implémenter un mécanisme de réintégration des productions scientifiques au sein de l'hébergement secondaire pour leur ré-utilisation tertiaire.

4.2. Recommandations sur les formats de métadonnées et données dérivées

- **Recommandation 4.2**

Éviter autant que possible l'utilisation des fichiers tabulaires pour le stockage des métadonnées.

- **Recommandation 4.3**

Utiliser un format structuré de type JSON ou les spécifications BIDS pour la réintégration des métadonnées.

- **Recommandation 4.4**

Utiliser les standards NIfTI/BIDS pour le stockage des segmentations et images dérivées lorsque cela est applicable.

- **Recommandation 4.5**

Utiliser la structuration BIDS pour les autres types de données en privilégiant les formats non propriétaires.

4.3. Recommandations sur la gouvernance et la documentation des données réintégrées

- **Recommandation 4.6**

S'accorder en amont sur les données dérivées pertinentes pouvant être réintégrées.

- **Recommandation 4.7**

Utiliser un système de documentation pour décrire les données réintégrées.

- **Recommandation 4.8**

Adjoindre une description exhaustive des producteurs des données secondaires.

- **Recommandation 4.9**

Définir dès la conception du projet un dispositif de conservation des résultats garantissant la traçabilité et la reproductibilité. En pratique, et sous réserve de compatibilité avec la durée prévue dans le protocole et dans la méthodologie de référence (MR) applicable, une conservation de 5 à 10 ans après publication est généralement adoptée

- Recommandations relatives à la valorisation, à la gouvernance scientifique et à la rémunération

5.1. Recommandations sur les principes de valorisation

- **Recommandation 5.1**

Affirmer que la donnée d'imagerie n'a pas de valeur marchande propre et que sa valeur découle du travail collectif nécessaire à sa réutilisation scientifique.

- **Recommandation 5.2**

Remplacer toute référence à la propriété des données par la notion de responsabilité du dépositaire.

5.2. Recommandations sur la gouvernance scientifique et la reconnaissance académique

- **Recommandation 5.3**

Rendre obligatoire la mise en place d'un registre des contributions et d'une charte de gouvernance des publications.

- **Recommandation 5.4**

Inscrire l'adoption d'une charte de liste d'auteurs et d'un comité de publication comme exigence de bonne pratique.

5.3. Recommandations sur la rémunération et la soutenabilité financière

- **Recommandation 5.5**

Mettre en place une grille nationale harmonisée de rémunération fondée sur les postes de travail.

- **Recommandation 5.6**

Adopter un modèle national de rémunération garantissant transparence, soutenabilité et équité entre établissements.