# Méthodes d'intégration sémantique de données patients multi domaines dans des architectures d'entrepôts de données de santé : une approche orientée cas d'usage

Pierre Lemordant

UNIVERSITÉ DE RENNES 1

# THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1

ECOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies de l'Information et de la Communication*
Spécialité : Informatique

Par
**Pierre Lemordant**

# Méthodes d'intégrations sémantiques de données patients multi domaines dans des architectures d'entrepôts de données de santé

Une approche orientée cas d'usage

**Thèse présentée et soutenue à Rennes, le 02/11/2022**
**Unité de recherche : Laboratoire Traitement du Signal et de l'Image, Equipe Données Massives en Santé**

**Rapporteurs avant soutenance** :

| | |
|---|---|
| Marc-Olivier GAUCI | Praticien Hospitalier, HDR, CHU de Nice |
| Marie-Christine JAULENT | Directrice de Recherche INSERM. LIMICS |

**Composition du Jury :**

| | | |
|---|---|---|
| Président : | Christophe AUBE | PU-PH, Radiologue, CHU d'Angers |
| Dir. de thèse : | Marc CUGGIA | PU-PH, Université de Rennes 1 |
| Co-dir. de thèse : | Bernard Gibaud | CRHC, LTSI UMR 1099, Inserm |

A mes parents, mon frère Vincent et mes sœurs Adèle et Claire

# Remerciements

Je remercie les membres du jury, Madame Jaulent et les docteurs Aubé et Gauci, de m'avoir fait l'honneur d'évaluer le résultat de ces 3 ans et 9 mois de thèse.

J'ai eu la chance d'effectuer une thèse CIFRE dans le cadre du partenariat entre l'équipe DOMASIA du LTSI et l'entreprise Enovacom. Je tiens à remercier particulièrement mon encadrant de thèse chez Enovacom, Mr Cyril Garde, pour le temps qu'il m'a permis d'accorder à ma thèse. J'exprime ma profonde reconnaissance à mes directeurs de thèse, Mr Bernard Gibaud qui m'a accompagné dans la première partie de cet effort et m'a aidé à me familiariser avec le domaine de la recherche et le Pr Marc Cuggia, directeur de l'équipe DOMASIA, qui a su ajouter à ses responsabilités celle de m'aider à mener ce travail à bien et m'a assisté et encouragé jusqu'au bout.

J'ai le plaisir de travailler dans cette équipe depuis maintenant presque 5 ans, ce qui m'a donné l'opportunité de rencontrer de nombreuses personnes sympathiques et inspirantes. Je remercie ainsi celles qui ont poursuivi leur route ailleurs, Françoise, Véronique, Marie-Lisen et particulièrement Canelle qui donnait à l'épreuve de la thèse l'apparence d'une simple formalité. Je remercie également ceux que le laboratoire a toujours la chance de compter dans ses rangs, Denis, Guillaume, Pascal, Christian, Christine et Marc. Enfin, je remercie ceux, nombreux, qui nous ont rejoint ces dernières années, Youenn, Morgane, Sandie, Vivien, Arthur, Samir, Audrey, Mathilde, Naïma, Mohamed, Noémie, Marie, Aurélie et Boris.

Je garde d'excellents souvenir de moments passés, au bureau, en déplacement, week-end d'intégration, dans les restaurants et bars de Rennes, devant la machine à café ou en réunion, avec chacun d'entre eux.

Je remercie mes amis qui ont enduré mes incertitudes et m'ont soutenu, chacun à leur façon, Guillaume, Gautier, Maxime, Marie, Guilhem, Yohann.L, Yohann.P, Youenn, Maryon, Clément, Thibault, Thomas, Kévin et Sébastien.

Enfin, je remercie ma famille, mes parents, mon frère Vincent et sa compagne Emeline ainsi que mes sœurs Adèle et Claire. J'aimerais pouvoir rendre à chaque personne citée ici un peu de ce qu'elle m'a apporté.

# Table des matières

# Résumé

Dans le domaine de l'imagerie médicale, les évolutions techniques (stockage de données, modalités d'imagerie, …) et méthodologiques (médecine personnalisée, évolution de l'imagerie diagnostique et interventionnelle, …) des dernières décennies ont fait apparaitre des enjeux majeurs concernant l'usage de l'imagerie dans le soin et dans la recherche, notamment de l'utilisation secondaire de cette source de données.

L'objectif de cette thèse a été de développer des méthodes afin de rendre intégrables et réutilisables ces données d'imagerie dans une solution d'entrepôt de données de santé et de mettre en œuvre cette intégration sémantique dans le cadre concret de la solution d'entrepôt développée dans notre laboratoire.

Dans ce travail, nous avons étudié les outils et méthodes d'alignement de terminologie locales et de références pour permettre la réutilisation et le partage des données d'imagerie. Nous avons conçu une preuve de concept d'outil de classification des examens d'imagerie utilisant le raisonnement ontologique. Enfin, nous avons développé et déployé un prototype de module d'intégration sémantique des données d'imagerie permettant de gérer le trajet des données depuis le PACS jusqu'à l'entrepôt.

# Abstract

In the field of medical imaging, technical (data storage, imaging modalities, etc.) and methodological (personalised medicine, evolution of diagnostic and interventional imaging, etc.) developments over the last few decades have raised major challenges concerning the use of imaging in care and research, particularly the secondary use of this data source.

The objective of this work was to develop methods to integrate and reuse this imaging data in a clinical data warehouse solution and to implement this semantic integration in the framework of the warehouse solution developed in our laboratory.

In this work, we studied tools and methods for local terminology alignment with reference terminology to enable the reuse and sharing of imaging data. We designed a proof of concept for an imaging classification tool using ontological reasoning. Finally, we developed and deployed a prototype module for semantic integration of imaging data to manage the data path from the PACS to the clinical data warehouse.

# Productions scientifiques liées à la thèse

Lemordant, P., Mougin, F., Cabon, S., Gandon, Y., Bouzillé, G., & Cuggia, M. (2022). Indexing Imaging Reports for Data Sharing: A Study of Mapping Using RadLex Playbook and LOINC. Studies in Health Technology and Informatics, 294, 312-316.

Lemordant, P., Gibaud, B., Garde, C., Delarche, S., Goudet, D., & Cuggia, M. (2020, September). Ontology-Based Classification of Radiological Procedures for Consistent Sharing in Clinical Data Warehouses. In *ICBO/ODLS* (pp. 1-11).

Lemordant, P., Bouzille, G., Mathieu, R., Thenault, R., Gibaud, B., Garde, C., ... & Cuggia, M. (2022). How to Optimize Connection Between PACS and Clinical Data Warehouse: A Web Service Approach Based on Full Metadata Integration. In *MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation* (pp. 27-31). IOS Press.

# Abréviations

ACR : American College of Radiology

AMM : Autorisation de Mise sur le Marché

ANR : Agence National de la Recherche

ANSI : American National Standards Institute

ANSM : Agence Nationale de Sécurité du Médicament

CCAM : Classification Commune des Actes Médicaux

CDC : Centre de Données Cliniques

CEN/TC251 : Comité Européen de Normalisation / Technical Committee 251

CHU : Centre Hospitalier Universitaire

CIC-IT : Centre d'investigation clinique et d'innovation technologique

CIM-10 : Classification Internationale des Maladies, 10ème version

CNIL : Commission Nationale de l'Informatique et des Libertés

CPT : Current Procedural Terminology

DICOM : Digital imaging and COmmunications in Medicine

DMP : Dossier Médical Partagé

eCRF : électronique Case Report Form

eHOP : entrepôt HÔPital

ETL : Extract, Transform & Load

FAIR : Findability, Accessibility, Interoperability, and Reusability

HUGO : Hôpitaux Universitaires du Grand Ouest

IRM : Imagerie par Résonance Magnétique

i2b2 : Informatics for Integrating Biology and the Bedside

I4DW : Imaging For DataWarehouse

JIRA : Japan Medical Imaging and Radiological Systems Industries Association

LOINC : Logical Observation Identifiers Names & Codes

LTSI : Laboratoire Traitement du Signal et de l'Image

NEMA : National Electrical Manufacturers Association

NIH : National Institute of Health

NLM : National Library of Medicine

NLP : Natural language processing

OCR : Optical Character Recognition

PACS : Picture Archiving and Communication System

PMSI : Programme de Médicalisation du Système d'Information

RCP : Réunion de Concertation Pluridisciplinaire

RECIST : Response Evaluation Criteria in Solid Tumors

REST : REpresentational State Transfer

RGPD : Règlement Général sur la Protection des Données

SNDS : Système National des Données de Santé

SNIIRAM : Système national d'information inter-régimes de l'assurance maladie

SNOMED-CT : Systematized Nomenclature of Medicine - Clinical Terms

SUV : Standardized Uptake Value

TI : Terminologie d'Interface

TR : Terminologie de Référence

UMLS : Unified Medical Language System

XML : eXtensible Markup Language

XNAT : eXtensible Neuroimaging Archive Toolkit

# Table des figures

# Introduction générale

L'évolution des technologies dans le domaine du traitement informatique des données et de leur stockage a permis de générer une immense masse de données, à un point tel qu'il est devenu impossible de les gérer avec les technologies actuellement disponibles. Le terme "big data" est apparu pour décrire ces données volumineuses et ingérables [1]. Le secteur de la santé est particulièrement concerné par les questions d'organisation et d'extraction du sens des données et a donc dû évoluer en profondeur sur les vingt dernières années avec l'apparition d'outils et de méthodes pour stocker ces données de natures inédites et massives (nouvelles technologies d'imagerie, données génomiques et autres, stockées dans des bases de données distribuées ou spécialisées …) et pour les analyser (machine learning, architecture de calcul distribué, langages de programmation propres à la science des données, …). Depuis la dématérialisation des données de santé, les données provenant du soin, de la recherche médicale et du domaine médico-administratif sont devenues une immense mine de données qui continue de croître.

Ces données et leur potentiel sont des enjeux majeurs pour la santé, car ils ouvrent des perspectives dans le développement de nouvelles méthodes diagnostiques, thérapeutiques ou d'aide à la décision que ce soit au niveau individuel ou populationnel. Cette masse de données et les nouvelles capacités de traitements ont permis à la médecine de faire évoluer la modélisation des processus et interactions à différentes échelles (génome, organes, exposome, …) et ainsi la définition de nouveaux marqueurs diagnostiques, thérapeutiques ou pronostiques qui permettent de mieux comprendre les pathologies et des processus physiopathologiques. Le but de cette approche est la médecine personnalisée [2], capable de capitaliser sur les connaissances extraites de milliers de cas pour aider au mieux un patient donné, par exemple avec des systèmes d'aide au diagnostic ou des systèmes permettant de simuler l'évolution d'une pathologie ou l'impact d'une intervention. Les capacités à mobiliser, manipuler et traiter ces données sont donc primordiales et concentrent les efforts de nombreux acteurs, publics ou privés, des grands groupes aux start-ups.

Cette utilisation des données de santé (générées pendant la pratique du soin) à des fins de recherche est une utilisation secondaire des données ou réutilisation. Si les technologies et méthodes permettant de gérer la masse de ces données sont de plus en plus disponibles, il reste des verrous importants à cette réutilisation des données en termes d'interopérabilité, d'intégration des données et d'analyse des données. En effet, ces données de vie réelle, n'ayant pas été acquises dans un but de recherche, doivent être extraites des systèmes d'information qui les génèrent, contextualisées, comprises, croisées et partagées. Sur ces

aspects, il existe encore des verrous technologiques et méthodologiques en matière de standardisation, de représentation sémantique et d'interopérabilité des données.

L'utilisation de l'imagerie médicale dans la pratique clinique courante a augmenté à la fois en quantité et en diversité au cours des dernières décennies. Alors que les progrès du matériel et des méthodes (logiciels, traceurs, …) d'imagerie font apparaître de nouvelles façons de visualiser les processus pathologiques, l'imagerie médicale joue un rôle important dans le diagnostic et le traitement des patients dans le cadre des soins cliniques de routine. Par ailleurs, en pratique [3] et en recherche [4] clinique, les informations sur les images sont de plus en plus condensées en un ensemble de mesures quantitatives. Ces mesures, des « biomarqueurs d'imagerie quantitative », permettent par exemple d'évaluer de façon homogène et reproductible l'évolution d'une tumeur en définissant précisément les critères sur les tailles des lésions et leurs évolutions [5].

L'augmentation de la quantité de données d'imagerie dans la pratique clinique de routine et l'importance croissante des biomarqueurs d'imagerie quantitative, laissent entrevoir l'opportunité de l'utilisation secondaire des données d'imagerie clinique acquises en routine pour soutenir la recherche en imagerie médicale [4].

Ma thèse se déroule dans le cadre d'un partenariat entre un laboratoire et une entreprise privée, un laboratoire commun (Labcom) : le LITIS (Laboratoire Interopérabilité Traitement et Intégration des données massives en Santé). Le Labcom LITIS a pour objectif de lever les verrous concernant l'interopérabilité, l'intégration et le traitement des données massives en santé via des innovations méthodologiques et technologiques et de les valoriser sur le plan scientifique et industriel. Pour cela, il associe l'unité de recherche UMR 1099-LTSI, un laboratoire public de l'INSERM qui possède une expertise sur le traitement et l'analyse des données biomédicales ainsi qu'une expertise clinique et la société Enovacom qui est leader dans le domaine des flux de données en santé.

L'équipe DOMASIA du LTSI, dans laquelle j'effectue ma thèse, a développé une solution d'entrepôt de données de santé destiné à la recherche, nommée eHOP, rassemblant les données de tout l'hôpital avec un objectif d'exhaustivité et mettant ces données à disposition des chercheurs. La plate-forme intègre en temps réel les flux de données des applications de soins (prescriptions et administration de médicament, comptes rendus d'urgence, d'imagerie …) et permet aux utilisateur un accès sécurisé, authentifié et tracé aux données propres à leur étude. Les projets menés grâce à eHOP sont la recherche non-interventionnelle (e.g épidémiologie), études de vigilance (e.g suivi thérapeutique en vie réelle), assistance aux professionnels de santé (e.g interprétation des signaux) ou l'évaluation des pratiques cliniques (e.g le suivi des trajectoires de soin). Dans le cadre l'évolution d'eHOP, nous recueillons les avis et besoins des chercheurs sur la solution au fur et à mesure des cas d'usages. Dans le cadre de cette thèse, on s'intéressera aux

perspectives offertes par l'utilisation de l'imagerie dans la recherche et aux besoins actuels des cliniciens de moyens pour accéder, traiter, représenter et analyser ces données d'imagerie. Afin de répondre aux besoins des cliniciens et des chercheurs, le Labcom LITIS va amener au développement de composants logiciels qui pourront être intégrés dans eHOP. Ce dernier est actuellement utilisé par le Réseau Inter-régional des Centres de Données Clinique (Ri-CDC), qui regroupe les 6 hôpitaux universitaires de l'Ouest (dont le CHU de Rennes) et l'institut de cancérologie de l'Ouest (ICO).

Dans ce contexte, l'objectif de ce travail est d'identifier les méthodes et outils pertinents pour extraire, intégrer et traiter les données d'imagerie en santé dans un objectif de réutilisation pour la recherche.

Dans un premier temps, nous ferons un tour d'horizon du domaine de la réutilisation des données d'imagerie de santé. Cette partie permettra de comprendre qui sont les acteurs de cette utilisation secondaire de données et quels sont leurs besoins et les verrous qu'ils rencontrent, chacun à leur niveau de la chaîne de traitement.

Une deuxième partie portera sur le champ de la représentation sémantique des données afin de comprendre comment conserver leur sens et pouvoir les aligner avec des référentiels standards, les rendant ainsi partageables et accessibles aux algorithmes, notamment d'IA. Dans cette partie, nous voulons identifier les verrous et leviers de la représentation sémantique dans le domaine de l'imagerie médicale. Nous proposerons aussi d'aborder le problème de l'alignement des données sur un référentiel standard en employant le raisonnement ontologique.

Enfin, nous montrerons comment nous avons pu mettre en œuvre ces méthodes d'intégration en prenant en compte les verrous et contraintes qu'imposent la gestion des données de santé. L'objectif spécifique de cette dernière partie est de mettre en œuvre un flux optimisé d'intégration sémantique des données d'imagerie dans un entrepôt de données de santé, de l'étude des besoins des chercheurs et cliniciens jusqu'à son évaluation.

# Première partie : L'utilisation secondaire des données d’imagerie en santé

# 1. Introduction

L'utilisation des données de santé hors du domaine du soin lui-même est une utilisation secondaire de données, ou réutilisation. Safran [6] définit l'utilisation secondaire des données comme étant "l'usage non direct pour le soin des informations de santé personnelles pour l'analyse, la recherche, la mesure de qualité/sécurité, la santé publique, le financement, accréditation ou certification de fournisseurs, le marketing et autres domaines incluant les activité strictement commerciales".

Les données ainsi ré-utilisées sont donc observationnelles et rétrospectives, en opposition à la recherche clinique conventionnelle qui repose sur des données collectées prospectivement sur des cohortes prédéfinies. Ces données dites “de vie réelle” sont accessibles à bas coût et en très grande quantité (avec par conséquent une grande puissance statistique) sans interférer avec le soin porté aux patients [7].

Le champ des données réutilisable est très large allant des données médico-administratives (e.g le Système national des Données de Santé en France réunissant des données en provenance d’établissements de santé, de médecins ou de pharmaciens) aux données de soins (par exemple au sein de l'hôpital : comptes rendus d’urgence ou d’opération, prescriptions, administrations, imagerie, etc.) en passant par les données renseignées par les patients eux même [8] (Enregistrements de constantes et d’activités obtenues via des dispositifs personnels (smartphones, smartwatches), données nutritionnelles comme les calories ingérées, etc. )

Par conséquent, les données nécessitent un décloisonnement avant leur réutilisation. Cette étape de décloisonnement passe par l’extraction des données depuis les systèmes qui les génèrent ou les stockent, l'harmonisation des données permettant l’interopérabilité entre les sources et enfin leur intégration dans un système d’exploitation transversale et/ou de partage à plus large échelle [9].

Dans le domaine de l’imagerie en particulier, l’explosion de la quantité de données combinée au développement de marqueurs quantitatifs pour interpréter ces images (contrairement à l’approche historiquement “qualitative” abordée par les radiologues) a créé l’opportunité d'analyse à grande échelle de ces données aujourd’hui cruciales pour le suivi des patients [4]. De plus, le coût et la logistique nécessaire pour employer les nouvelles techniques d’imagerie (d’IRM par exemple) rendent particulièrement intéressant la réutilisation de données plutôt que leur prospection [10].

Dans cette partie, nous allons d'abord nous pencher sur l'importance et l'impact de la réutilisation des données en listant ses enjeux. Nous verrons ensuite les défis qu'implique cet usage secondaire à travers la liste des verrous. Enfin, nous aborderons les solutions actuelles, leurs différentes approches, leurs spécificités et leurs limites.

# 2. Les Enjeux

## 2.1. Le développement d'outils d'aide aux radiologues

Les radiologues ont de nombreuses tâches à effectuer dans le cadre de leur activité et font face à de nombreux défis [11] :

- Augmentation des demandes et de la production d'examens d'imagerie
- Examens nécessitant de plus en plus de temps pour leurs interprétations
- Des équipements de plus en plus rapides et produisant de plus en plus d'informations à analyser
- Une organisation en permanence perturbée (demandes d'examens en urgence, sollicitation de l'expertise non programmée)
- Staff, réunions de concertation pluridisciplinaire (RCP)
- Travaux de recherche ou exigences réglementaires de plus en plus contraignantes.

Dans ce contexte, les techniques d'imagerie diagnostiques par rayon X, IRM et échographie fournissent une grande quantité d'informations que le radiologue ou un autre professionnel de santé doit analyser et évaluer de manière exhaustive en peu de temps [12]. Un des objectifs de la recherche en imagerie est d'aider le radiologue dans ses activités variées.

La Figure 1 présente différentes tâches effectuées par les radiologues et pour chacune, des exemples d'outils développés grâce à l'IA et permettant de la simplifier.

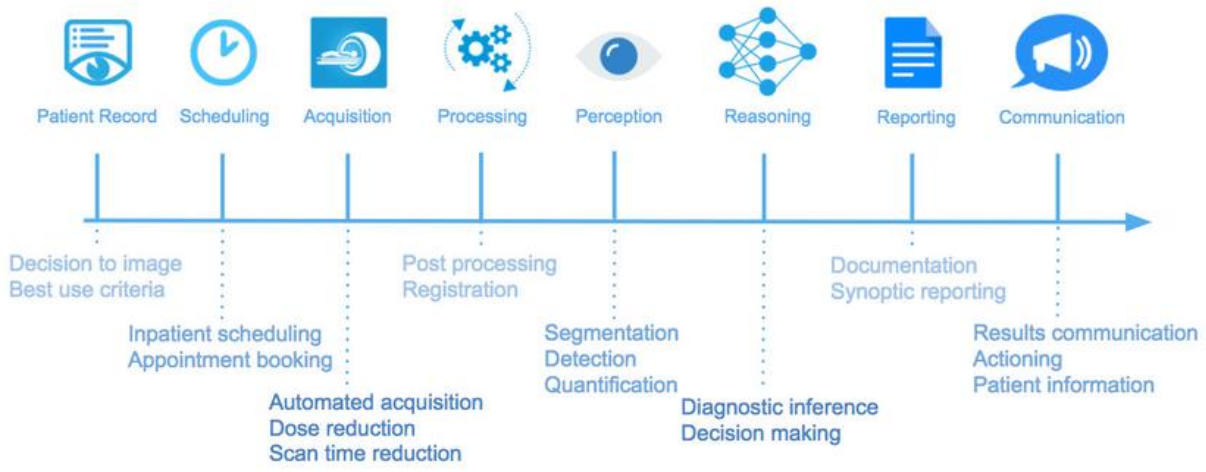



*Figure 1 : Les étapes d'un flux de diagnostic radiologique et les applications potentielles de l'intelligence artificielle (IA) à chaque étape* [13]

Les cliniciens peuvent, par exemple, avoir besoin d'effectuer de la segmentation d'images, c'est-à-dire de définir précisément la forme et/ou le volume des organes ou des signes pathologiques sur les images. L'objectif de cette segmentation d'image est de simplifier et de transformer la représentation d'une image de sorte qu'elle soit plus lisible et plus facile à analyser [14]. La segmentation consiste à extraire des régions d'intérêt (ROI) afin d'identifier des zones anatomiques et les mesurer, par exemple pour positionner virtuellement des implants modélisés par ordinateur (CAO) chez un patient. La figure 2 présente un exemple d'application de l'IA à la segmentation de tumeurs. La segmentation d'image permet aussi de supprimer les détails indésirables d'un scanner, comme l'air, et permet d'isoler différents tissus, comme les os et les tissus mous [15]. Elle peut être utilisée pour des interventions guidées par l'image, la radiothérapie, ou l'amélioration des diagnostics radiologiques [16,17]. La segmentation des images médicales peut être une tâche fastidieuse, et les progrès récents des techniques logicielles d'intelligence artificielle (IA) facilitent l'exécution des tâches en routine. De nombreux travaux portent sur la conception d'algorithmes de segmentation automatique [16,18].

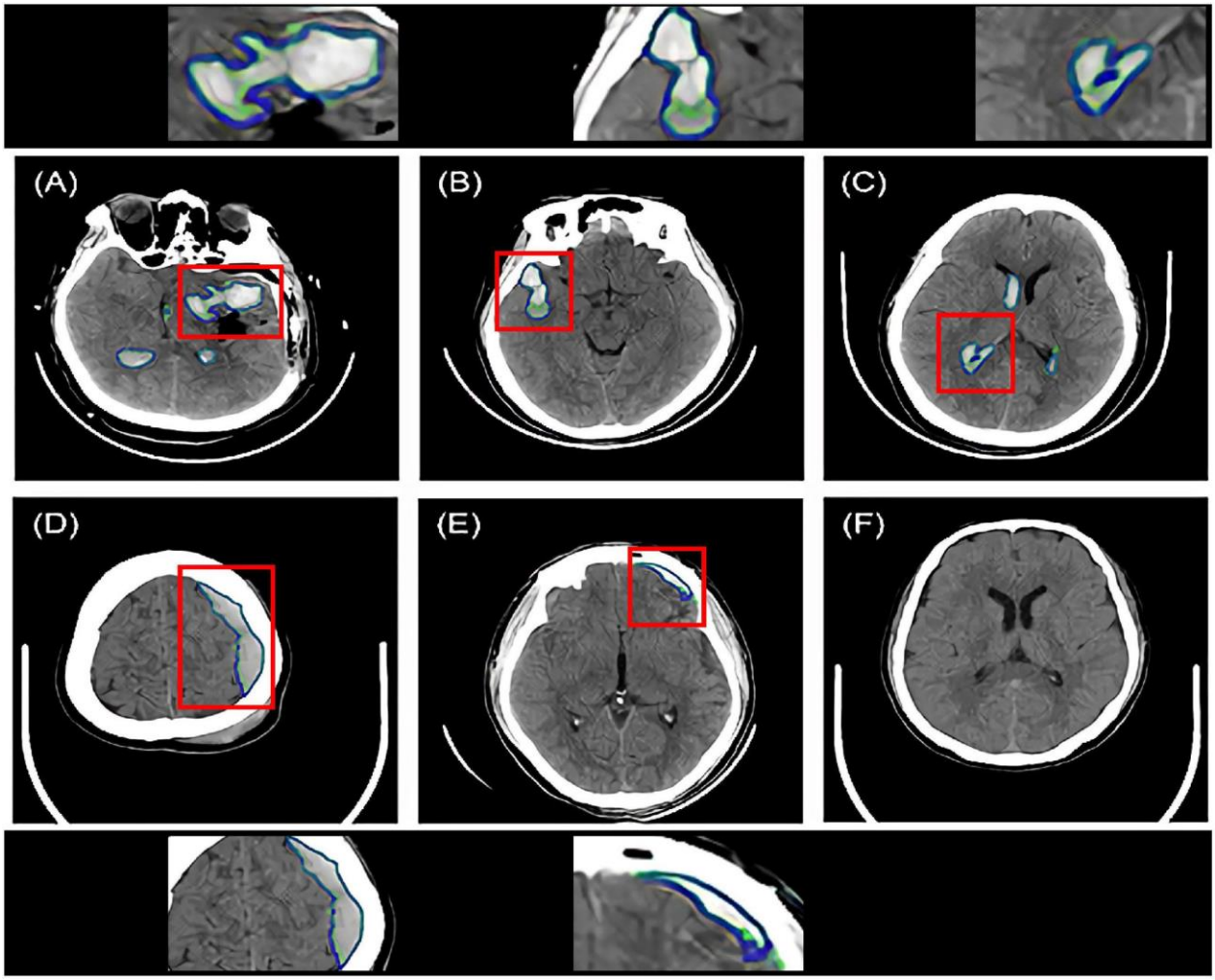



*Figure 2 : Exemples de segmentation d'hémorragie intracrânienne effectuée par l'algorithme de Jun Xu et al [19] à partir de six patients représentatifs. Les zones de l'hémorragie segmentée par trois neuroradiologues certifiés sont représentées en vert, tandis que les segmentations du modèle sont représentées en bleu. Les cases rouges indiquent les régions de segmentation agrandies. Le modèle apprend à trouver les contours des hémorragies et à quantifier leurs volumes*

Plusieurs études ont démontré que l'utilisation de systèmes d'IA aidant à la lecture et l'interprétation des images peut améliorer l'efficacité du radiologue en termes de temps, de qualité du diagnostic (meilleures sensibilité et spécificité), par exemple pour la détection du cancer du sein sur des mammographies [20–26].

L'intelligence artificielle commence à démontrer son intérêt pour aider les cliniciens dans l'analyse sémiologique et le diagnostic lui-même [27]. La détection assistée par ordinateur (Computer Assisted Detection, CADe), également appelée diagnostic assisté par ordinateur (Computer Assisted Diagnostic, CADx), sont des systèmes qui assistent les médecins dans l'interprétation des images médicales [28]. La communauté des radiologues voit donc dans les outils développés par l'IA un support pour faire face à de nombreux défis [11].

## 2.2. Faire avancer la médecine personnalisée

A partir des années 2000, la médecine a pu tirer parti des progrès techniques pour s'inscrire dans une démarche améliorant la prise en charge des pathologies dans le cadre de la médecine 4P [29], c'est à dire prédictive, préventive, personnalisée et participative (par opposition à l'approche réactive consistant à agir au fur et à mesure de l'apparition et de l'évolution de la pathologie).

L'aspect "personnalisé" de cette nouvelle médecine repose sur une stratification des patients, c'est-à-dire l'analyse d'un nombre toujours plus grand de caractéristiques des patients et de leurs pathologies pour les classer dans des catégories toujours plus spécifiques et les traiter de la façon la plus appropriée. La génétique joue un rôle très important dans ce paradigme, c'est-à-dire la définition de profils génétiques. L'imagerie permet, elle, de dresser le profil des traits observables (caractéristiques physiques à notre échelle ou à l'échelle cellulaire ou encore moléculaire) appelé "profil phénotypique" des patients ou de leurs pathologies [30] grâce à des mesures quantitatives des caractéristiques des patients, les biomarqueurs. La Figure 3 présente les étapes types de l'extraction de biomarqueurs

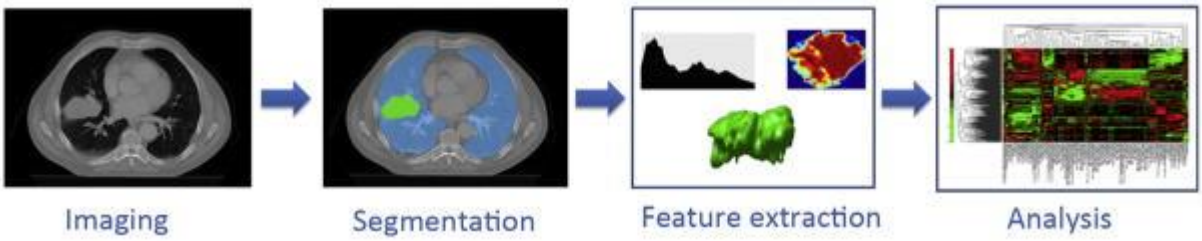



*Figure 3 : Le flux de travail de la radiomique. A partir des images médicales, une segmentation est effectuée pour définir la région d'intérêt, ici une tumeur. Les caractéristiques (basées sur l'intensité, la texture et la forme de la tumeur) sont extraites puis sont utilisées pour leur pouvoir pronostique, ou liées au stade, ou à l'expression génétique [31]*

Le National Institute of Health (NIH) défini un biomarqueur comme étant "une caractéristique qui est objectivement (c'est à dire avec une précision et reproductibilité suffisantes) mesurée et évaluée comme un indicateur de processus biologiques normaux, ou pathologiques ou de la réponse biologique à une intervention thérapeutique [32]. Il existe trois types de biomarqueurs :

- Paramètres biochimiques histologiques détectés sur un échantillon tissulaire obtenus lors d'une biopsie ou d'une chirurgie
- Paramètres biochimiques ou cellulaires détectés dans le sang ou l'urine
- Paramètres anatomiques, fonctionnels ou moléculaires détectés par imagerie médicale : l'imagerie quantitative

Ce sont ces derniers qui nous intéressent dans ce travail, dont la Figure 4 donne plusieurs exemples concrets. Ces biomarqueurs d'imagerie sont rassemblés en 3 catégories :

- Paramètres anatomiques, par exemple les critère RECIST (Response Evaluation Criteria in Solid Tumors) qui sont basés sur la mesure du plus grand diamètre des lésions [33].
- Paramètres fonctionnels, ils permettent l'estimation des processus biologiques comme la prolifération, l'hypoxie, l'angiogenèse, l'envahissement et la mort cellulaire [34]. Par exemple La mesure SUV (Standardized Uptake Value) quantifie la fixation tissulaire du traceur radioactif au sein du tissu tumoral dans une région d'intérêt
- Paramètres moléculaires. De nouvelles méthodes d'acquisition et de traitement permettent la compréhension des pathologies à l'échelle moléculaire [35].

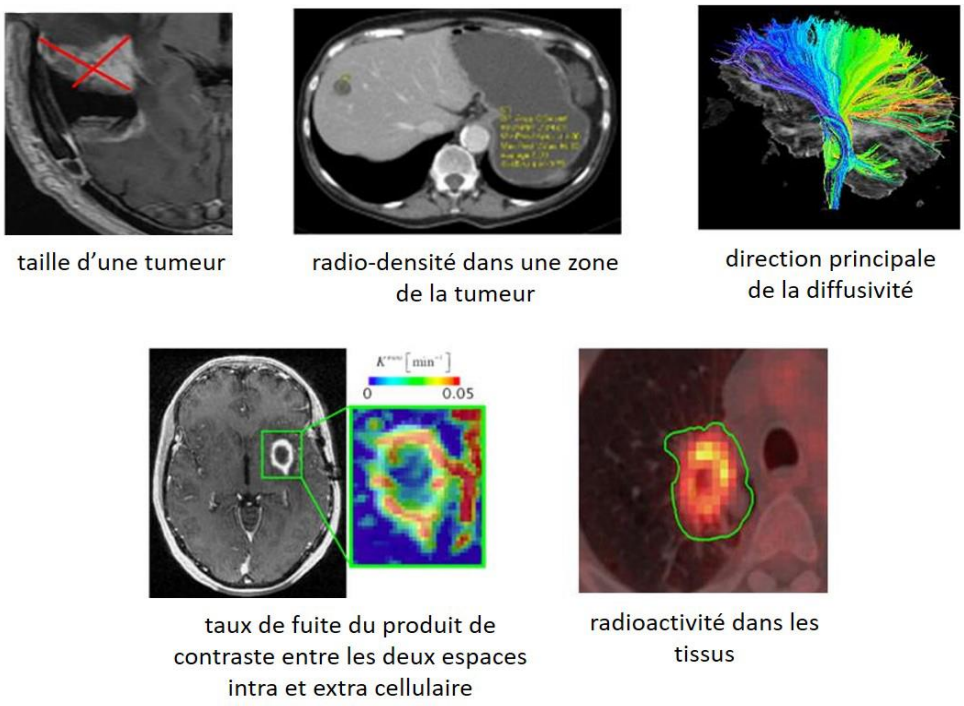



*Figure 4 : Exemples de caractéristiques numériques mesurées en imagerie quantitative* [36]

Ces indicateurs peuvent ensuite être liés au profil génétique de la tumeur et à son anatomopathologie. Comparés à grande échelle, ces biomarqueurs devraient permettre d'anticiper la réponse au traitement en administrant, pour un patient et un profil tumoral donné, le meilleur traitement, puis d'en suivre l'évolution selon les biomarqueurs pertinents, au-delà des mesures de dimensions actuelles [11].

L'amélioration de la médecine de précision, ou personnalisée, est donc un enjeu de l'imagerie quantitative et de l'utilisation des biomarqueurs, dans le soin comme dans la recherche.

## 2.3. La validation des modèles in silico

La modélisation in silico, dans laquelle des modèles informatiques sont développés pour modéliser un processus pharmacologique ou physiologique, est une évolution numérique de l'expérimentation in vitro contrôlée. La modélisation in silico combine les avantages de l'expérimentation in vivo et in vitro, sans avoir à se soumettre à certaines considérations éthiques et au manque de contrôle associés aux expériences in vivo [11,37]. Les essais cliniques in silico mettent en place des cohortes virtuelles ou des études de cas afin de tester la sécurité et l'efficacité des interventions médicales en utilisant des modèles informatiques

simulant les patients et/ou les maladies. Ces essais cliniques in silico peuvent exploiter les données cliniques du monde réel pour représenter par exemple des formes anatomiques 3D ou des résultats cliniques [38]. Le projet Européen PRIMAGE [39], qui a pour objectif de faciliter la prise de décision dans la gestion clinique de cancers pédiatriques en développant des modèles d'évolution des tumeurs, est un bon exemple de développement de modèle in-silico utilisant, entre autres, des données rétrospectives. Ces modèles seront construits en utilisant des biomarqueurs et des paramètres dynamiques extraits d'échographies, de scanners ou d'IRM acquis dans le contexte du soin [40].

Même dans les cas où ces modèles in-silico sont construits sans données provenant du soin, il est parfois nécessaire d'utiliser des données de vie réelle pour identifier les divergences avec les modèles et ainsi corriger ces derniers [41].

On retrouve cette approche avec les jumeaux numériques [42]. Les jumeaux numériques de patients (il existe des jumeaux numériques d'organes précis, de pathologie, de virus, …) peuvent être constitués à partir des données de vie réelles [43] et permettent de tester virtuellement un traitement ou une intervention sur un patient, cf figure 5. Il est aussi possible de créer un modèle d'un organe, comme par exemple un cœur [44], à partir de données d'un grand ensemble de patients [45] (ici encore, potentiellement des données de vie réelle, dont l'imagerie) puis, à partir des données d'un patient, personnaliser ce modèle pour en faire le double numérique de son cœur afin de faire des tests in silico.

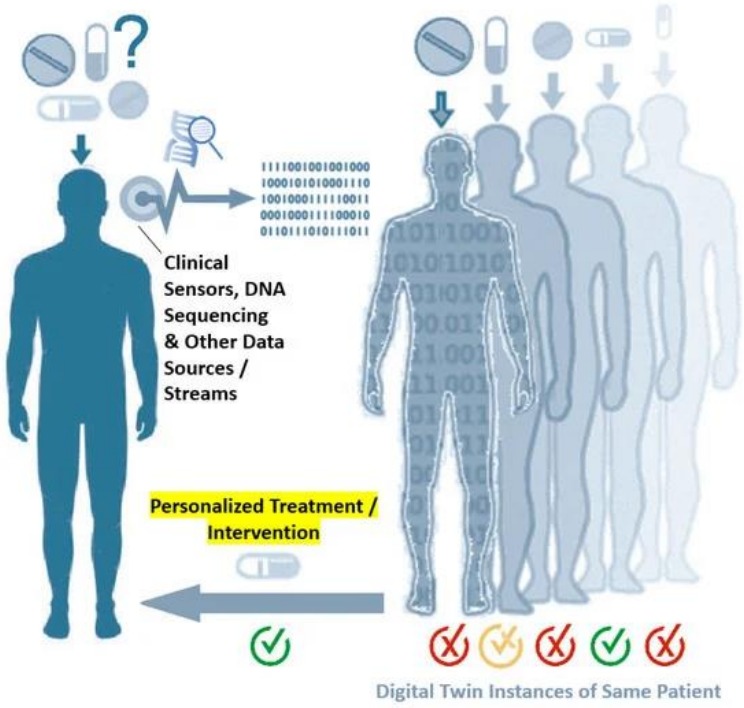



*Figure 5 : Plusieurs instance du jumeau numérique d'un patient peuvent être utilisés pour tester les réponses du patient réel aux traitement et intervention envisageables* [43]

## 2.4. L'amélioration de l'expérience des patients

Les techniques d'imagerie évoluent afin d'obtenir des images plus faciles à analyser, plus rapidement et avec moins d'effets nocifs sur le patient [11,46]. Par exemple, lors de reconstructions scanographiques, des algorithmes permettent de diminuer le bruit dans l'image. Plusieurs logiciels basés sur l'IA permettent des durées d'acquisitions plus courtes et des injections de produit de contraste plus faibles grâce à des post-traitements appliqués sur les images [47,48]. Les données de vie réelles sont indispensables pour vérifier que ces stratégies de réduction de dose n'impactent pas la qualité des images [49].

## 2.5. La création de cohorte

Les études de cohorte sont un suivi dans le temps d'un ou plusieurs groupes de patients [50]. En pratique, on établit un groupe de sujets exposés au facteur de risque et un groupe non exposé au facteur de risque. Les deux groupes vont être suivis (i.e., études longitudinales)

puis comparés vis à vis d'un critère de jugement (décès, apparition d'une pathologie, etc) [51]. On parlera de cohorte historique (ou rétrospective) si la survenue de l'exposition au(x) facteur(s) et de la maladie, a déjà eu lieu au moment où le chercheur débute son enquête. L'utilisation secondaire des données de vie réelles (dont les données d'imagerie) permet la création de telles cohortes rétrospectives. Par exemple, dans [52], les scanners des poumons de patients ayant un diagnostic confirmé de COVID-19 sont extraits et utilisés pour étudier la capacité des caractéristiques des images à prédire si l'infection sera stable ou progressive (ces caractéristiques sont des features d'imagerie utilisés par ailleurs pour évaluer l'hétérogénéité de cancers) .

## 2.6. Les essais cliniques

Les essais cliniques évaluent la sécurité et l'efficacité d'un produit de santé (médicament, dispositif ou thérapie cellulaire et génique) chez des volontaires sains ou malades. Si son efficacité est prouvée, et s'il est sans danger, le médicament pourra obtenir une autorisation de mise sur le marché (AMM) délivrée par l'Agence nationale de sécurité du médicament et des produits de santé (ANSM) [53]. Ces travaux sont prospectifs et les examens (d'imagerie par exemple) effectués dans ce cadre ont un objectif de recherche avant tout. Cependant, les données de vie réelles peuvent par exemple être utilisées pour trouver les patients éligibles à l'essai.

## 2.7. Le partage à large échelle

De nombreux projets visent à rassembler des données et proposent de rassembler les données de façon centralisée, de travailler sur des données agrégées, ou encore d'amener les algorithmes vers les données par une approche fédérée. L'utilisation secondaire des données à large échelle est fondamentale pour la recherche et tous les pays développés mettent en place des stratégies afin de collecter et tirer parti de ces masses d'informations [54].

A l'échelle nationale, le Health data hub [55] a pour objectif de garantir l'accès aisé et unifié, transparent et sécurisé, aux données de santé pour améliorer la qualité des soins et l'accompagnement des patients. De nombreux projets et "data challenges" portés par le HdH concernent l'imagerie. En 2022, 5 des 6 "data challenges" proposés mobilisent des données d'imagerie [56].

Les sujets de recherches précis, comme la neuro-imagerie par exemple, sont tournées vers le partage de données afin de mener des études avec une quantité de données ayant suffisamment de puissance statistique. Cette approche est par exemple concrétisée par le projet France Life Imaging [57] à l'échelle nationale, dont le nœud "Analyse et gestion de l'information" a pour objectif la mise en place d'une infrastructure logicielle et matérielle de gestion des données d'imagerie médicale en provenance de différents centres de recherche clinique, et utilisera les installations de stockage de données et de traitement de l'information déjà existantes, c'est-à-dire CATI, Shanoir et ArchiMed.

# 3. Les verrous

## 3.1. La quantité de données

Les éléments de base des données d'imagerie médicale sont les examens d'imagerie, constitués du contenu de(s) (l')image(s), des métadonnées et du compte rendu de l'examen. L'ensemble de données mobilisé pour une étude doit comporter suffisamment d'examens d'imagerie pour répondre à la question posée [46,58]. Le développement d'algorithmes d'IA par exemple, nécessite des ensembles de données importants pour que ces algorithmes soient performants, généralisables et statistiquement fiables, de l'ordre de centaines de milliers ou de millions [17,59].

La masse de données brutes est aujourd'hui immense et leur gestion posent des problèmes techniques de stockage et de gestion des flux de données [12,46]. Cependant, pour le développement des algorithmes, l'ensemble de données lui-même et chaque examen d'imagerie doivent être décrits et étiquetés avec précision en fonction de l'étude. Dans le cas de l'apprentissage supervisé, le "gold standard", c'est-à-dire l'information à faire deviner au modèle et qui doit être connue pour l'ensemble des données d'apprentissage et de test, doit être aussi précis et reproductible que possible [60]. Nous allons aborder ce point avec l'annotation et la qualité des données.

## 3.2. La qualité des données

### 3.2.1. La segmentation des images

Dans l'imagerie diagnostique, l'information brute se trouve sous la forme d'une trame de pixels ou de voxels mais une représentation de haut niveau (tissus, organes, lésions, etc.) est souhaitable pour estimer des descripteurs numériques (tailles, volumes des organes, valeurs de contrastes et textures, etc), tels que ceux décrits pour les domaines de la radiomique vus précédemment [61]. Il est donc nécessaire d'obtenir autant d'annotations cliniques que possible, sous la forme de segmentations manuelles et de délimitations de régions d'intérêt (Region Of Interest, ROI), pour l'entraînement des modèles selon des techniques d'apprentissages automatiques supervisés [62]. Ces annotations d'imagerie médicale sont

chronophages, coûteuses et doivent être réalisées par des spécialistes capables de comprendre et d'interpréter les images [61,63]. Dans la mesure où l'annotation ne rentre pas dans la finalité première de l'usage et de la collecte de données (c'est à dire le soin), elle n'est pas pour le clinicien une priorité [62].

### 3.2.2. Des métadonnées pauvres ou difficile à enrichir

Au cours du soin, une grande partie des données est recueillie de façon non structurée (tels que les comptes rendus d'examens d'imagerie le plus souvent dictés) et les quelques données médicales structurées et codées (selon des terminologies de références) au cours des activités d'un département de radiologie ont surtout pour vocation de permettre la facturation des actes [62] (usage de la CIM10 et de la CCAM pour générer le PMSI). La norme DICOM, utilisée quasi universellement pour structurer et échanger les données d'imageries comprenant les images et métadonnées, et sur laquelle nous reviendrons plus en détail, est riche en spécifications qui sont sous-utilisées par les constructeurs d'équipements d'imagerie [60]. Cela rend les données moins faciles à trouver (Findable), car les métadonnées accompagnant les images sont alors moins riches (c'est à dire moins détaillées et exhaustives concernant les informations descriptives du contexte, leur qualité et leurs caractéristiques [64]).

Il est aussi possible d'extraire des informations structurées depuis les comptes rendus d'imagerie rédigés en texte libre avec des approches d'analyse sémantique (Natural Language Processing, NLP) [46]. Si un grand nombre de solutions ont été développées pour la langue anglaise, elles sont moins nombreuses dans les autres langues [65].

Dans la partie 2 de cette thèse, nous proposons une approche pour pallier le manque d'information dans les données d'imagerie brutes (DICOM) en utilisant le raisonnement ontologique sur les données disponibles.

## 3.3. La nature sensible des données

En Europe, le Règlement général de protection des données RGPD [66], appliqué depuis le 25 mai 2018, encadre l'utilisation des données à caractère personnel. Ces données personnelles de santés sont définies par le RGPD comme étant "les 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” [67]. Les études sur des données de santé doivent utiliser des données anonymisées ou se conformer au RGPD.

### 3.3.1. Anonymisation et pseudonymisation

Le RGPD définit l’anonymisation comme étant “un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible" [68]. Il s’agit d’une solution, parmi d’autres, pour pouvoir exploiter des données personnelles dans le respect des droits et libertés des personnes permettant ainsi aux acteurs d’exploiter et de partager leur « gisement » de données sans porter atteinte à la vie privée des personnes. Le bon fonctionnement de la méthode d’anonymisation appliquée incombe au responsable de traitement à l’initiative de l’opération. Concernant l’imagerie médicale, il est nécessaire de réaliser une analyse d’impact relative aux risques de réidentification des patients car l’anonymisation des images n’est pas aujourd’hui bien définie juridiquement [46]. L’anonymisation ne doit pas être confondue avec la pseudonymisation qui est, elle, réversible et consiste à remplacer les données directement identifiantes (nom, prénom, etc.) d’un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.). Les individus ne sont plus identifiables directement mais les données concernées conservent tout de même un caractère personnel. La pseudonymisation constitue une des mesures recommandées par le RGPD pour limiter les risques liés au traitement de données personnelles. La Figure 6 résume la différence entre les deux notions d’anonymisation et de pseudonymisation.

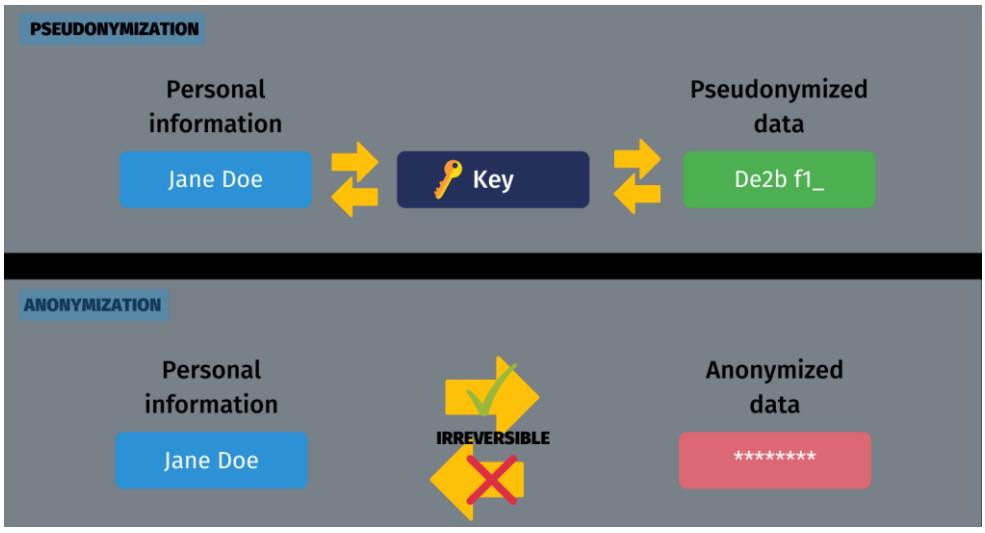



*Figure 6 : Différence entre pseudonymisation et anonymisation [69]*

### 3.3.2. Anonymisation en imagerie

Dans le cas de l'imagerie, l'anonymisation passe par la suppression, dans les comptes-rendus, des informations potentiellement identifiantes par des approches de traitement du langage [70,71]. Les métadonnées des images sont aussi à traiter afin de retirer les champs d'informations personnelles des patients ou les champs privés (dont seuls les constructeurs de l'appareil d'imagerie utilisé peuvent lire le contenu).

Les images elles même peuvent aussi contenir des données identifiantes. D'une part, certaines informations sont parfois incorporées à l'image comme le nom du patient sur des ultrasons par exemple. Il existe des approches de détection et reconnaissance des zones de texte sur une image (Optical Character Recognition, OCR) qui permettent de masquer les éléments textuels dans les images [72,73]. D'autre part, certaines caractéristiques physiques des patients, visibles sur les acquisitions d'imagerie, sont identifiantes. Pour le visage en particulier, il existe des techniques de suppression, modification ou remplacement des caractéristiques faciales, telles que le "defacing" (retirer la partie du visage sur une image 3D de la tête d'un patient), la "suppression du crâne", le "masquage/brouillage du visage" ou encore le remplacement du visage par un visage "par défaut" [74]. Cependant, ces méthodes limitent les analyses possibles qui peuvent être effectuées sur les données, en particulier pour les procédures qui tirent parti des géométries de la tête (comme la segmentation automatique) [75]. Les méthodes d'anonymisation trouvent aujourd'hui leurs limites avec les IRM fonctionnelles du cerveau, qui permettent de visualiser le connectome (un plan complet des connexions neuronales d'un cerveau [76]). En effet, plusieurs travaux ont pu démontrer qu'il est possible d'extraire des caractéristiques discriminantes de ces IRM fonctionnelles encéphaliques et donc de ré-identifier les patients [77,78].

Nous aborderons l'anonymisation des données d'imagerie respectant le RGPD dans la partie 3 de ce manuscrit, lors de l'intégration des données à un entrepôt.

## 3.4. Le cloisonnement

Les études multicentriques nécessitent de collecter des données de domaines différents (clinique, radiologique, biologique, etc). Ces données, réparties physiquement et utilisant des référentiels différents, doivent être décloisonnées et alignées sur des référentiels communs.

### 3.4.1. Rassembler les données

Afin de concevoir des modèles de prédiction applicables largement, il faut obtenir des données couvrant l'ensemble de la population cible. Certains algorithmes d'IA par exemple, sont très performants pour l'aide au diagnostic dans le contexte où ils ont été développés, sur des données d'imagerie bien harmonisées, mais se montrent peu efficaces quand ils sont utilisés dans des populations différentes [79,80]. Par exemple, des algorithmes de détection de cancer du sein développés avec une base de données chinoise, montrent un important taux d'échec sur une population européenne [11]. Il existe de nombreuses autres sources de biais possible lorsque l'on se base sur un ensemble trop restreint de données d'imagerie [17] : différences d'âge, les proportions d'origines ethniques et de sexe, l'utilisation d'équipements d'imagerie différents (fournisseurs, types d'images, protocoles d'acquisition), la prévalence des maladies ou même des variations dans les pratiques locales. Pour obtenir des modèles généralisables à large échelle, il faut donc exploiter des données représentatives en mobilisant plusieurs sources distinctes et faire face aux contraintes légales, techniques ou de sécurité qui représentent un cloisonnement physique.

Par ailleurs, il est en général nécessaire pour mener une étude de croiser des données de différentes natures. Au sein même d'un établissement, les données brutes ne sont pas directement utilisables depuis le système d'information clinique. La mise en place d'une étude en imagerie consiste à identifier les patients pertinents pour l'étude, extraire les informations -souvent non structurées- de leurs dossiers puis extraire les images adéquates du serveur d'images (Picture Archiving and Communication System, PACS). Ces silos de données doivent être ouverts afin de permettre aux chercheurs un accès commun aux données de natures différentes (compte-rendu, prescription, imagerie, …).

Dans la partie 3, nous aborderons le problème du désilotage des données d'imagerie en présentant une solution pour intégrer ces données d'imagerie dans un entrepôt de données de santé.

### 3.4.2. La variabilité des standards et de leur usage

Afin de réutiliser des données d'imagerie de différentes sources, il faut nécessairement que celles-ci soient (ou soient rendues) interopérables c'est-à-dire que, d'une façon ou d'une autre, elles sont rendues homogènes en termes de sémantique et de format [79].

Si DICOM permet de normaliser universellement le format des données d'imagerie, en définissant une liste d'attributs relatifs à l'examen, respectant des standards syntaxique et

sémantiques (SNOMED-CT, LOINC ou des listes de valeurs définies par DICOM), de nombreux attributs DICOM restent optionnels et ne sont pas utilisés en pratique.

De plus, des informations importantes sont souvent indiquées uniquement dans des champs de "description" en texte libre [81]. Les fournisseurs de matériel d'imagerie utilisent souvent des annotations et des structures de données propriétaires (dans des champs DICOM privés), et qui de fait, ne sont pas interopérables [82]. Même le contenu de certains attributs publics, dont DICOM impose qu'ils soient renseignés, peut varier selon les constructeurs. Par exemple, l'attribut du type d'image ("ImageType"), a pour valeur un ensemble de plusieurs éléments relatif à des caractéristiques de l'image dont les trois premiers sont bien décrits et restreints par la norme DICOM mais les constructeurs peuvent ensuite ajouter dans cette liste des valeurs qui leur sont propres et n'ont de sens que pour leurs systèmes [81].

De nombreuses ressources sémantiques existent pour décrire les données d'imagerie selon certaines techniques (e.g microscopie[83]), pathologies (e.g la maladie d'Alzheimer[84]) ou centrées sur les biomarqueurs d'imagerie (e.g Quantitative Imaging Biomarker Alliance (QIBA)[85]). Ces ressources ne sont pas toujours facilement interopérables entre elles comme nous le verrons dans la suite de ce travail. A cela s'ajoute la barrière de la langue car ces ressources sont conçues an anglais, pour certaines partiellement traduite vers d'autres langues.

On propose de contribuer à lever ce verrou dans la partie deux de ce manuscrit en étudiant les solutions d'alignement de terminologie locale sur une terminologie de référence ou de classification des examens.

Synthèse

L'utilisation secondaire des données d'imagerie de santé nécessite de pouvoir extraire les données depuis le PACS en conservant leur sens et d'aligner les données sur les référentiels qui permettront de croiser des données avec d'autres domaines et/ou de partager ces données à plus large échelle. Il est donc crucial de pouvoir décrire précisément et formellement la signification des données collectées pour garantir leur valeur (en conservant le contexte de la donnée) et pour les rendre accessibles aux machines qui pourront ainsi les appréhender et permettront de les comparer, les partager et les analyser automatiquement.

# Deuxième partie : Méthodes pour l'intégration sémantique de l'imagerie en santé

# 1. La représentation sémantique de l'imagerie en santé

La représentation de données de santé, l'imagerie et l'interopérabilité sont des domaines vastes et en constante évolution, dont les écosystèmes sont donc très fournis. Afin de pouvoir appréhender au mieux les travaux présentés dans cette thèse, il est nécessaire de faire un tour d'horizon des outils et solutions mis en place aujourd'hui pour répondre aux enjeux et essayer de lever les verrous que nous avons vu dans la partie précédente.

## 1.1. Définitions

Une terminologie est un vocabulaire composé de termes utilisés dans un domaine spécifique [36]. Chaque terme est associé à une définition textuelle, les algorithmes n'ayant donc pas accès à la sémantique de ces termes. Une terminologie structurée hiérarchiquement est une taxonomie (ou hiérarchie de type). Selon W.Ceusters [86], les terminologies servent à la communication entre humain et entre les humains et les machines, quand les ontologies - que nous verrons par la suite - servent à représenter la réalité pour les machines et à la communication entre elles.

### 1.1.1. Terminologie d'interface

Les terminologies d'interface (TI) sont des "collections systématiques de phrases (termes) liées au soin qui représentent les informations des patients rentrées par les cliniciens dans les logiciels métier". Les TI sont par exemple utilisées pour la saisie informatisée des prescriptions des médecins : les médicaments, les tests et les procédures pouvant être commandés sont nombreux et les termes se rapportant à ces éléments peuvent être non intuitifs. Dans le contexte d'une spécialité médicale, utilisant des acronymes, des termes spécifiques, il faut que les utilisateurs locaux s'entendent sur un ensemble de termes et de définitions, ce que l'utilisation d'une TI peut résoudre [87].

Les termes des TIs sont les éléments constitutifs des notes cliniques et sont également utilisés comme valeurs textuelles pour la saisie de données structurées [88] par exemple dans

des formulaires du dossier patient informatisé. Cependant, ces terminologies ne sont pas adaptées à la recherche mais plus au soin :

- Ces termes doivent être manipulables très facilement par les cliniciens dans la pratique et sont donc aussi courts que possible et parfois ambigus hors contexte. Les abréviations et les acronymes jouent un rôle majeur.
- Des terminologies d'interface différentes peuvent donner des sens différents à un même terme. En effet, les significations des termes peuvent être différentes selon les groupes d'utilisateurs, les spécialités médicales ou les dialectes régionaux. Pour un gynécologue une IVG sera une Interruption Volontaire de Grossesse, et pour un cardiologue une Insuffisance Ventriculaire Gauche. La signification des termes des TI peut aussi changer au fil du temps.

### 1.1.2. Terminologie de Référence

Il existe de très nombreuses terminologies en santé. Le Unified Medical Language System (UMLS), un réseau sémantique rassemblant de nombreux vocabulaires contrôlés dans les sciences biomédicales [89], compte aujourd'hui 222 terminologies ou vocabulaires différents [90]. Certaines de ces terminologies sont des terminologies de référence, définies par Rosenbloom et al comme des "terminologies conçues pour fournir des représentations exactes et complètes des connaissances d'un domaine donné, y compris ses entités et ses idées, ainsi que leurs interrelations, et sont généralement optimisées pour prendre en charge le stockage, la récupération et la classification des données cliniques" [91].

L'objectif des TR de vouloir couvrir l'ensemble de leur domaine entrave leur utilisation par les cliniciens et ne les préserve pas d'un manque de termes dans des domaines de connaissances spécifiques, c'est pourquoi elles ne permettent pas de répondre aux besoins des cliniciens de structurer les données dans les dossiers médicaux électroniques [87].

Au contraire des TI, les TR doivent fournir des unités de représentation bien définies (les "concepts", "classes", "descripteurs" …). La stabilité d'une TR repose sur des labels non ambigus ainsi que des définitions, des liens vers des normes externes et des définitions ontologiques formelles, généralement basées sur des logiques de description. La terminologie de référence SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) par exemple est décrite sous forme d'ontologie [92,93] et LOINC (Logical Observation Identifiers Names and Codes) décrit l'ensemble de ce que l'on peut « tester, mesurer ou observer concernant le patient » [94]. De plus, il a été montré que l'utilisation de logique de description formelle facilite le mapping avec d'autres terminologies [95] ce qui correspond au rôle des terminologies de référence.

### 1.1.3. Ontologie

Le mot “ontologie” provient de la philosophie et désigne la “philosophie de l’être”. Studer [96] rassemble les définitions de T.Gruber [97] et Borst [98] et définit l’ontologie comme étant “une spécification formelle et explicite d’une conceptualisation partagée”, cette définition met l’accent sur trois aspects très importants des ontologies. Les ontologies proposent des définitions **explicites**, signifiant que les concepts et les relations sont définies de façon déclarative. Elles sont **formelles**, donc décrites sans ambiguïté et interprétable par une machine et ont pour but de représenter une vision **partagée**, donc consensuelle d’un domaine et d’être transmises et rassemblées avec d’autres ontologies pour fusionner les représentations de différents domaines.

Selon Werner Ceusters [86], une ontologie est une représentation d'un domaine de réalité préexistant qui :

1. Reflète les propriétés des objets dans son domaine de telle manière qu'il existe une corrélation systématique entre la réalité et la représentation elle-même,
2. Est intelligible pour un expert du domaine
3. Est formalisée d'une manière qui lui permet de supporter le traitement automatique de l'information

Par exemple, pour relier les concepts d’index, de doigt et de main, on pourra définir les relations suivantes : un index est un doigt (relation de subsomption), un doigt fait partie de la main ; par ailleurs, il en découle qu’un index fait partie de la main. Créer une ontologie demande de caractériser les propriétés essentielles des objets [99]. Guarino [100] propose de distinguer trois catégories d’ontologies :

- Les ontologies de haut niveau (ou ontologies fondamentales) fournissent les concepts de base pour la représentation du savoir indépendamment d'un domaine particulier. Les entités manipulées à ce niveau sont les objets, les qualités, les processus, etc.
- Les ontologies de domaine décrivent le vocabulaire d’un domaine particulier comme la médecine, ou des domaines plus spécifiques comme la radiologie.
- Les ontologies d'application décrivent un modèle particulier utilisé par des professionnels d'un domaine pour réaliser leur activité.

## 1.2. Les ressources pour la représentation sémantique de l'imagerie en santé

Il existe une variété d'outils et de technologies pour accéder aux ressources de contenu liées à l'imagerie et les utiliser. Dans le domaine biomédical, il existe deux grandes familles de formats de représentation des connaissances : UMLS (Unified Medical Language System) et la fonderie OBO.

### 1.2.1. L'UMLS

Le système de langage médical unifié (UMLS) est une bibliothèque de plus de 200 ressources de contenu biomédical conçue et gérée par la National Library of Medicine (NLM). Développé avant la démocratisation des ontologies, L'UMLS contient des terminologies et des systèmes de classification. L'UMLS contient un méta-thésaurus qui aligne les différents termes et codes provenant de nombreux vocabulaires, notamment CPT, ICD-10-CM, LOINC, SNOMED CT. Le méta-thésaurus cherche à assurer l'intégration sémantique des concepts entre les ontologies et les vocabulaires ; un seul identifiant unique de concept (CUI) dans le méta-thésaurus UMLS peut renvoyer à des concepts dans plusieurs vocabulaires qui le composent.

L'UMLS contient aussi le "réseau sémantique" qui spécifie les relations qui relient les principaux groupements des types sémantiques (organismes, structures anatomiques, etc.) et du lexique SPECIALIST, contenant plus de 200 000 termes et utilisé pour aider au traitement du langage naturel.

### 1.2.2. La fonderie OBO et l'ontologie fondatrice BFO

Le NCBO (National Center for Biomedical Ontology) est un consortium de biologistes, cliniciens, informaticiens et de spécialistes des ontologies qui promeuvent le développement du web sémantique [101] et développent des technologies permettant de gérer des informations et des connaissances biomédicales via des ontologies, de sorte que les connaissances et les données sont sémantiquement interopérables et utiles pour faire progresser la science biomédicale et les soins [102]. Les logiciels et les technologies du NCBO comprennent entre autres le BioPortal, et la "OBO foundry".

La fonderie OBO favorise le développement collaboratif d'ontologies biomédicales, couvrant les domaines allant des molécules aux organismes, en passant par le niveau de la cellule. Ces ontologies sont développées autour de l'ontologie fondatrice BFO (Basic Formal Ontology) et de RO (Relation Ontology) [103] pour l'établissement des liens entre les concepts. BFO a pour but de soutenir les ontologies de domaine développées pour la recherche scientifique [104,105] et ne contient pas de termes physiques, chimiques, biologiques ou autres qui relèveraient d'un domaine des sciences en particulier [106]. La Figure 7 illustre l'organisation de BFO.

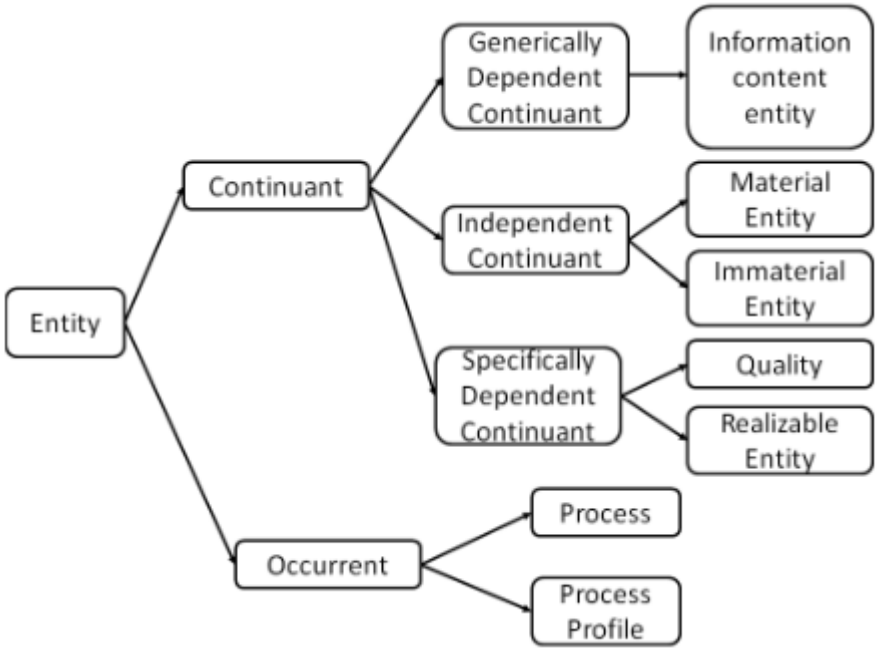



*Figure 7 : Les classes de haut niveau de BFO [106]*

Le BioPortal est une application Web, accessible via des services web RESTful, pour la recherche, le partage, la visualisation et l'analyse des ontologies biomédicales, de terminologies et d'annotations basées sur des ontologies. BioPortal permet d'établir des correspondances entre les termes, d'exporter en masse les correspondances, de visualiser les termes et les relations au sein des ontologies, de prendre des notes et de naviguer dans plusieurs ontologies par le biais d'onglets [107]. BioPortal donne accès à plus de 600 ontologies et terminologies biomédicales, dont SNOMED CT, ICD, CPT, FMA et RadLex [108].

### 1.2.3. La terminologie SNOMED CT

La SNOMED CT (SNOMED Clinical Terms) est une collection organisée de termes médicaux fournissant des codes, des termes, des synonymes et des définitions utilisés dans la documentation et les rapports cliniques. La couverture complète de la SNOMED CT comprend : les résultats cliniques, les symptômes, les diagnostics, les procédures, les structures anatomiques, les organismes, les substances, les produits pharmaceutiques, les dispositifs et les spécimens. Elle est l'une des terminologies cliniques les plus complètes avec plus de 350 000 termes

La SNOMED CT est maintenue et distribuée par la SNOMED International, un organisme international de développement de normes créé en 2007 et situé à Londres. Elle est disponible en plusieurs langues. Utilisée dans plus de 50 pays, cette terminologie nécessite l'obtention d'une licence est n'est donc pas librement accessible.

### 1.2.4. La terminologie LOINC

LOINC (Logical Observation Identifiers Names & Codes) est une terminologie de référence internationale pour le codage des observations et des documents électroniques. LOINC a été conçu en 1994 par le Regenstrief Institute, une organisation de recherche médicale américaine à but non lucratif en réponse à une demande d'accès public et gratuit à une base électronique codant des soins cliniques et des résultats de laboratoire afin d'avoir une base de référence publique et gratuite [109]. LOINC est également utilisée pour la codification des données relatives au DMP.

### 1.2.5. RadLex

La terminologie RadLex a été développée par la Radiological Society of North America (RSNA) afin de fournir une terminologie uniforme pour la pratique clinique, la recherche et l'éducation en imagerie médicale [110] . Lancé en 2005, RadLex comprend aujourd'hui plus de 46000 classes et couvre tous les domaines de la radiologie, comme les modalités d'imagerie, l'anatomie, la pharmacologie, etc. RadLex importe des vocabulaires provenant d'ontologies comme le FMA (Foundational Model of Anatomy [111]) pour l'anatomie.

En outre, la RSNA a conçu le RadLex Playbook [112], une nomenclature uniforme des procédures radiologiques, créée en combinant les termes RadLex pour les techniques d'imagerie, les parties du corps ciblées et les indications cliniques et destiné à servir de terminologie pour les procédures radiologiques. Ce playbook a été conçu afin de répondre au besoin du Dose Index Registry (DIR) de l'American College of Radiology (ACR), un registre de données permettant aux établissements d'imagerie de comparer leurs mesures de dose de tomodensitométrie aux valeurs régionales et nationales [113], de pouvoir calculer des statistiques inter-établissements. Ce système de codage normalisé permet de répondre aux problèmes de variations dans la dénomination des procédures (par exemple, en raison de l'utilisation de synonymes, d'abréviations et d'acronymes) [114] .

Chaque entrée du Playbook, dont la Figure 8 présente un extrait, comprend (entre autres champs) un identifiant (RPID) permettant d'identifier de manière unique chaque procédure, une brève description de la procédure (lisible par un humain) et un ensemble d'identifiants RadLex (RID) qui définissent collectivement le RPID. L'objectif du Playbook est de permettre aux établissements d'imagerie de partager des informations en associant la procédure au RPID au lieu des codes et des descriptions d'examen propres à chaque établissement.

| RPID | SHORT_NAME | AUTOMATED_SHORT_NAME | MODALITY | BODY_REGION | MODALITY_MODIFIER | ANATOMIC_FOCUS | RIDS |
|---|---|---|---|---|---|---|---|
| RPID2599 | | XRAY LE 1-2VWS ANKLE BILAT | XR | LOWER EXTREMITY | 1 - 2 VIEWS | ANKLE | RID10345\|RID13060\|0\|RID2638\|0\|0\|0 |
| RPID2600 | XR Ankle 1-2V | XRAY LE 1-2VWS ANKLE | XR | LOWER EXTREMITY | 1 - 2 VIEWS | ANKLE | RID10345\|RID13060\|0\|RID2638\|0\|0\|0 |
| RPID2601 | | XRAY GUIDE MAJ JNT ASP | XR | | GUIDANCE | MAJOR JOINT | RID10345\|RID13060\|0\|0\|0\|0\|0\|0\|RID |

*Figure 8 : Extrait du playbook. Chaque entrée du playbook représente un examen d'imagerie identifié par un RPID (RadLex Playbook ID) ayant un nom, un nom court, puis différentes caractéristiques qui sont autant de colonnes (modalité, région du corps, focus anato anatomique, etc.) donc les valeurs sont des termes de l'ontologie RadLex. La dernière colonne contient la liste des RID (RadLex ID) correspondant aux termes RadLex utilisés dans la ligne*

Une nouvelle version du Playbook appelée LOINC/RSNA Radiology Playbook est le résultat d'un travail d'harmonisation entre RadLex et la terminologie LOINC (Logical Observation Identifiers Names & Codes). Ce travail d'harmonisation financé par l'Institut national d'imagerie biomédicale et de bioingénierie (NIBIB) et le ministère de la Défense a été achevé en septembre 2017. Le LOINC/RSNA Radiology Playbook contient plus de 40000 termes.

### 1.2.6. Le standard DICOM

DICOM (Digital Imaging and COmmunications in Medicine), est un standard pour la gestion des données d'imagerie médicale créé en 1985 par l'ACR (*American College of Radiology*) et la NEMA (*National Electric Manufacturers Association*).

Ces deux comités mettent régulièrement à jour la norme DICOM avec l'aide d'autres comités d'experts internationaux tels le JIRA au japon, l'ANSI aux USA, le CEN/TC251 en Europe [115]. L'objectif de DICOM est de standardiser la transmission de l'information entre les différents appareils de radiologie. Le standard définit un format de fichier ainsi qu'un protocole de communication et est aujourd'hui universellement accepté.

Afin de normaliser les des méthodes de connexion, de transfert et d'identification, chacune des machines respectant la norme doit fournir un Document de Conformité (Conformance Statement). L'interconnexion des différents appareils se base sur les classes SOP (Service Object Pair) qui définissent des services proposés par un appareil et les informations que les fichiers DICOM devront contenir pour communiquer.

Les données DICOM sont organisées en 4 niveaux : Patient, Examen (Study), Séries et Image, comme illustré sur la figure 9.

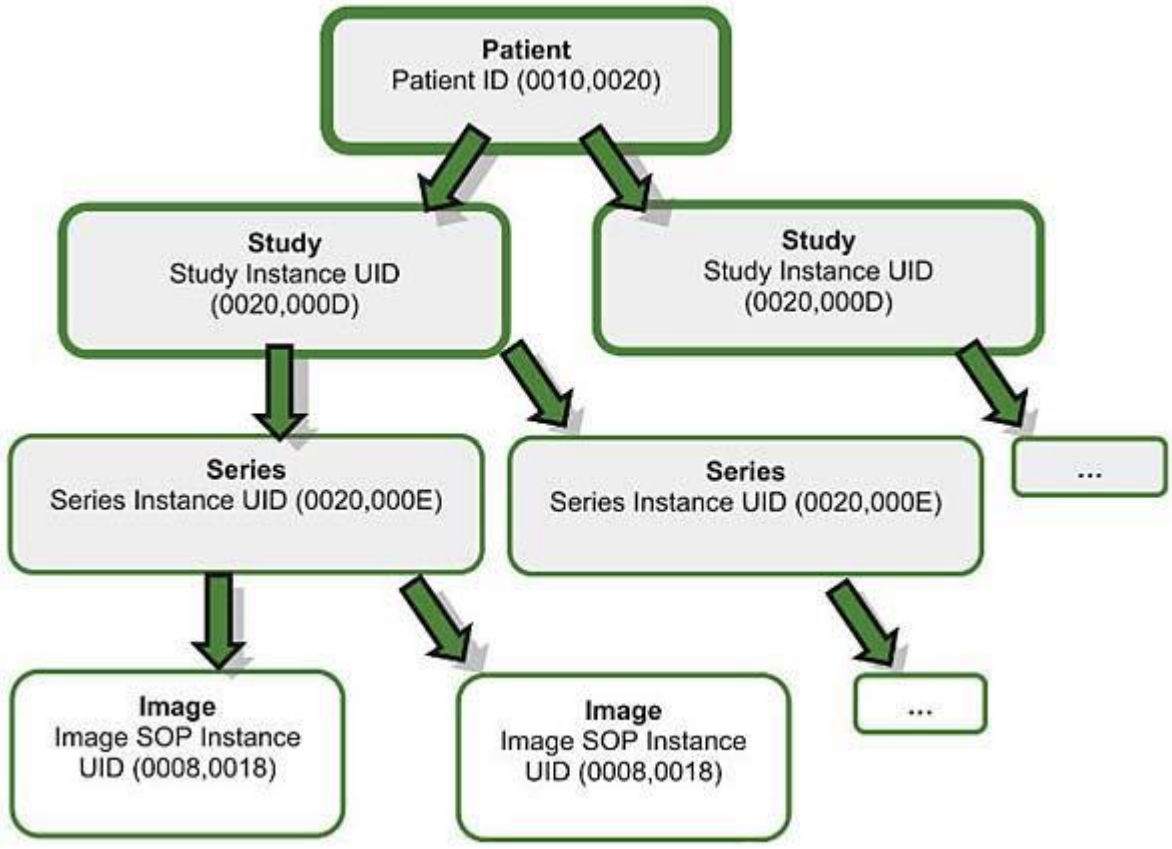



*Figure 9 : La hiérarchie de l'information dans DICOM*

Chaque image est stockée dans un fichier au format ".dcm". Un examen représente donc autant de fichiers que d'images acquises. Les fichiers DICOM contiennent un ensemble d'attributs (*DICOM Data Elements*) c'est-à-dire des associations clé et valeur auxquels s'ajoutent les informations sur la taille et le type de la donnée. Chaque attribut est donc défini comme suit :

- Un *tag* (la clé) composée de deux nombres codés sur 16 bits, l'un représentant le Group et l'autre l'élément. Les attributs sont ainsi "rangés" par groupes d'éléments. Les groupes numérotés pairs sont des éléments définis par la norme DICOM, les attributs publics. Les groupes de numéros impairs peuvent être définis par les utilisateurs du format de fichier (les constructeurs généralement), mais doivent se conformer à la même structure que les éléments standard, ce sont les attributs privés [116].
- Le type (*Value Representation*) est optionnel. Il est représenté sur deux caractères (par exemple : UI = Unique identifier, CS = coded string, US = unsigned short, …).
- La taille (*length* ) informe sur le nombre d'octets occupé par la valeur.

- un champ de valeur (value field) contenant les informations correspondantes. Si la longueur du champ de valeur est indéfinie, un élément de délimitation de séquence marque sa fin.

Les modules d'information DICOM (DICOM *Information Modules)* sont utilisés pour regrouper les attributs en unités logiques et structurées, par exemple un module “patient”, “appareil” (device), “Examen d’étude clinique” (Clinical Trial Study). Au sein du module, les attributs ont un “type” définissant la façon dont ils doivent être renseignés, ils peuvent être obligatoires avec une valeur non nulle, obligatoires avec une valeur potentiellement nulle, optionnels ou obligatoires à certaines conditions.

Les DICOM *Information Entities* (IEs) représentent les objets du monde réel tel que les patients, dispositifs médicaux, études cliniques etc. DICOM spécifie pour chaque IE les modules qu'il doit inclure. Les modules au sein d’un IE peuvent être obligatoires, conditionnels, facultatifs ou définis par l'utilisateur. Par exemple, l'IE “Patient” doit inclure le module “patient”, le module “specimen identification” et le module “clinical trial”. Il y a un IE pour chaque élément du modèle de donnée “Patient”, “Study”, “Series” et “Image” et beaucoup d’autres comme “Visite”, “équipement”, “essai clinique”, “procédure”, etc.

Enfin, les objets DICOM sont définis selon un IOD (Information Object Definition), c’est-à-dire une liste de IE. Un IOD doit contenir les quatre principaux IE’s: Patient, Study, Series et Image. Un fichier DICOM respecte donc un IOD, il existe par exemple des IODs pour les images de scanner, les images d’IRM, les vidéos d’endoscopie en temps réel ou les signaux respiratoires (resp. “CT Image IOD”, “MR Image IOD”, “Real-Time Video Endoscopic Image IOD” et“Respiratory Waveform IOD”). La figure 10 représente l’IOD “MR Image IOD”

| IE | Module | Reference | Usage |
|---|---|---|---|
| Patient | Patient | C.7.1.1 | M |
| | Clinical Trial Subject | C.7.1.3 | U |
| Study | General Study | C.7.2.1 | M |
| | Patient Study | C.7.2.2 | U |
| | Clinical Trial Study | C.7.2.3 | U |
| Series | General Series | C.7.3.1 | M |
| | Clinical Trial Series | C.7.3.2 | U |
| Frame of Reference | Frame of Reference | C.7.4.1 | M |
| Equipment | General Equipment | C.7.5.1 | M |
| Acquisition | General Acquisition | C.7.10.1 | M |
| Image | General Image | C.7.6.1 | M |
| | General Reference | C.12.4 | U |
| | Image Plane | C.7.6.2 | M |
| | Image Pixel | C.7.6.3 | M |
| | Contrast/Bolus | C.7.6.4 | C - Required if contrast media was used in this image |
| | Device | C.7.6.12 | U |
| | Specimen | C.7.6.22 | U |
| | MR Image | C.8.3.1 | M |
| | Overlay Plane | C.9.2 | U |
| | VOI LUT | C.11.2 | U |
| | SOP Common | C.12.1 | M |
| | Common Instance Reference | C.12.2 | U |

*Figure 10 : Exemple d'implémentation d'un Information Object Definition : Les modules impliqués dans l'IOD "MR Image"* [117(p3)]

# 2. Mapping de la terminologie d'interface sur une terminologie de référence

## 2.1. Contexte et objectif

Comme nous l'avons vu, les études en santé, en particulier en imagerie, nécessitent des données de qualité et en importante quantité. Un des points centraux de la préparation des données est donc l'alignement des données issues d'un établissement donné, donc alignées sur une terminologie d'interface (TI), sur une terminologie de référence (TR). Cette terminologie de référence doit conserver au maximum le sens des informations d'origine contribuant ainsi à la qualité des données et permettant leurs interopérabilités. Cette interopérabilité permettra ensuite de croiser les données avec d'autres sources mais aussi de les exploiter dans le cadre d'études multicentriques [88,91,108,118].

Dans ce travail, nous abordons la question du mapping entre les terminologies d'interface et de référence afin d'amener les données du soin vers la recherche en s'assurant que les examens d'imagerie considéré dans les études sont classés de façon fiable, c'est-à-dire en conservant le sens des données. L'objectif est ici de dresser la liste des obstacles à cette tâche, provenant des constructions des terminologies d'interface ou de référence, ou de la qualité de nos données.

Nous travaillons dans cette étude avec deux versions du playbook RadLex, un ensemble de plus de 4400 terme représentant des examens d'imagerie conçu par des spécialistes et décrits par des concepts tirés de l'ontologie RadLex. Une des vocations de ce playbook est de permettre l'alignement des descriptions d'examens (une des métadonnées des examens d'imagerie) sur un référentiel dans le but de mener des études multicentriques.

Plusieurs travaux se sont intéressés au mapping des terminologies en radiologie et ses obstacles en considérant les scanners et/ou les IRMs. Ces approches s'intéressent à l'amélioration de la terminologie d'interface par son dédoublonnage[119], l'amélioration de la terminologie de référence en enrichissant cette dernière [120] ou, comme dans notre cas, à l'automatisation du procédé de mapping [121].

Le travail que nous avons effectué pour répondre à cette problématique, et décrits dans les parties suivantes, a fait l'objet d'un article présenté à la conférence MIE 2022 et se trouve en annexe 1 de ce document.

## 2.2. Matériel

La TI considérée dans notre étude sera constituée des titres des comptes-rendus d'examen d'imagerie disponibles dans l'entrepôt eHOP. En effet, l'attribut DICOM « Study Description » présent dans les métadonnées DICOM auxquels nous avons aussi accès est parfois tronqué (dû à la limite de caractère dans le champ) ou composé de mots répétés selon l'origine de l'examen, voir absent. Les titres les comptes-rendus sont quant à eux toujours renseignés et ont été conçus par les radiologues et utilisés dans le système d'information radiologique et dans les dossiers de santé. De plus, les comptes-rendus sont disponibles depuis 2004 dans notre entrepôt et représentent donc une plus grande quantité de données : 1467000 comptes-rendus pour 486000 patients.

Comme on a pu le voir, il existe plusieurs ressources pour la représentation des données d'imagerie. Parmi elles, le LOINC/RSNA Radiology Playbook a été conçu spécifiquement pour représenter les examens d'imagerie et a fait l'objet d'une fusion avec LOINC, qui fait partie des terminologies déjà utilisées dans l'entrepôt (Pour représenter les types de document notamment). Nous avons fait le choix de travailler avec ce playbook mais aussi avec la version du playbook précédent la fusion avec LOINC afin de voir comment cette fusion impacte la couverture de notre TI. Nous étudions donc l'alignement de notre TI avec le playbook RadLex version 2.5 et le LOINC/RSNA Radiology Playbook version 2.71.

## 2.3. Méthodes

La première étape est de nous assurer que la TI permet bien de refléter le contenu des examens. L'intégration des données d'imagerie à l'entrepôt, effectuée par ailleurs, nous permet de vérifier cette concordance. Nous pouvons donc vérifier que les métadonnées, en particulier l'attribut « Body Part examined », sont cohérent avec le titre du rapport associé.

Afin de lister les obstacles à l'alignement des terminologies, nous allons l'effectuer manuellement pour relever les différents cas où il n'est pas trivial et répertorier les origines de ces problèmes. Ce travail étant manuel, nous avons réduit notre TI aux 200 termes les

plus fréquents. Cet ensemble de 200 termes représente 73.2% du total des examens d'imagerie présents dans l'entrepôt, soit 1073886 comptes-rendus.

Nous procédons ensuite au mapping manuel des termes de notre terminologie locale sur les TRs choisies sans mettre en œuvre de connaissances expertes du domaine médical, l'objectif n'étant pas simplement l'alignement mais d'étudier pour chaque cas les raisons de la réussite ou de l'échec de celui-ci.

Afin de classer les différents cas de succès ou d'échec d'alignement, nous nous somme basé sur les travaux de Humphreys et al. [122] pour définir les différentes éventualités :

● *Exact match*, Correspondance exacte : le code de la TR correspond exactement à la procédure, par exemple, « Ultrasound - Abdomen-Kidney » correspond parfaitement à « US ABD KIDNEY » (RPID1992)

● *Broader RT term issue*, La terminologie de référence est trop « générale » : le meilleur candidat de la TR avait un sens plus large que le terme local. Par exemple, le Playbook L/R ne dispose pas d'un code contenant tous les éléments de « CT - Torse Abdomen Pelvis Crâne »

● *Narrower RT term issue*, La terminologie de référence est trop « spécialisée » : certains termes de la TR spécifient des informations supplémentaires qui ne sont pas dans le terme local. Par exemple, le Playbook RadLex spécifie toujours des informations sur l'agent de contraste dans l'IRM du sein et cela empêche de trouver une correspondance exacte pour le terme local « IRM – seins »

● *No exact match,* Pas de correspondance exacte : le terme local utilise un concept qui n'est pas encore défini ou jamais utilisé dans la TR. Par exemple « hémosidérose » qui n'est jamais utilisé dans les deux playbooks (mais est défini dans l'ontologie RadLex (RID5203)) ;

Enfin, nous avons revu avec un expert notre tableau d'alignements afin de valider les cas de correspondances exactes, vérifier la justesse des traductions, expliquer les acronymes utilisés localement et s'assurer que plusieurs zones anatomiques inclues dans un terme du playbook sont redondantes entre elles et peuvent bien être mises en correspondance avec une zone anatomique précisée dans un titre de compte rendu. Par exemple « MR ABD LIVER » (RPID2211) inclut « abdomen » et « foie » qui désignent une même zone anatomique, le terme devient donc une correspondance exacte avec le titre de compte-rendu « IRM – Foie ».

## 2.4. Résultats

D'une part, ce travail nous a permis de relever des problèmes dans la façon dont sont représentés les examens dans l'hôpital. Notamment, un titre de rapport d'examen peut concerner uniquement une région anatomique, par exemple « Radiographie de l'épaule » alors que l'examen lui-même contient des acquisitions de régions supplémentaires (en plus de l'épaule). La figure 11 présente les résultats de requêtes sur des examens ayant pour titre « Radiographie – Poignet » et « Radiographie – Genou », les carrés orange rassemblent les informations de *series* DICOM appartenant à un même examen. On voit par exemple qu'un rapport intitulé « Radiographie – Poignet » concerne un examen où des images du rachis cervical ont aussi été acquises. En effet, les cliniciens décident parfois en salle de radiologie de faire des acquisitions supplémentaires, notamment pour les examens radiologiques dans le cadre d'un traumatisme.

| ACCESS_NUMBER | SERIES_DESC | BODYPARTEXAM | TITRE |
|---|---|---|---|
| [redacted] | LATERAL | POIGNET | Radiographie - Poignet |
| [redacted] | AP | RACHIS CERVICAL | Radiographie - Poignet |
| [redacted] | AP | RACHIS LOMBAIRE | Radiographie - Poignet |
| [redacted] | AP | RACHIS CERVICAL | Radiographie - Poignet |
| [redacted] | LATERAL | POIGNET | Radiographie - Poignet |
| [redacted] | LATERAL | POIGNET | Radiographie - Poignet |
| [redacted] | LATERAL | POIGNET | Radiographie - Poignet |
| [redacted] | LATERAL | POIGNET | Radiographie - Poignet |
| [redacted] | LATERAL | POIGNET | Radiographie - Poignet |

| ACCESS_NUMBER | SERIES_DESC | BODYPARTEXAM | TITRE |
|---|---|---|---|
| [redacted] | AP | COUDE | Radiographie - Genou |
| [redacted] | AP | GENOU | Radiographie - Genou |
| [redacted] | OTHER | GENOU | Radiographie - Genou |
| [redacted] | AP | EPAULE | Radiographie - Genou |
| [redacted] | AP | CHEVILLE | Radiographie - Genou |
| [redacted] | LATERAL | GENOU | Radiographie - Genou |
| [redacted] | AP | GENOU | Radiographie - Genou |
| [redacted] | AP | GENOU | Radiographie - Genou |
| [redacted] | AP | OMOPLATE | Radiographie - Genou |

*Figure 11 : Un examen contient parfois plusieurs acquisitions de différentes parties du corps alors que le titre du rapport ne mentionne qu'une d'entre elles*

Enfin, la terminologie d'interface (l'ensemble des titres de rapports d'imagerie) contient assez peu d'informations et est redondante, ce qui provient de l'évolution au cours du temps de ces termes, définis par les professionnels eux-mêmes.

D'autre part, nous avons pu lister les obstacles que devront surmonter les algorithmes de mapping automatique, des problèmes de traductions depuis le français, aux couvertures du domaine différentes selon les terminologies.

| Mapping category | RadLex Playbook | LOINC/RSNA (L/R) Radiology Playbook |
|---|---|---|
| Exact match | 61(57.5%) | 73 (68.8 %) |
| Broader RT term issue | 18 (17.0 %) | 13 (12.3 %) |
| Narrower RT term issue | 17 (16.0 %) | 10 (9.4 %) |
| No exact match | 10 (9.4 %) | 10 (9.4 %) |

*Figure 12 : Résultats de l'alignement de notre terminologie d'interface sur les playbooks*

La plupart des cas où la terminologie de référence était trop spécifique (*Narrower RT term issue*) était dûe à la mention de l'agent de contraste dans les termes de la TR, alors que cette information n'était pas spécifiée dans les termes locaux. Cependant, cette information n'est pas fournie de manière homogène dans les TRs, par exemple dans le Playbook L/R, les termes « CT Chest », « CT Abdomen » et « CT Chest and Abdomen and pelvis » (resp. 24627-2, 41806-1, 87869-4) ne mentionnent pas d'agent de contraste, alors que les termes décrivant les scanners du « thorax et de l'abdomen » précisent toujours l'utilisation d'un agent de contraste (« with », « without » ou « without and with » resp. 42275-8, 42276-6, 42277-4). La revue avec l'expert a permis de confirmer que certains examens sont réalisés avec et/ou sans agent de contraste dans la pratique sans que le titre du rapport ne le précise ce qui empêche une correspondance parfaite (*exact match*) bien que la RT définisse le terme qui correspond exactement à l'examen effectué.

Le cas où la terminologie de référence était trop spécifique (*Broader RT term issue*) s'est produit lorsque les termes les plus proches dans la TR ne comprenaient pas tous les mots correspondant au titre local. Nous avons observé que le niveau de spécification des termes de la TR peut varier en fonction de la modalité, entre autres. Par exemple, le terme « MR Lower Extremity Joint » existe dans le Playbook L/R (24687-6), mais il n'y a pas d'équivalent exact pour « Ultrasound - Lower Extremity Joint ».

En cas d'absence de correspondance, la raison la plus fréquente était la spécification par le terme local de la raison de l'examen, ou une procédure qui n'était pas mentionnée dans les playbooks. Par exemple, les concepts « tuberculose » ou « cystographie » ne sont jamais utilisés dans le Playbook L/R (mais existent dans l'ontologie RadLex, RID34878 et RID29116). Un autre exemple est l'utilisation d'une nouvelle technique d'imagerie (par exemple, le système d'imagerie EOS™ récemment décrit) qui n'a pas encore été ajoutée dans la TR.

Enfin, la terminologie de référence présente aussi des opportunités d'amélioration. Lors du mapping, nous avons pu constater que des termes a priori utiles - puisqu' utilisés en pratique à Rennes - n'étaient pas définis (c'est à dire non pré-coordonnés) alors que des termes similaires, parfois très proches, étaient pré-coordonnées. A l'inverse, certaines associations de concepts (i.e précoordination de termes) sont manquantes alors que des termes plus précis, rassemblant au moins les concepts du terme manquant, existent.

## 2.5. Discussion

L'objectif du travail présenté ici était d'étudier les liens entre la terminologie d'interface utilisée dans le cadre des soins (ici les titre des comptes-rendus d'imagerie) et des terminologies de référence reconnues dans le domaine de la recherche en imagerie (les playbooks RadLex et Loinc). Le rapprochement manuel de ces terminologies permet de mettre en lumière les obstacles à prendre en compte pour le développement d'outils automatiques de mapping mais aussi les points d'améliorations potentiels, dans la chaîne de traitement des informations de l'hôpital, qui permettrait de faciliter ce travail de mapping.

Les leviers pour l'amélioration de la description des examens sont donc nombreux. Les titres d'examens peuvent être complétés par d'autres données tirées du rapport ou des métadonnées d'imagerie pour plus de précision. L'extraction et la traduction des informations en français vers l'anglais peut être amélioré en affinant les ressources utilisées pour interpréter le vocabulaire local (ensemble des acronymes, diminutifs, expression) et pour la traduction (UMLS, traduction d'expression composées) et en utilisant des approches plus fines comme du NLP [70] qui permet de mieux considérer le contexte d'utilisation des termes, leurs significations, leurs inter-relations etc.

Notre étude étant basée sur les données d'un seul hôpital, certains des problèmes identifiés peuvent être spécifiques à notre institution. De plus, nous avons utilisé un ensemble limité de termes locaux et nous avons donc pu manquer des problèmes de mapping liés à des examens moins courants. D'autres travaux sur la mise en correspondance des TI avec les TR dans différents domaines mentionnent les mêmes problèmes de granularité différente entre les TI et les TR [119–121], ou de termes qui n'ont de sens que dans l'institution locale [121]. Ces observations soulignent l'importance de suivre les bonnes pratiques lors de la création de TI basées sur des sous-ensembles tirés de TR. Cette approche de création des TI permet également d'identifier les termes qui manqueraient à la TR pour la faire évoluer efficacement.

Une approche compositionnelle, c'est-à-dire la définition de règles de compositions de nouveaux termes, permettrait d'éviter ces écueils en donnant la possibilité de former des termes adéquats tant que les éléments qui le constituent sont présents dans l'ontologie de référence et que la grammaire compositionnelle le permet. La SNOMED propose une approche compositionnelle ce qui offre plus de possibilité pour le mapping avec d'autres terminologies[123]. Enfin, comme cela a été souligné par Rosenbloom et al. [124], il est préférable de se baser sur une terminologie de référence pour former la terminologie d'interface afin de faciliter le rapprochement entre les deux terminologies.

## 2.6. Perspectives

Les résultats de cette étude sont un socle qui permettra de développer un outil d'alignement de nos données d'imagerie sur une terminologie de référence, dans le but de mener des études multicentriques.

L'objectif était de comprendre comment notre terminologie d'interface peut s'aligner sur les playbooks RadLex et de lister les verrous à cet alignement, aussi le mapping est réalisé en se basant sur l'équivalence entre mots, avec une traduction mot à mot. Le problème de la traduction automatique et de la mise en œuvre de connaissance experte automatique a fait l'objet d'un autre travail que nous présentons dans la suite de ce manuscrit. Dans l'article qui suit, l'approche est au contraire d'avoir une solution de classification des données brutes, en français, automatisée de bout en bout. Pour cela, on emploi une base de connaissance pour la traduction des termes clinique (UMLS) et on met en œuvre un raisonneur automatique basé sur les ontologies. L'objectif ici est plus une preuve de concept, d'où l'approche automatisée.

# 3. Classifier les examens grâce au raisonnement ontologique

## 3.1. Contexte et objectifs

L'alternative au mapping de la terminologie local sur la terminologie de référence, est le déploiement d'un outil de classification automatique individuel, effectuant la classification de chaque examen individuellement à partir de ses informations.

Dans notre contexte, ayant accès aux métadonnées de chaque examen, nous voulons étudier la mise en place d'une telle solution de classification "basée sur les données". Pour effectuer cette classification automatique des examens d'imagerie, nous avons choisi la terminologie du playbook RadLex qui, comme vu précédemment, est adossé à l'ontologie RadLex ce qui facilite l'interopérabilité mais permet aussi d'effectuer des raisonnements automatiques sur les données. Cette capacité à décrire les examens à travers une ontologie ouvre des perspectives d'utilisation de ces descriptions formelles des données pour des cas d'études utilisant l'intelligence artificielle ou du data mining. L'approche utilisée dans ce travail est d'utiliser cette capacité de raisonnement automatisé pour la classification elle-même.

Comme on l'a vu précédemment, si le playbook RadLex utilise les classes de l'ontologie RadLex pour décrire les examens, ces derniers ne sont pas intégrés dans l'ontologie elle-même. Nous avons donc ici pour objectifs (1) de constituer une ontologie des procédures de radiologie en intégrant le playbook dans l'ontologie Radlex et (2) de concevoir une preuve de concept de classification automatique des examens d'imagerie à partir des données brutes (DICOM)

La conception d'une ontologie permettant de représenter à la fois les types d'examens et les concepts de l'imagerie (i.e la fusion du playbook dans l'ontologie RadLex) ainsi que l'outil de classification automatique, que nous allons décrire dans les parties suivantes, ont fait l'objet d'un article présenté à la conférence ICBO 2020 et qui se trouve en annexe 2 de ce document.

## 3.2. Matériel

Comme on l'a vu, le playbook RadLex que nous utilisons contient plus de 4400 procédures d'imagerie identifiées par un RPID (RadLex Playbook ID) et décrites par 26 champs (par exemple « MODALITY », « POPULATION », « BODY_REGION », « BODY_REGION_2 », « MODALITY_MODIFIER », …) dont les valeurs sont des RID (RadLex ID) faisant référence à des concepts de l'ontologie RadLex.

Pour des raisons d'optimisation, nous avons d'abord réduit l'ontologie RadLex à l'ensemble des classes utilisées par le playbook ainsi que toutes les classes qu'elle subsument ou qui les subsument. L'ontologie ainsi réduite contient 12103 classes.

Afin de pouvoir traiter nos données locales, en français, avec RadLex, nous devons les traduire vers l'anglais. Dans RadLex, la propriété *UMLS_ID* renseigne, quand cela est possible, l'identifiant UMLS correspondant à une classe. L'UMLS, présenté plus haut, contient les traductions de nombreux termes médicaux en différentes langues, dont le français. Pour chaque classe, notre outil de traduction automatique cherche dans l'UMLS l'élément correspondant en utilisant la propriété *UMLS_ID* si elle est disponible. Sinon, il cherche dans l'UMLS un alignement exact entre la propriété *Preferred_name* de la classe et un label dans l'UMLS. Quand un ou plusieurs éléments de l'UMLS correspondent à la classe RadLex, celle-ci est enrichie avec des propriétés *Synonym_french* qui prennent les valeurs les labels en français de l'UMLS.

Les outils pour extraire cette sous-partie de l'ontologie, ajouter les synonymes en français des nom des classes et pour effectuer toutes les autres tâches de manipulation automatique des ontologie décrites ci-après dans la partie « Méthode », ont été développés en Java en utilisant la librairie Jena [125].

## 3.3. Méthode

### 3.3.1. Conception de l'ontologie des procédures

Nous intégrons le playbook RadLex dans l'ontologie RadLex en recréant automatiquement les relations entre les termes du playbook et leurs concepts associés dans l'ontologie. Ainsi un terme du playbook devient une classe de l'ontologie, et les colonnes du playbook deviennent des propriétés permettant de faire le lien entre ces nouvelles classes et les

concepts de l'ontologie qui leur sont associés. Les champs textuels du playbook deviennent alors des propriétés d'annotation (*LONG_NAME*, *AUTOMATED_SHORT_NAME* et *AUTOMATED_LONG_NAME*) tandis que les champs contenant des RPID deviennent des propriétés d'objet, faisant ainsi le lien avec les concepts de l'ontologie. Afin de se conformer à la pratique de l'ontologie RadLex, nous avons défini une propriété d'annotation *Preferred_name* à laquelle nous attribuons la valeur de la colonne SHORT_NAME (si elle n'est pas renseignée, nous utilisons la valeur de *AUTOMATED_SHORT_NAME*). Les champs contenant des RIDs (de *MODALITY* à *VIEW_4*) et ayant une valeur dans le playbook ont été convertis en restriction *owl:someValuesFrom* sur la définition *subClass of* de la nouvelle classe (cf Figure 13, les restrictions *has_ANATOMIC_FOCUS*, *has_BODY_REGION*, *has_MODALITY* et *has_MODALITY_MODIFIER*). Certaines de ces colonnes ont plusieurs versions numérotées. Par exemple, 5 colonnes sont définies pour la région du corps, de *BODY_REGION* à *BODY_REGION_4* car une procédure d'imagerie peut couvrir plusieurs régions du corps. Nous n'avons pas tenu compte d'une hiérarchie au sein de ces colonnes et avons défini une seule propriété d'objet pour ces cas. Par conséquent, chaque champ *BODY_REGION_X* renseigné a entraîné la définition d'une nouvelle restriction *has_BODY_REGION* sur la classe nouvellement créée. Le seul élément spécifique est le champ POPULATION pour lequel nous avons également déclaré une restriction *owl:allValuesFrom*, car le patient ne peut pas être dans plusieurs catégories à la fois.

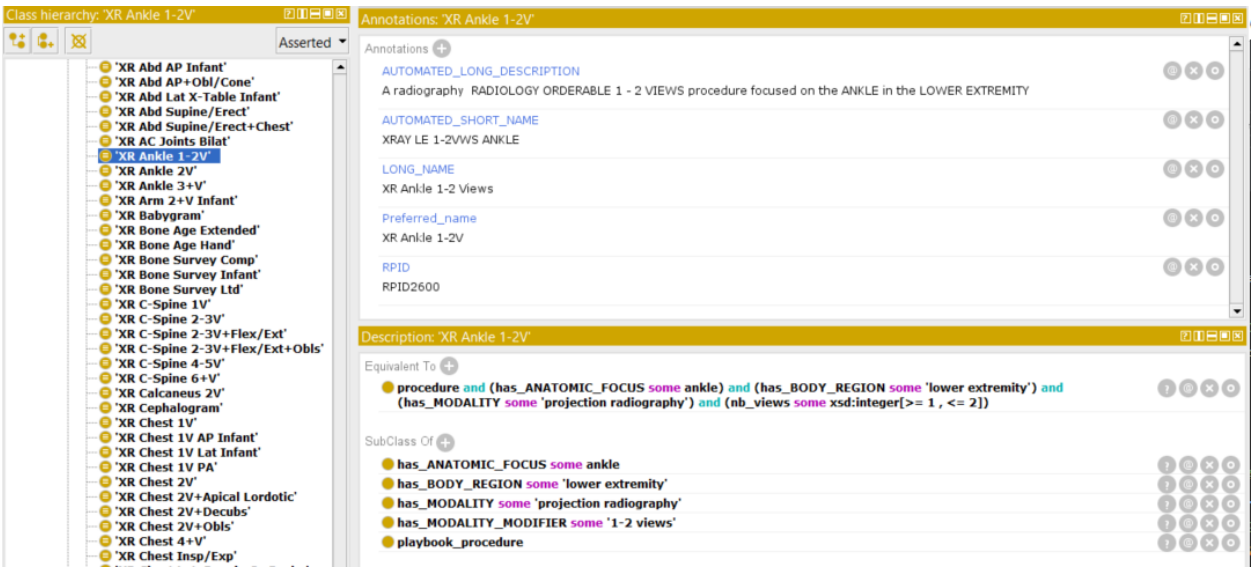



*Figure 13 : Le playbook est mergé dans l'ontologie RadLex. Les termes du playbooks deviennent des classes, sous-classes de la classe procédure. Les colonnes du playbook sont devenues des restrictions qui lient ces nouvelles classes avec les concepts RadLex. Par exemple, la colonne BODY_REGION est devenue une restriction has_BODY_REGION.*

Nous décrivons la clause *owl:equivalentClass* pour regrouper toutes ces restrictions en ajoutant le fait que la classe est une sous-classe de la classe *procedure* (celle-ci étant déjà définie dans RadLex).

La colonne *MODALITY_MODIFIER* contenant parfois des informations sur le nombre de vues utilisées pour la procédure, nous avons défini la propriété de données *nb_views*, plus significative puisque cela permet de manipuler directement un nombre. Ainsi, les valeurs de *MODALITY_MODIFIER* décrivant le nombre de vues ont été spécifiquement converties en contraintes en utilisant cette propriété de données. Sur la figure 13, on voit que la colonne du playbook *MODALITY_MODIFIER* avec la valeur « 1 – 2 views » a été convertie en une contrainte sur la propriété de données *nb_views* dans la clause *owl:equivalentClass* de l'ontologie résultante. Toutes ces classes nouvellement créées deviennent des sous-classes de *playbook_procedure* définie comme une sous-classe de *procedure*.

### 3.3.2. Classification par le raisonnement ontologique

Une fois notre ontologie des procédures finalisée, nous mettons en place la classification automatique. Comme on l'a vu, le standard DICOM décrit les métadonnées qui accompagnent les images médicales par une séquence d'attributs de différents types, dont un type de séquence pouvant contenir lui-même d'autre attributs.

La première étape de la classification est de créer une instance de la classe *procedure* dans notre ontologie puis de l'enrichir en extrayant des métadonnées les concepts associés à des classes RadLex. Nous extrayons les informations permettant la classification depuis un ensemble défini de métadonnées pertinentes pour chaque « Dicom Study ». Cette association est rendue possible par la traduction en Français faite préalablement sur l'ontologie RadLex et la définition de certaines règles ad-hoc. Par exemple, les valeurs « B », « L » et « R » de l'attribut DICOM « Image Laterality » sont associé par ces règles respectivement aux concepts RadLex « bilateral » (RID5771), « left » (RID5824) et « right » (RID5825). De la même façon, en utilisant les attributs « Study Date » et « Patient's Birth Date », nous définissons une restriction *has_POPULATION* en fonction de l'âge du patient (par exemple, en ciblant la classe "néonatal" si le patient a moins d'un mois).

Pour les métadonnées contenant du texte (e.g « Study Description » ou « Series Description ») nous comparons chaque groupe de 1, 2 ou 3 mots avec les valeurs des propriétés *Preferred name*, *Acronym*, *Synonym* et *Synonym_french* de toute notre ontologie RadLex réduite. Afin de gérer les négations, le groupe de mot n'est pas pris en compte s'il contient un mot comme « non », « sans », « ss » (raccourci parfois utilisé pour « sans »), etc. Les acronymes ou les raccourcis sont remplacés par le ou les mots dont ils découlent. La

liste d'acronymes est définie manuellement, l'acronyme « TAP » est par exemple remplacé par les mots « thorax », « abdomen » et « pelvis » ou « mammo » par « mammographie ».

Une fois la liste des classes RadLex correspondantes établie, nous créons une instance de chacune de ces classes puis nous cherchons parmi les propriétés d'objets que nous avons créé celle qui permet de représenter la relation entre la procédure et le concept. Ainsi le RID « RID10312 » de la classe « MRI » n'étant utilisé dans le playbook que dans la colonne *MODALITY,* l'instance de la classe « MRI » créée est liée à notre instance de *procedure* par la propriété d'objet *has_MODALITY.*

Chaque acquisition devient donc une instance de la classe *procedure* dont on renseigne les liens à d'autres concepts que l'on a pu obtenir grâce aux métadonnées. Enfin, on lance le raisonneur qui va, à partir des données renseignées, trouver les classes les plus adaptées à cette instance de procédure et donc la classifier. Par exemple, une instance de procédure reliée par la relation *has_MODALITY* à une instance du concept de scanner (« computed tomography » RID10321) et relié par relation *has_BODY_REGION* une instance du concept d'os (« bone_organ » RID13197) sera considéré comme une instance de la classe de procédure « CT BONE » RPID1241 ajoutée dans l'ontologie depuis le playbook. Nous avons utilisé le raisonneur HermitT [126] pour remplir cette tâche.

La figure 14 ci-dessous permet de résumer visuellement ce processus. L'ontologie RadLex d'origine est représentée en noir, on y voit les concepts de « procédure », « tomodensitométrie » et « tête ». En vert sont représentés les nouvelles classes extraites du playbook, le concept « procédure playbook », sous classe de « procédure » contient chacune des classes correspondant à un élément du playbook. Lorsque notre outil traite un fichier DICOM, il créé une instance de procédure, ici *Examen_001* et pour chaque information extraite des métadonnées qu'il peut faire correspondre à un concept de l'ontologie, il créé une instance de cette classe qu'il lui associe. Ici les concepts de « tomodensitométrie » et de « tête » ayant été trouvés, il les a instanciés (respectivement *Tomodensitométrie_001* et *Head_001*) et à associé ces instances à *Examen_001*. Enfin, le raisonneur ontologique traite ces informations et déduit que l'examen *Examen_001* est une instance de la classe « CT Head ».

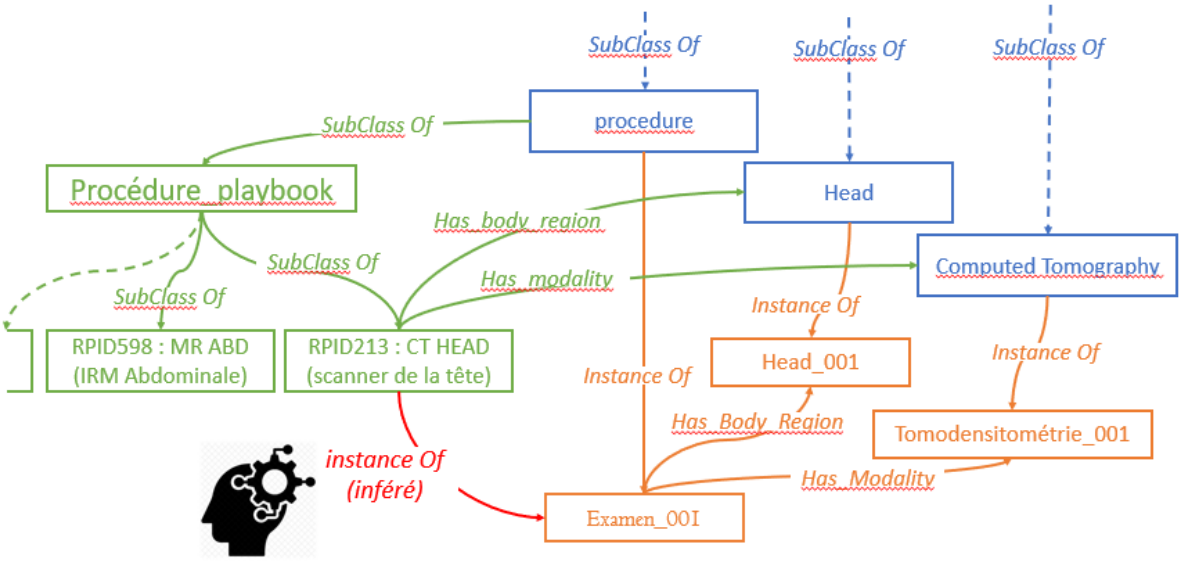



*Figure 14 : Résumé du processus de classification ontologique*

## 3.4. Résultats

### 3.4.1. L'ontologie des procédures

La traduction vers le français a permis de trouver 4623 synonymes français, 1301 classes ayant été traduites avec au moins un synonyme français. Cependant, nous avons parfois détecté des erreurs dues à l'appariement qui est fait sur les acronymes. Ainsi, MRI a été associé à « Mauritanie » ou à « arriération mentale ».

Cette ontologie contient 16506 classes dont 4402 étant le résultat de l'intégration d'une ligne du playbook. Cette ontologie est disponible en ligne sur le lien suivant : https://github.com/pierrelemordantUR1/ICBO_2020/blob/master/playbook_ontology_v7_fr.owl

### 3.4.2. La classification automatique

Nous avons appliqué notre classifieur sur les données d'imagerie anonymisées d'IRM, scanneur et ultrasons acquis sur une journée dans deux institutions, le CHU de Pontchaillou et le centre Eugène Marquis. La figure 15 ci-dessous décrit les instances de *procedure* créées automatiquement à partir des données DICOM brutes extraites du PACS.

| | Institution 1 | | Institution 2 | |
|---|---|---|---|---|
| our Object Property | number of individuals using this object Property | number of different Radlex classes targeted by this object Property | number of individuals using this object Property | number of different Radlex classes targeted by this object Property |
| has_MODALITY | 122 | 4 | 75 | 5 |
| has_MODALITY_MODIFIER | 58 | 2 | 63 | 2 |
| has_BODY_REGION | 41 | 6 | 57 | 9 |
| has_ANATOMIC_FOCUS | 29 | 11 | 41 | 5 |
| has_REASON_FOR_EXAM | 4 | 3 | 8 | 5 |
| has_POPULATION | 4 | 1 | 2 | 1 |
| has_TECHNIQUE | 0 | 0 | 26 | 1 |

*Figure 15 : Caractéristiques des instances de la classe procedure conçues à partir des examens d'imagerie DICOM*

La requête sur le PACS du CHU sur une journée a retourné 122 examens décrits en moyenne par plus de 2 propriétés d'objet. Au CEM, 75 examens ont été retournés, décrits par 5 propriétés d'objet en moyenne.

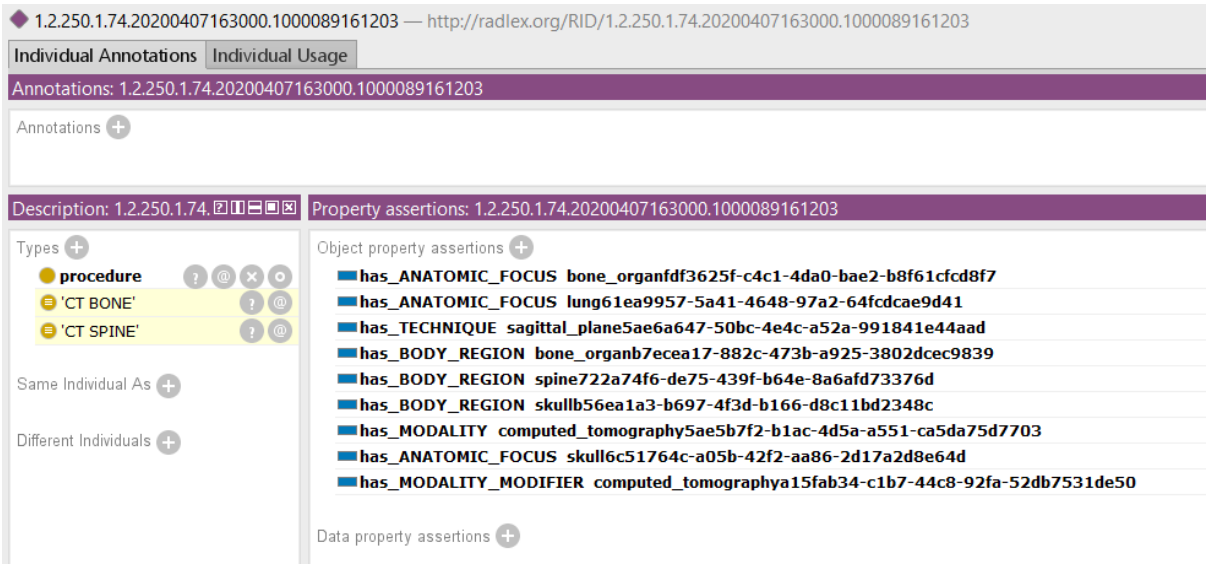



*Figure 16 : Résultat de la classification automatique d'un examen vu dans protégé*

La Figure 16 illustre le résultat d'une classification. Le classifieur a déterminé que l'instance de *procedure* était aussi une instance des classes *CT BONE* et *CT SPINE*, ajoutées depuis le playbook.

Sur les 122 examens du CHU, 4 ont été classés dans 2 procédures différentes du Playbook, 92 ont été classés dans 1 procédure et 26 n'ont pas été classés. Dans le cas des examens non classés, la seule information recueillie était la modalité de la procédure qui était « US » (ultrasons) alors que le playbook ne définit pas de procédure dont on sait simplement qu'elle a une modalité « ultrasons » (comme c'est le cas pour les IRM). Sur les 75 examens du CEM, 42 ont été classés dans une classe, 30 dans deux classes et 3 dans 4 classes différentes.

Pour les deux institutions, le temps d'exécution de la requête pour retrouver les métadonnées des examens était inférieur à 30 secondes et le temps de création des instances à partir des données brutes était d'environ cinq secondes. Le raisonneur HermiT prend 20 secondes pour charger et classer nos instances de *procedure*.

## 3.5. Discussion

### 3.5.1. L'ontologie des procédures

L'intégration du playbook dans l'ontologie RadLex permet de manipuler les procédures, comme tous les concepts de l'ontologie, avec les outils des manipulations automatique des ontologies. Cette intégration a une valeur ajoutée comme nous l'avons montré avec notre cas d'usage. Si l'ontologie RadLex a l'avantage de couvrir largement le domaine des procédures d'imagerie, il existe cependant quelques réserves. La description de l'anatomie, par exemple, pourrait être améliorée même si un effort a été fait pour mettre en correspondance certains termes avec la FMA[127] . Plus globalement, il serait intéressant d'avoir un alignement de RadLex avec des ontologies formelles telles que BFO et d'autres ontologies centrales de l'Open Biological and Biomedical Ontologies Foundry (OBO)[110].

La richesse des propriétés d'annotation « Preferred_name », « Acronym » et « Synonym » est importante car nous nous y fions pour faire correspondre les données sources et les RIDS, mais elle doit être améliorée. Par exemple, notre programme n'a pas pu traiter le mot « épaule » car le Playbook n'utilise pas la classe RadLex « shoulder » (RID39518) mais « shoulder girdle » (RID1852), or cette dernière n'a pas pu être traduite par notre méthode. Avec 1301 classes traduites sur les 12103 que compte l'ontologie réduite que nous avons

conçu, notre système de traduction demande à être amélioré car notre utilisation de l'UMLS a montré ses limites.

Comme on l'a montré dans le premier chapitre de cette partie, les entrées du Playbook couvrent différemment les modalités. Une procédure « MR » dans le playbook permet d'attribuer une classe à tous les examens IRM même si aucune autre information n'est disponible, mais une telle procédure n'existe pas pour l'échographie. Pour le scanner (CT), une telle entrée existe (RPID88) mais son Preferred_name est « CT Guide needle place », ce qui révèle que cette procédure était destinée à modéliser le « CT Guidance for Needle Placement », bien que cette motivation ne soit pas explicite, par exemple dans la colonne REASON_FOR_EXAM du Playbook.

Le playbook étant conçu manuellement, il se concentre davantage sur la classification que sur la description, le nombre de colonnes mappées reste réduit. Il est parfois difficile d'avoir une définition claire du rôle d'une colonne, par exemple MODALITY_MODIFIER qui pointe vers des classes de plusieurs types différents. La colonne REASON_FOR_EXAM est définie comme suit : « Informations sur l'indication clinique, le diagnostic du patient, l'état clinique (par exemple, postopératoire), une mesure prévue, une anatomie modifiée (par exemple, endogreffe), ou un autre objectif de l'étude (par exemple, dépistage) ». Il y a manifestement un chevauchement important entre les colonnes MODALITY_MODIFIER_X, REASON_FOR_EXAM_X et TECHNIQUE, résultant peut-être aussi du travail de fusion avec la LOINC.
Enfin, le Playbook RadLex n'est pas utilisé en France, ni aucun autre vocabulaire de procédures radiologiques, mais des projets de regroupement interinstitutionnel pourraient amener progressivement RadLex ou un autre vocabulaire standard à s'imposer.

### 3.5.2. La classification automatique

Cette méthode de classification, basée sur les données plutôt que sur les terminologies et s'appuyant sur le raisonnement ontologique, permet de faire une démonstration simple des capacités des ontologies à rendre un domaine comme la radiologie accessible à un algorithme. On exploite ici concrètement cette modélisation du domaine.

Certains cas d'alignement illustrent la capacité du raisonnement ontologique. Par exemple, un examen dans lequel nous avons trouvé le terme « mammo » était associé au RID « mammographie ». Ce dernier est une sous-classe du RadLex « projection radiographique » (RID10345) qui est une classe cible potentielle de la propriété d'objet has_MODALITY, l'instance de *procedure* nouvellement créée a donc été liée à une instance de

*mammographie* via has_MODALITY. Le raisonneur a classé l'examen avec le code de procédure « X-RAY » (RPID2501) car il avait une projection radiographique comme modalité.

Nous avons constaté que les métadonnées DICOM ne sont pas toujours renseignées ou fiables, et remplies de façon différentes selon les établissements. Ce problème de qualité des données DICOM provient du fait qu'elles sont peu utilisées par les soignants. Les champs de description sont ceux dont notre solution tire le plus d'information, ils ne sont cependant pas renseignés de la même façon dans les deux institutions. L'une utilise notamment des adjectifs comme « encéphalique » que notre système n'arrive pas à interpréter. Une solution plus poussée pour extraire l'information des descriptions pourrait améliorer notre système et le rendre plus généralisable. Enfin, ces descriptions indiquent rarement une absence, par exemple il est rare que l'absence d'injection de produit de contraste soit précisée, par défaut nous n'avons donc pas d'information sur le produit de contraste. Enfin, une évaluation plus poussée nécessiterait des données DICOM en français pré-annotées. L'évaluation actuelle ne permet d'estimer les capacités du système à classifier correctement que sur un ensemble très reteint de données.

## 3.6. Perspectives

Il existe de nombreuses façons d'améliorer notre ontologie, sa traduction et notre preuve de concept, mais l'approche axée sur les données donne déjà des résultats intéressants. L'utilisation d'un tel système offre de nouvelles opportunités dans le cadre d'un entrepôt de données de santé étant donné l'importance de l'interopérabilité et la place croissante de l'imagerie.

Le système d'intégration des données d'imagerie développé dans le cadre de cette thèse et présenté à la partie 3 de ce manuscrit a été conçu pour accueillir des méthodes d'enrichissement automatique des données, comme celui que pourra constituer un classifieur automatique basé sur le travail présenté ici. De plus, avec l'intégration de l'imagerie a eHOP, il sera possible d'appliquer la classification sur les données eHOP en prenant en entrée le rapport de l'examen en plus des métadonnées DICOM. Enfin, en améliorant et en validant la précision de notre programme, nous pouvons envisager d'autres utilisations. Si notre système est incapable de classer une procédure radiologique ou s'il la classe dans plusieurs classes très différentes, il se peut qu'il s'agisse d'un nouvel examen non encore couvert par le Playbook, ou que le fichier DICOM soit mal rempli. Avec une précision accrue, il pourrait être utilisé pour aider à détecter de nouvelles utilisations ou des erreurs.

Synthèse de la deuxième partie

Les problèmes de qualité des données de vie réelle sont un frein à leur réutilisation. Dans notre cas, on peut constater que les métadonnées d'imagerie sont parfois peu renseignées ou de façon différente d'un examen à l'autre (spécialité, constructeur du matériel, …)

Des méthodes existent pour enrichir ces données automatiquement (par un raisonnement ontologique par exemple, comme on a pu le montrer dans cette partie). Il peut y avoir plusieurs méthodes d'enrichissement pour un même type de données et ces méthodes peuvent évoluer, c'est pourquoi il est préférable, dans le cadre de l'utilisation secondaire des données, de disposer des données brutes. Les référentiels (terminologies, ontologies) du domaine de l'imagerie sont limités dans leur couverture des données même si de nouvelles approches (comme la post-coordination par exemple) apparaissent pour mieux les décrire. Pour cette raison aussi, il est nécessaire, quand ces données sont utilisées pour la recherche, de conserver les données d'origines pour pouvoir utiliser selon les études à mener le référentiel le plus adéquat.

Nous abordons à présent, dans la troisième partie de ce manuscrit, la question du trajet des données d'imagerie depuis le soin vers un entrepôt et sa mise en place.

# Troisième partie : Application pour la génération de cohorte

# 1. L’imagerie dans les entrepôts de données de santé

## 1.1. La réutilisation des données d’imagerie en santé

A la fin des années 90 et au début des années 2000, avec l’évolution de l’imagerie et l’adoption à large échelle du standard DICOM et des PACS, il est devenu possible d’intégrer l’imagerie dans la recherche de manière informatisée.

Des solutions sont apparues [128–132] pour aider les chercheurs à gérer différentes opérations nécessaires à la réalisation d’études sur les données d’imagerie. On peut citer par exemple les opérations :

- d’intégration de différents types de fichiers d’imagerie (DICOM ou fichier bruts)
- de désidentification des métadonnées, l’indexation des métadonnées
- de gestion des accès par rôles
- de stockage et transfert de données selon règlements en vigueur
- d’audit
- d’extraction des caractéristiques des images

La Figure 17 décrit la préparation des données d’imagerie, dont les étapes 2 à 5 nous concernent ici, et nécessitent les opérations évoquées précédemment.

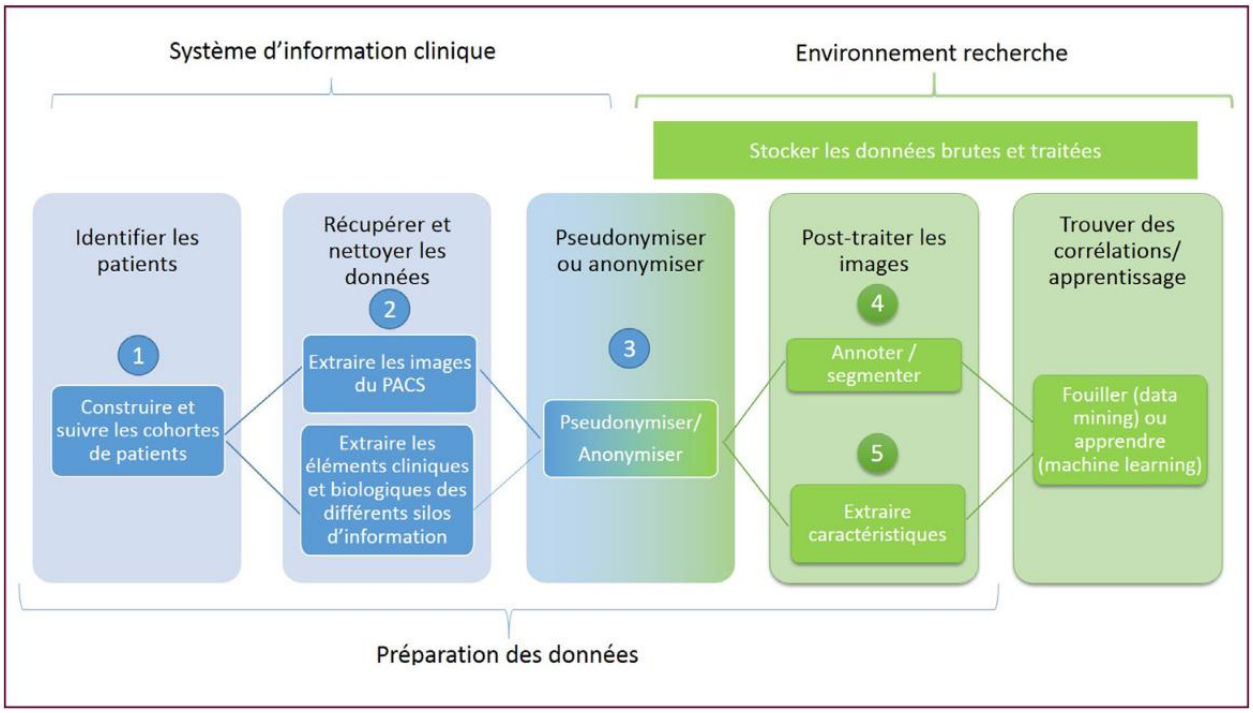



*Figure 17 : Le processus de préparation des données pour la recherche en imagerie* [46]

Certaines de ces solutions sont plutôt orientées vers la gestion des études sur un domaine précis, et proposent un environnement permettant de gérer les données d'imagerie, parfois en permettant de se connecter directement aux appareils d'acquisition. D'autres sont orientées vers l'intégration des données d'imagerie dans des entrepôts où elles seront rassemblées avec d'autres données clinico-biologiques dans le but de mener des études rétrospectives sur un large ensemble de données. Dans les deux cas, elles proposent souvent de faciliter le partage à plus large échelle. Un ensemble d'acteurs ont par ailleurs formalisé il y a quelques années un ensemble de bonnes pratiques pour la gestion et le partage des données, particulièrement applicable dans le cadre de la recherche, les principes FAIR.

Nous allons tout d'abord présenter les principes FAIR qui synthétisent un certain nombre de verrous que nous avons pu voir et permettent de comprendre comment ils peuvent être levés par des pratiques concrètes. Nous verrons ensuite les solutions (projets de recherches et industriels) apparues pour l'utilisation secondaire des données d'imageries, pour les dé-siloter puis pour les intégrer à des entrepôts. Enfin, nous aborderons les projets de partage à large échelle de ces données.

## 1.2. Présentation des principes FAIR

En 2015, un ensemble d'acteurs provenant du domaine universitaire, de l'industrie, des organismes de financement et les éditeurs scientifiques se sont réunis pour concevoir un ensemble de principes concis et mesurables que doivent respecter les données afin d'être utilisables au mieux dans un contexte de réutilisation pour la recherche [133] Ces principes appelés "FAIR" : Foundable, Accessible, Interoperable, Reusable (Requêtables, Accessibles, Interopérables et Réutilisables) représentent donc les objectifs à atteindre pour les données destinées à la recherche et mettent l'accent sur l'actionnabilité de la machine (c'est-à-dire la capacité des systèmes informatiques à trouver, accéder, inter-opérer et réutiliser des données avec un minimum, voire sans, intervention humaine)

Concernant l'interopérabilité, on en distingue en général trois types [134] : l'interopérabilité technique qui dont le but est de permettre la communication, l'interopérabilité syntaxique dont dépend la capacité à savoir communiquer et l'interopérabilité sémantique qui vise à savoir se comprendre. Le rapport entre la sémantique et la syntaxe est le même qu'entre les notions de fond et de forme. L'interopérabilité syntaxique se met en place par la définition d'unités d'information dans les flux de données échangés, par exemple dans notre cas via des normes telles que le protocole d'échange de DICOM ou HL7 FHIR. L'interopérabilité sémantique est la capacité des systèmes à échanger des données de manière significative [135]. Les systèmes doivent alors partager le même vocabulaire mutuellement compris ou créer des correspondances ou des mappages entre leurs différents vocabulaires [136].

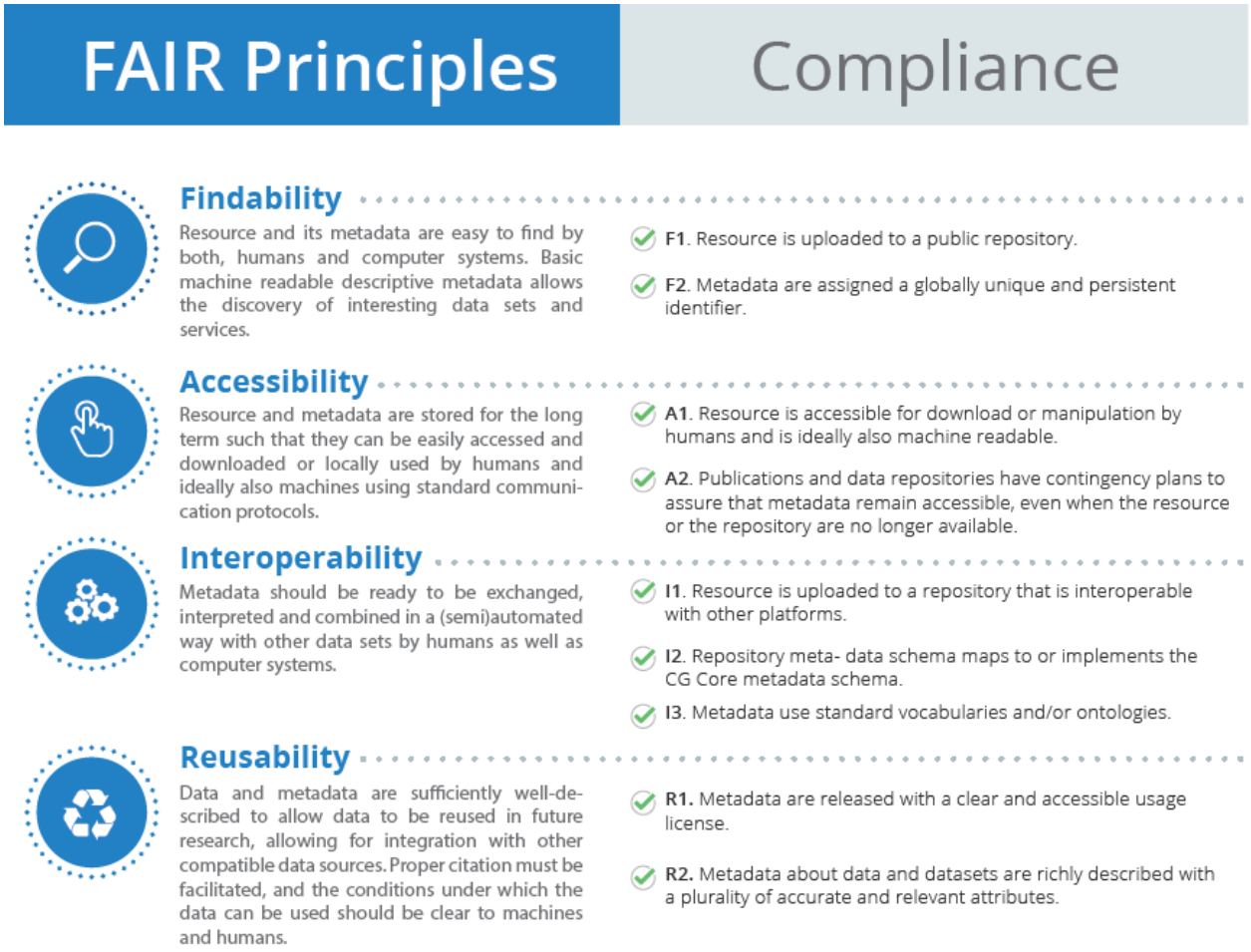



*Figure 18 : Les définitions des principes FAIR selon le "Consultative Group on International Agricultural Research" (CGIAR), un partenariat mondial réunissant des acteurs recherche en agroalimentaire* [137]

Nous reprenons ici les définitions des principes FAIR, qui sont autant de verrous à lever pour une intégration efficace des données d'imagerie [60,138,139] :

Findable - Facile à trouver

- Attribution à chaque élément de donnée (patient, document, images, prescription, etc.) d'un identifiant unique au monde et permanent. Des métadonnées sont indexées et peuvent être recherchées facilement
- Stockage de données dans des entrepôts accessibles via des protocoles standard.

Accessible

- Les données FAIR ne sont pas nécessairement des données ouvertes mais doivent être accessibles via un protocole de communication ouvert, libre, et d'usage

universel. Si nécessaire, ce protocole gère des procédures d'authentification et d'autorisation.

- Les métadonnées doivent rester accessibles même si les données ne le sont plus.
- Les éléments de métadonnées essentiels, recommandés et facultatifs doivent pouvoir être traités et vérifiés par des machines.

Interopérables

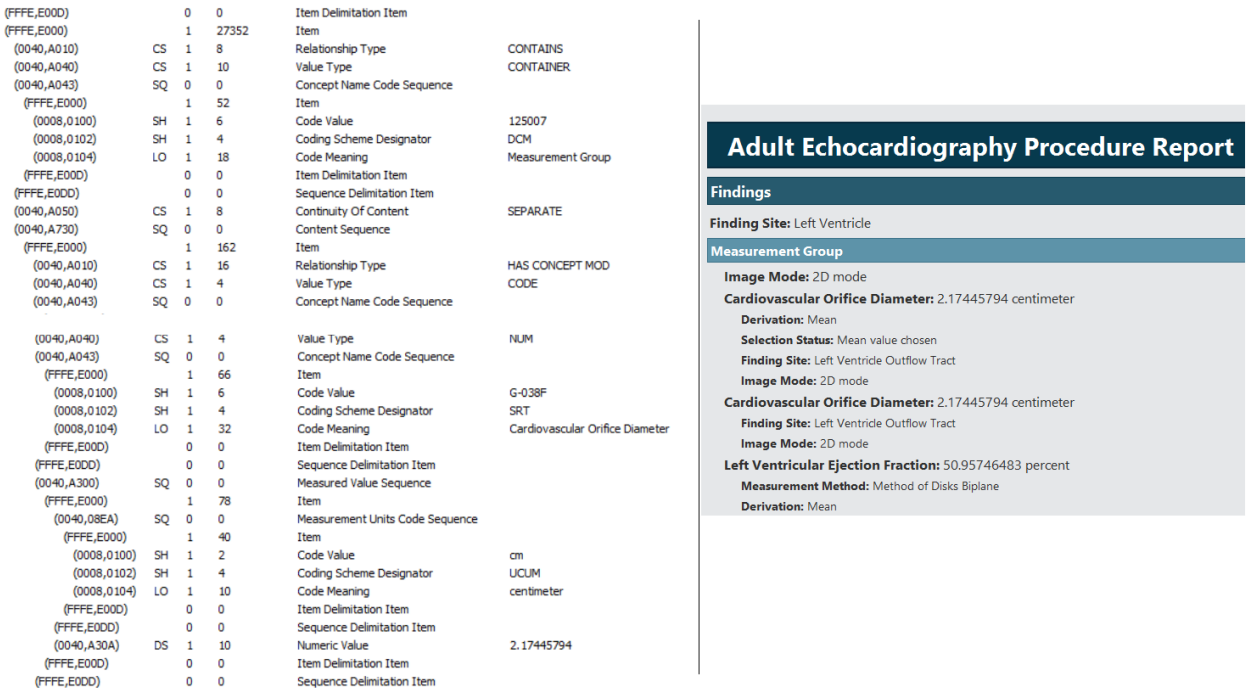
- La description des éléments de métadonnées doit suivre les directives de la communauté qui utilisent un vocabulaire ouvert et bien défini (interopérabilité sémantique).
- Les données et les métadonnées doivent pouvoir être récupérées dans une variété de formats accessibles aux humains et aux machines (interopérabilité syntaxique).

Réutilisables

- Description des données à l'aide de métadonnées riches, qui doivent au maximum être conservées.
- L'origine des données est renseignée et les données sont "citables" pour soutenir le partage des données et reconnaître leur valeur.
- Les données sont accessibles à des conditions claires, en utilisant des licences comme Creative Commons par exemple

## 1.3. Les rapports structurés DICOM

Les rapport structurés DICOM (DICOM structured reporting ou DICOM SR) sont des structures de données couvrant le domaine des observations d'imagerie incluses dans les rapports cliniques [140]. Ces observations d'imagerie concernent principalement les modalités d'images DICOM, les images dérivées, les résultats de segmentation, les mesures (taille, volume, surface, etc.), les cartes paramétriques, etc. Les données de DICOM SR sont sérialisées en syntaxe DICOM (triplet, <tag, longueur, valeur>) et organisées sous forme d'arbre [140,141] comme on peut le voir sur la figure 19 ci-dessous.

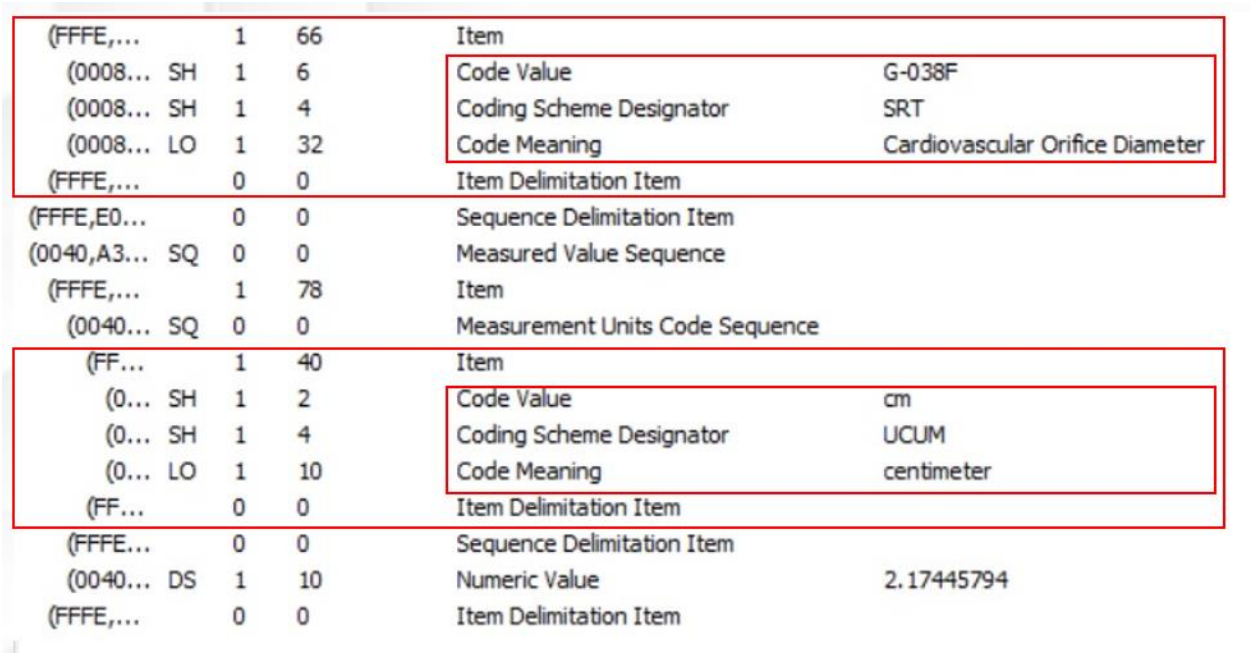


| | | | | | |
|---|---|---|---|---|---|
| (FFFE,E00D) | | 0 | 0 | Item Delimitation Item | |
| (FFFE,E000) | | 1 | 27352 | Item | |
| (0040,A010) | CS | 1 | 8 | Relationship Type | CONTAINS |
| (0040,A040) | CS | 1 | 10 | Value Type | CONTAINER |
| (0040,A043) | SQ | 0 | 0 | Concept Name Code Sequence | |
| (FFFE,E000) | | 1 | 52 | Item | |
| (0008,0100) | SH | 1 | 6 | Code Value | 125007 |
| (0008,0102) | SH | 1 | 4 | Coding Scheme Designator | DCM |
| (0008,0104) | LO | 1 | 18 | Code Meaning | Measurement Group |
| (FFFE,E00D) | | 0 | 0 | Item Delimitation Item | |
| (FFFE,E0DD) | | 0 | 0 | Sequence Delimitation Item | |
| (0040,A050) | CS | 1 | 8 | Continuity Of Content | SEPARATE |
| (0040,A730) | SQ | 0 | 0 | Content Sequence | |
| (FFFE,E000) | | 1 | 162 | Item | |
| (0040,A010) | CS | 1 | 16 | Relationship Type | HAS CONCEPT MOD |
| (0040,A040) | CS | 1 | 4 | Value Type | CODE |
| (0040,A043) | SQ | 0 | 0 | Concept Name Code Sequence | |
| (0040,A040) | CS | 1 | 4 | Value Type | NUM |
| (0040,A043) | SQ | 0 | 0 | Concept Name Code Sequence | |
| (FFFE,E000) | | 1 | 66 | Item | |
| (0008,0100) | SH | 1 | 6 | Code Value | G-038F |
| (0008,0102) | SH | 1 | 4 | Coding Scheme Designator | SRT |
| (0008,0104) | LO | 1 | 32 | Code Meaning | Cardiovascular Orifice Diameter |
| (FFFE,E00D) | | 0 | 0 | Item Delimitation Item | |
| (FFFE,E0DD) | | 0 | 0 | Sequence Delimitation Item | |
| (0040,A300) | SQ | 0 | 0 | Measured Value Sequence | |
| (FFFE,E000) | | 1 | 78 | Item | |
| (0040,08EA) | SQ | 0 | 0 | Measurement Units Code Sequence | |
| (FFFE,E000) | | 1 | 40 | Item | |
| (0008,0100) | SH | 1 | 2 | Code Value | cm |
| (0008,0102) | SH | 1 | 4 | Coding Scheme Designator | UCUM |
| (0008,0104) | LO | 1 | 10 | Code Meaning | centimeter |
| (FFFE,E00D) | | 0 | 0 | Item Delimitation Item | |
| (FFFE,E0DD) | | 0 | 0 | Sequence Delimitation Item | |
| (0040,A30A) | DS | 1 | 10 | Numeric Value | 2.17445794 |
| (FFFE,E00D) | | 0 | 0 | Item Delimitation Item | |
| (FFFE,E0DD) | | 0 | 0 | Sequence Delimitation Item | |



*Figure 19 : Les DICOM Structured Reports sont organisés hiérarchiquement via les attributs "Sequence" et "Item". Des logiciels adaptés permettent de visualiser le rapport*

Cette description standardisée des observations d'imagerie repose sur la définition de "SR templates" (modèles de compte rendus structurés). Ces modèles décrivent l'organisation des données en spécifiant les concepts et les relations codés ainsi que leurs significations. Des ressources terminologiques externes comme la SNOMED CT ou Radlex sont souvent utilisées pour coder les informations contenues dans ces rapports (cf Figure 20); quand les termes ne sont pas disponible dans ces ressources, la terminologie propre à DICOM, appelée "DICOM controlled terminology" (DCT), est utilisée.



| | | | | | |
|---|---|---|---|---|---|
| (FFFE,... | | 1 | 66 | Item | |
| (0008... | SH | 1 | 6 | Code Value | G-038F |
| (0008... | SH | 1 | 4 | Coding Scheme Designator | SRT |
| (0008... | LO | 1 | 32 | Code Meaning | Cardiovascular Orifice Diameter |
| (FFFE,... | | 0 | 0 | Item Delimitation Item | |
| (FFFE,E0... | | 0 | 0 | Sequence Delimitation Item | |
| (0040,A3... | SQ | 0 | 0 | Measured Value Sequence | |
| (FFFE,... | | 1 | 78 | Item | |
| (0040... | SQ | 0 | 0 | Measurement Units Code Sequence | |
| (FF... | | 1 | 40 | Item | |
| (0... | SH | 1 | 2 | Code Value | cm |
| (0... | SH | 1 | 4 | Coding Scheme Designator | UCUM |
| (0... | LO | 1 | 10 | Code Meaning | centimeter |
| (FF... | | 0 | 0 | Item Delimitation Item | |
| (FFFE... | | 0 | 0 | Sequence Delimitation Item | |
| (0040... | DS | 1 | 10 | Numeric Value | 2.17445794 |
| (FFFE,... | | 0 | 0 | Item Delimitation Item | |

*Figure 20 : Exemple d'utilisation de terminologies externes dans un DICOM Structured Report. Ici la terminologie SNOMED est utilisée (Coding Scheme Designator "SRT") pour indiquer la nature de la mesure "Cardiovascular Orifice Diameter" avec le code "G-038F". La terminologie UCUM est utilisée pour les unités de mesure, ici le centimètre.*

## 1.4. Le protocole de communication de DICOM

DICOM spécifie un protocole pour l'échange de messages qui fournit le cadre de communication pour les services DICOM. Le protocole DICOM est compatible avec les protocoles TCP (Transmission Control Protocol) et IP (Internet Protocol) permettant aux applications DICOM de communiquer sur Internet [142].

Ce protocole définit des services qui peuvent être échangés entre deux machines comme la vérification (Test de communication), le stockage d'images ou l'échange de listes d'examens. Chaque équipement pouvant être « Fournisseur » du service (SCP = Service Class Provider) et/ou « Utilisateur » du service (SCU = Service Class User). Un équipement fournit un certain nombre de services, et peut utiliser un certain nombre de services sur d'autres machines.

Une classe SOP (Service-Object-Pair) DICOM spécifie la combinaison d'un IOD et de l'ensemble des services qui sont utiles dans un but donné. Les classes SOP (telles que la classe SOP Basic Modality Worklist) sont spécifiées dans les classes de service en fonction de leur objectif [142].

Les transactions de messages utilisant DICOM commencent par l'établissement d'une association, c'est-à-dire un canal ouvert pour l'échange de messages entre deux appareils qui utilisent le protocole DIMSE (DICOM Message Service Element) pour générer et recevoir des messages DICOM. Lors de l'établissement de l'association, les deux dispositifs s'accordent sur une compréhension commune des structures d'information qui seront échangées et des services qui seront invoqués. Des paramètres supplémentaires essentiels à l'interopérabilité, tels que l'ordre des octets et la méthode de compression des données, sont également négociés. Les classes de services DICOM prennent en charge plusieurs domaines d'application généraux dont la gestion des images sur le réseau (Network image management) qui nous concerne particulièrement, et illustré sur la figure 21.

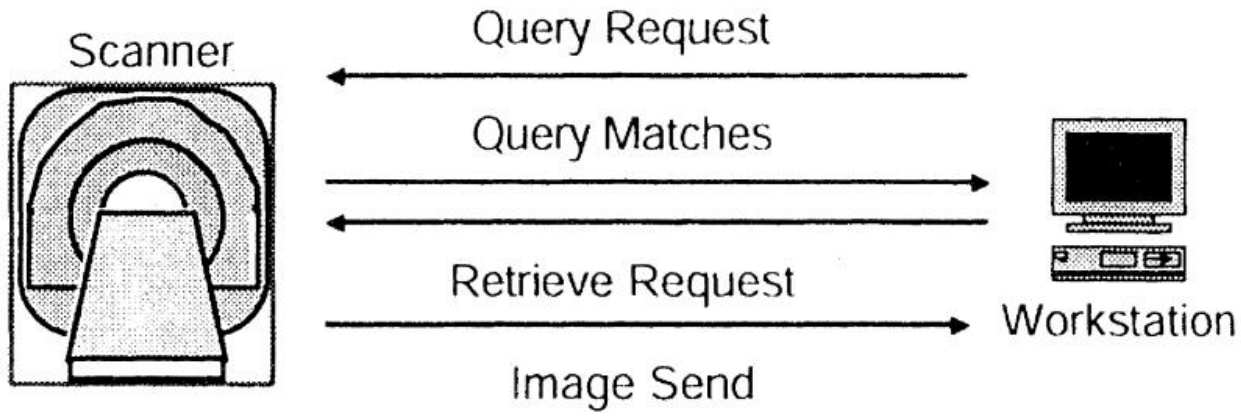



*Figure 21 : Recherche et récupération d'images (Query/Retreive). Un logiciel (ici une station de travail, à droite) envoie un message demandant des images dont les métadonnées correspondent à un ensemble de clé/valeurs fournies. Le scanner renvoie la liste des identifiants des images correspondantes. Connaissant ces identifiants, l'utilisateur de la station de travail sélectionne les images pertinentes dans la liste affichée et envoie une commande pour récupérer les images (Retreive). Le logiciel de la station de travail envoie alors un message au scanner, énumérant les identifiants des images demandées. Le scanner envoie les images, une par une, à la station de travail, en utilisant le service de stockage DICOM (Store)*[142]

DICOM ne précise pas les mécanismes de vérification ou de test à mettre en œuvre pour vérifier la conformité mais toutes les implémentations DICOM doivent être détaillées par une déclaration de conformité (conformance statement). Un utilisateur averti peut savoir si deux appareils sont interopérables en lisant ces déclarations de conformité.

# 2. Conception d'un système d'intégration des données d'imagerie

## 2.1. Problématique et objectifs

L'équipe Données Massives en Santé DOMASIA du laboratoire LTSI de l'INSERM, dans laquelle je réalise ma thèse, a développé une solution d'entrepôt de données biomédicales nommée eHOP [143]. eHOP peut intégrer des données structurées ou non structurées comme les données textuelles et met en œuvre des méthodes de recherche d'information et traitement automatique du langage. Une interface de requête permet de créer des cohortes de patients et faire du pré-screening en utilisant des critères de recherche portant sur les dates, types d'examens, valeurs des résultats numériques, etc.

Dans un premier cas d'usage concernant l'imagerie en cardiologie, il a été nécessaire de retrouver des examens au format DICOM depuis le PACS afin de croiser ces données avec celles de l'entrepôt. Des scripts ad hoc ont été conçu pour filtrer les données DICOM exportées sur leurs métadonnées puis extraire les données d'intérêt. Cette expérience a permis de mettre en évidence la nécessiter d'intégrer les données d'imagerie directement à eHOP en permettant d'effectuer des requêtes sur un ensemble le plus large possible de métadonnées et a donné des pistes sur la façon de la mettre en œuvre pour répondre aux futurs cas d'usage.

Afin de pouvoir répondre aux demandes des chercheurs désirant utiliser les données d'imageries, nous avons décidé de mettre eHOP en capacité d'intégrer les données d'imagerie. Cette tâche comporte principalement deux volets.

Le volet technique regroupe les problématiques de connexion et interfaçage avec les outils du soin, de performance dans la récupération des données sans influencer le fonctionnement des outils liés au soin et l'évolution du logiciel eHOP. Concernant l'interface avec le soin, nous avons conçu un outil autonome, que nous avons appelé "Imaging For Data Warehouse" (I4DW) et que nous présentons en détail dans l'article suivant. L'évolution du logiciel eHOP passe par l'adaptation de son interface, avec la possibilité de visualiser les images relatives aux examens consultés, des modifications dans le modèle de données pour représenter le lien vers les images d'origine et la mise en place d'un trajet des images jusqu'à l'utilisateur.

Le volet sémantique concerne la représentation, au sein d'eHOP, des données d'imagerie intégrées en préservant le sens de ces données et en rendant les données d'imagerie

manipulables via le requêteur d'eHOP. L'objectif dans cette partie est de proposer un moyen efficace aux utilisateurs finaux d'effectuer des recherches sur les données potentiellement complexes de l'imagerie et de pouvoir partager les données entre centres si besoin en conservant toutes les informations des examens avec leur lien vers des terminologies de référence. Ce travail a fait l'objet d'un article présenté à la conférence MedInfo 2021, en Annexe 3 de ce manuscrit.

## 2.2. Etat de l'art

### 2.2.1. Les solutions dédiées à l'imagerie

En 2002, Wong et al [132] conçoivent un framework d'entrepôt de données dédié à la neuroimagerie permettant de travailler sur des données extraites des rapports textuels, les métadonnées DICOM, les images et des features extraites de ces dernières. Depuis, plusieurs systèmes plus ou moins élaborés ont été conçus pour mener ce genre d'études, dont on peut citer :

- eXtensible Neuroimaging Archive Toolkit (XNAT) [130], est une plateforme logicielle d'imagerie open source développée par le groupe de recherche en neuro-informatique de l'université de Washington. XNAT facilite les tâches communes de gestion, de productivité et d'assurance qualité pour l'imagerie et les données associées. Grâce à son extensibilité, XNAT peut être utilisé pour soutenir un large éventail de projets basés sur l'imagerie. Cette solution est tournée vers le partage des données entre institutions puis l'open-data. Plusieurs solutions de réutilisation de données d'imagerie utilisent XNAT comme socle [129].
- Archimed [144], conçu au Centre d'investigation clinique et d'innovation technologique (CIC-IT) de Nancy propose un modèle calqué sur celui de DICOM, orienté étude. Respectant le protocole DICOM pour le transfert de fichiers d'imagerie, il peut se connecter directement aux équipements d'imagerie.
- SHANOIR (*SHA*ring *N*eur*OI*maging *R*esources, Next Generation) [145], une plateforme web (open-source) extensible et customisable pour la recherche clinique et préclinique créée par l'équipe de recherche Empenn (signifiant "cerveau" en Breton) conjointement affiliée à l'Inria et l'Irisa. Shanoir-NG a un modèle de données basé sur l'ontologie OntoNeuroLog [146] facilitant la réutilisation et l'intégration des données et permet de fédérer plusieurs instances de la plateforme. La solution est donc tournée vers le partage de données et les études multi-centriques.

- CATI [147] (Centre pour l'Acquisition et le Traitement des Images) initié en 2010 par le plan Alzheimer français (2008-2012), est une plateforme nationale conçue pour soutenir les études de neuro-imagerie multicentriques à grande échelle , de l'acquisition des images aux résultats d'analyse. En étroite collaboration avec les sociétés françaises de radiologie et de médecine nucléaire, CATI représente un réseau de plus de 50 sites français.
- Langer et al. [148,149] ont développé un Entrepôt de données DICOM qui intègre les données DICOM d'un PACS ou de l'appareil d'acquisition via les outils de l'hôpital, les stocke et fournit leurs métadonnées pour une analyse complète. Ce travail met l'accent sur l'harmonisation des métadonnées renseignées par les fournisseurs (constructeurs de matériel d'imagerie) différents.

On peut noter que le domaine de la neuroimagerie a été un terreau fertile pour la conception de ces solutions car XNAT, Shanoir et CATI-DB en sont issues ; XNAT et SHANOIR ayant évolué pour traiter aujourd'hui tous les organes.

### 2.2.2. Les solutions d'intégration à un entrepôt

Dès 2002, Cohen et al. [150] décrivent l'intégration et l'utilisation des métadonnées DICOM provenant d'un PACS via HL7 et XML dans le dossier médical électronique. Rubin. L et al proposent en 2008 le système RadBank [151] qui extrait automatiquement des données structurées des rapports d'examens d'imagerie pour les ajouter à un entrepôt de données de santé. Ils évoquent dans leurs perspectives la valeur des images elles-mêmes, indiquant qu'il serait possible de "lier les données dans RadBank aux images du PACS en stockant une référence aux images avec les données du rapport". Cette approche sera mise en œuvre plus tard par plusieurs solutions.

- Radiomics Enabler ® [152] par exemple, développée par Medexprim [153] en collaboration avec des chercheurs et des cliniciens du CHU de Toulouse est une application web qui permet à un utilisateur d'effectuer une recherche multicritères sur le PACS, de filtrer les résultats et de sélectionner les acquisitions pertinentes. Une API (Application Programming Interface) standard, permet d'intégrer les données dans un entrepôt de données cliniques afin de créer des cohortes de patients avec un accès complet à leurs données cliniques et d'imagerie.
- A l'hôpital universitaire de Tampere, en Finlande, a été développé le Tampere Research Archival System (TARAS) [154], composé d'un PACS de recherche et d'un

système d'archivage des données de recherche (CRDAS) qui intègre des données cliniques sur les patients provenant d'autres branches médicales.

- Murphy et al. [155] ont développé un module pour l'entrepôt i2b2, mi2b2, qui permet une extraction interactive d'images à partir de plusieurs PACS, sur la base d'une sélection de patients effectuée lors d'étapes précédentes. Ils n'ont pas intégré les résultats récupérés à partir du PACS directement dans un DWH, mais fournissent des outils pour transférer les images vers un dépôt dédié à l'étude, par exemple une instance XNAT.
- Dans [4], les auteurs ont mis en place un ensemble de solutions logicielles assemblées en modules pour réaliser les différentes tâches du data reuse en imagerie : sélection, anonymisation, suppression des caractéristiques faciales, contrôle qualité des métadonnées et des images, mais aussi extraction quantitative de biomarqueurs d'imagerie.
- Enfin, en 2020, Kaspar et al [156] ont réalisé la mise en œuvre d'un composant technique permettant d'intégrer les métadonnées d'imagerie du PACS clinique dans un entrepôt de données. Ils ont prouvé la faisabilité d'une alimentation de routine via des requêtes sur un ensemble restreint des principales métadonnées DICOM. Cependant, sur des ensembles de patients identifiés pour une étude, il est possible de retrouver l'exhaustivité des métadonnées. Nous reviendrons sur ces travaux, dont le contexte et les objectifs sont similaires aux nôtres, dans le troisième chapitre de cette thèse.

Aucune des solutions existantes ne permets de faire une requête sur l'ensemble des métadonnées DICOM. En effet, elles proposent au mieux de faire une requête de type « CFIND » c'est-à-dire portant sur une dizaine d'attributs DICOM spécifiques, indexés dans les PACS. La duplication des données d'imagerie elles-mêmes posent des problèmes techniques et de sécurité que nous ne pouvons pas gérer. Nous avons donc fait le choix d'une solution d'intégration systématiques des métadonnées à eHOP en conservant les identifiants des images afin d'y accéder ensuite ponctuellement ou par lot (lors d'un export). Nous détaillons l'implémentation de notre système dans les parties qui suivent.

## 2.3. Méthode et Résultats

La première étape a été de lister les besoins des spécialistes (radiologues, cliniciens et data scientists) afin de concevoir un premier cahier des charges. Le premier de ces besoins a été l'indexation de toutes les métadonnées disponibles afin de ne pas être limité dans les requêtes sur l'imagerie et donc dans les possibilités d'études. Nous avons donc choisi de récupérer les fichiers DICOM entièrement et de parser leur contenu pour extraire

l'intégralité des métadonnées. Le besoin de voir les images sur l'interface eHOP a également été abordé, ce qui a nécessité des développements sur eHOP afin d'adapter son interface. Parmi les besoins évoqués, on trouve aussi la possibilité d'enrichir les données en créant des variables agrégées, composées à partir d'autre métadonnées, qui nous a conduit à proposer un système de plugin permettant de définir des traitements à appliquer aux données lorsqu'elles transitent par le module.

Le module est donc avant tout une passerelle entre le PACS et l'environnement de l'entrepôt (l'outil d'ETL pour charger les métadonnées dans la base de données d'eHOP et l'application eHOP pour y afficher les images). Il permet aux outils de recherche (utilisant des technologies actuelles : http, REST) de solliciter facilement le PACS utilisant des protocoles qui n'ont pas été conçus de façon suffisamment flexible pour répondre aux besoins de la recherche en imagerie [101,155]. La figure 22 présente les trajets des métadonnées chargées en routine d'une part et des images transférées ponctuellement lorsqu'un utilisateur veut les afficher d'autre part.

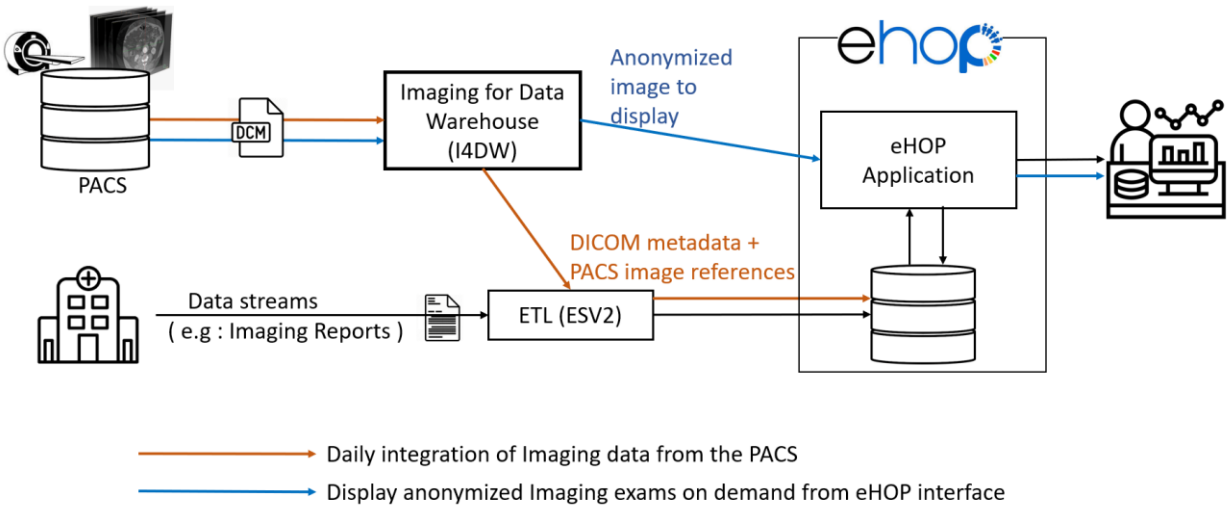



*Figure 22 : Le fonctionnement global du module imagerie. En routine, l'ETL interroge le module pour récupérer des métadonnées et les intégrer. Lorsqu'un utilisateur veut voir une image, l'application eHOP interroge le module pour obtenir les fichiers d'imagerie à afficher. Ces derniers sont anonymisés dans le module*

### 2.3.1. Intégration au modèle de donnée eHOP

Le modèle de données eHOP est orienté document (cf Figure 23). Les patients sont liés à des séjours pendant lesquels des documents sont générés. Les documents ayant un contenu textuel (non structuré) sont indexés avec les fonctions d'Oracle texte ce qui permet de les retrouver à partir de recherches textuelles plus ou moins élaborées. Quand des informations structurées peuvent être extraites des documents, elles sont enregistrées dans une table dédiée à ces éléments structurés où elles sont alignées avec des terminologies adéquates.

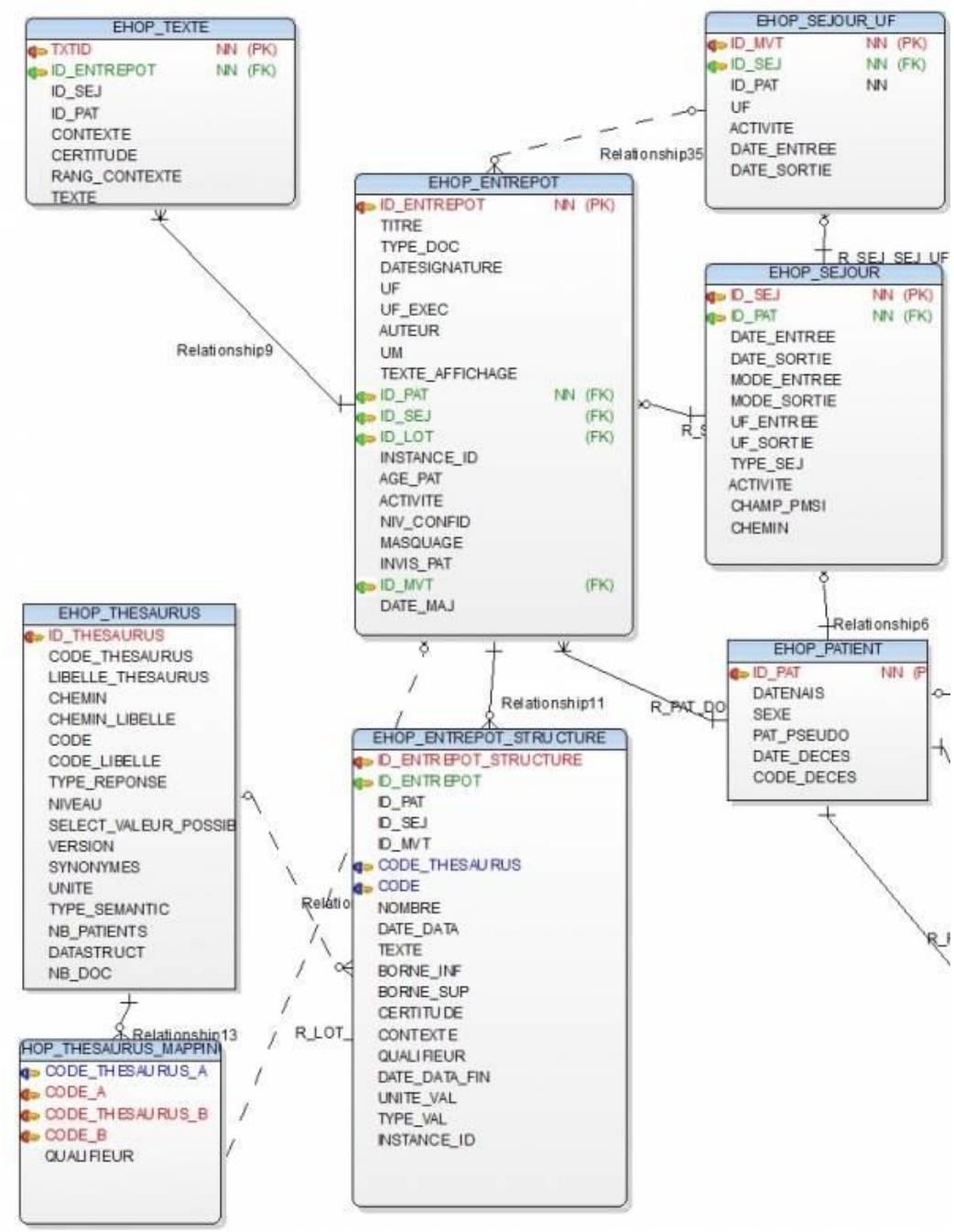



*Figure 23 : le modèle eHOP orienté autour du document. EHOP_PATIENT et EHOP_SEJOUR contiennent respectivement les patients et les séjours. EHOP_ENTREPOT est la table des documents. Les parties non structurées (texte libre) sont indexées dans EHOP_TEXTE alors que les éléments structurés sont stockés dans la table EHOP_ENTREPOT_STRUCTURE et alignés avec les terminologies maintenues dans EHOP_THESAURUS*

Comme on l'a vu, un examen d'imagerie produit un rapport textuel qui est déjà intégré dans eHOP (compte rendu Xplore) et des acquisitions (series DICOM) rassemblés dans le PACS en tant que *DICOM study*. En se basant sur le compte-rendu, qui contient un code identifiant

de l'examen, on retrouve la study DICOM correspondante dans le PACS et on charge chacune des series (acquisition d'imagerie ou rapport structuré) qu'elle contient comme un nouveau document eHOP. Les documents représentant l'examen, c'est-à-dire le compte-rendu et les series, sont reliés entre eux dans eHOP pour permettre l'affichage d'une vue commune où ils seront tous rassemblés (que nous verrons dans la partie suivante sur l'interface). La figure 24 ci-dessous présente cette organisation des documents d'imagerie dans eHOP.

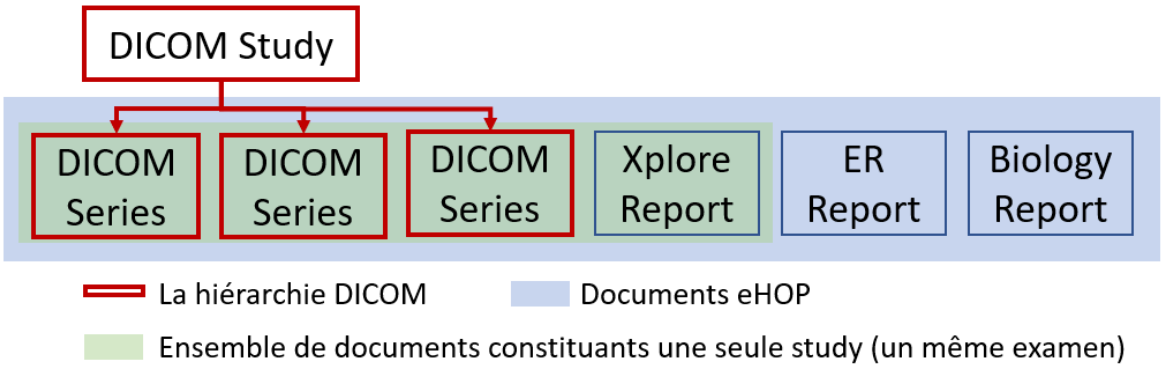



*Figure 24 : Organisation des documents d'imagerie dans eHOP*

Comme expliqué précédemment, le module I4DW est indépendant d'eHOP, c'est via son système de plugin qu'il permet de s'adapter facilement et d'extraire et de transformer les données brutes DICOM selon les besoins. C'est donc un plugin dédié à eHOP qui est ajouté dans I4DW et qui permet de passer des données DICOM au fichier d'intégration pivot propre à eHOP.

### 2.3.2. Ajout à l'interface d'eHOP

Les données générées par un examen d'imagerie sont les fichiers d'imagerie DICOM stockés dans le PACS mais aussi le compte rendu de l'examen (cf Figure 25). Les comptes rendus d'examens sont des documents rédigés principalement en texte libre et sont intégrés dans eHOP depuis 2004. Dans eHOP, nous voulons permettre à l'utilisateur d'avoir une vue au niveau de l'examen. Par conséquent nous avons ajouté la possibilité pour tous les documents liés à l'imagerie (série ou compte rendu), d'afficher une vue rassemblant l'ensemble des documents du même examen (cf Figure 26).

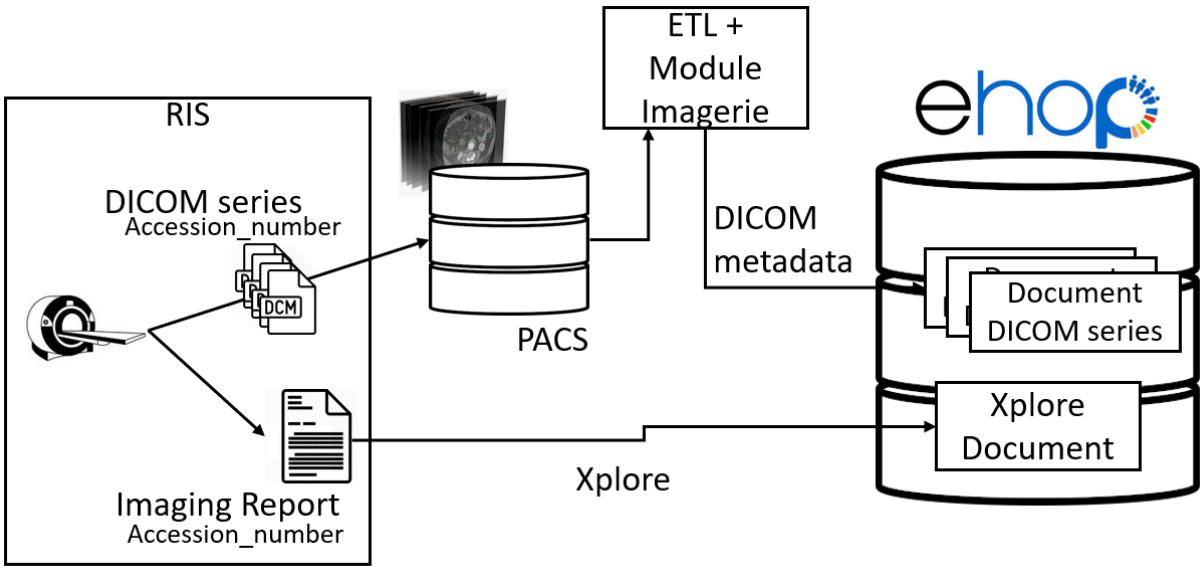



*Figure 25 : Un même examen d'imagerie produit des données qui sont stockées différemment, les données DICOM dans le PACS et le compte rendu dans la solution de l'hôpital, en l'occurrence le logiciel Xplore. Lors de l'intégration dans eHOP, nous voulons à nouveau rassembler tous ces éléments*

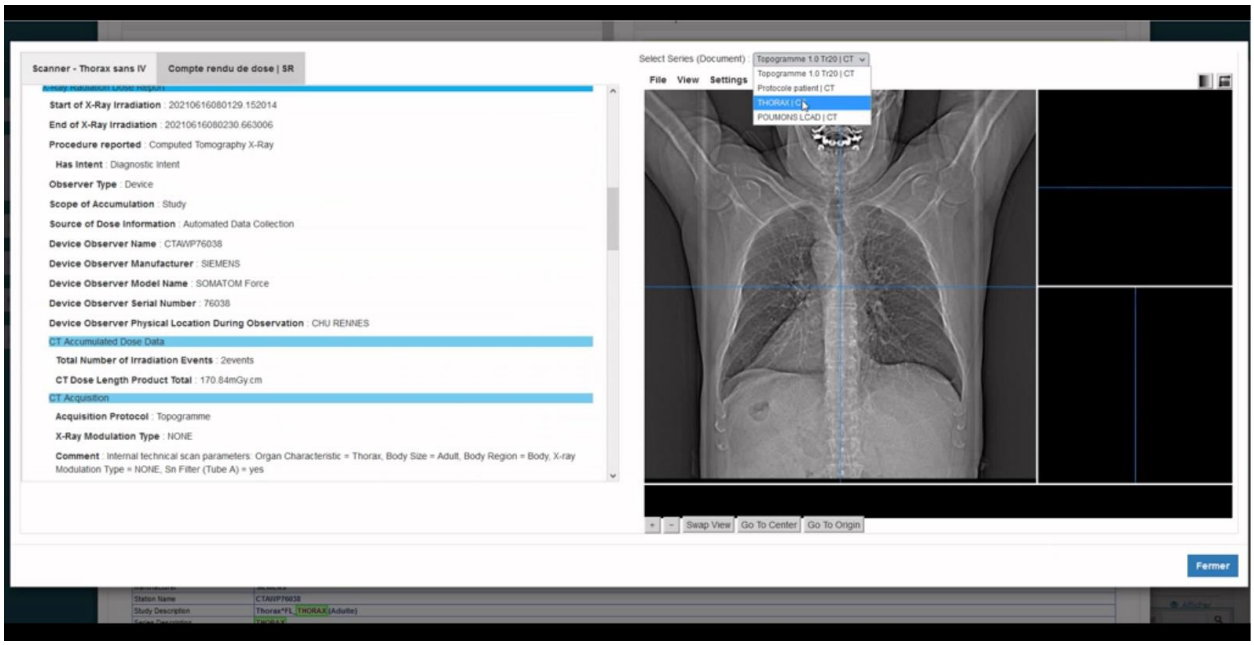



*Figure 26 : La vue "DICOM Study" permet d'accéder au contenu textuel dans la partie gauche (rapport structurés et compte rendus d'imagerie) tout en observant les acquisitions d'imagerie dans la partie de droite*

### 2.3.3. Adaptation de la terminologie DICOM

Comme nous l'avons vu dans la partie 1 de ce travail, les données DICOM peuvent être organisées de façon complexe, en particulier dans les rapport structurés (Structured Report). Ainsi un même attribut DICOM, selon sa position dans le document, peut avoir un sens différent. Afin de permettre aux utilisateurs de manipuler ces données, nous ne pouvions pas utiliser l'ensemble des attributs DICOM comme une simple terminologie et nous avons dû conserver le contexte de chaque attribut lors de l'intégration. Pour cela, nous avons défini dans eHOP une terminologie que nous avons appelée DCMEHOP construite au fur et à mesure de l'intégration de nouvelles données dans eHOP comme cela est représenté sur la figure 27.

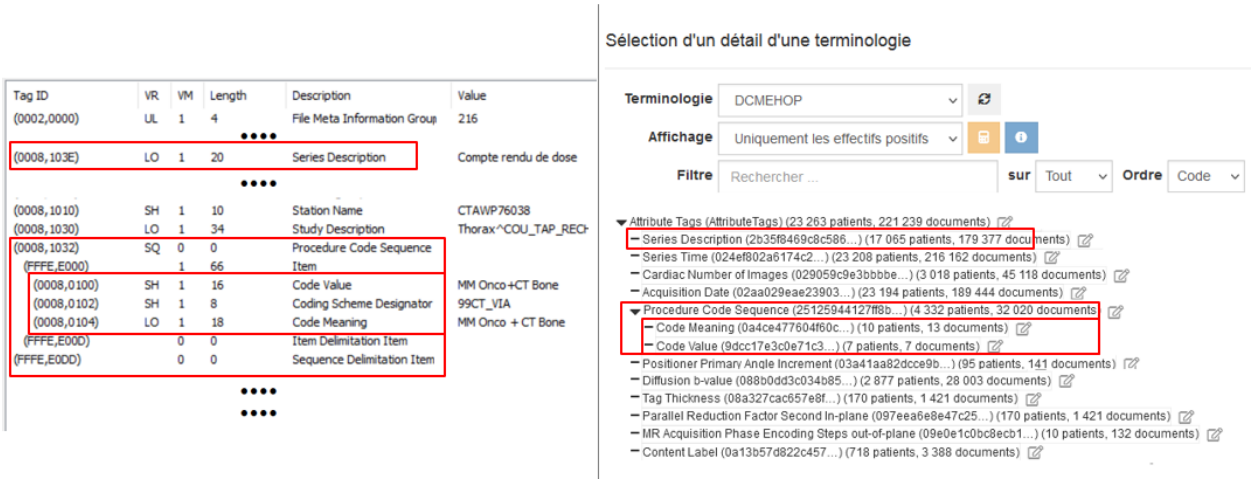



*Figure 27 : À gauche un extrait de fichier DICOM présentant les attributs, organisés hiérarchiquement via un attribut de type "sequence". A droite, on constate que la terminologie DCMEHOP a elle-même intégré ces attributs et cette organisation*

Dans les séries DICOM contenant de l'imagerie, les métadonnées sont presque toutes alignées sur le même niveau, on y trouve la modalité (le type d'examen : IRM, Scanner ou autre) la description de l'examen (Study Description) de la série (Series Description). Parmi les exceptions, c'est-à-dire les attributs de type « séquence » qui créent un niveau de hiérarchie en contenant eux même plusieurs attributs, on peut citer l'attribut « Reason For Performed Procedure Code Sequence » qui contient les raisons de l'examen. Cependant, certaines séries sont de type « Rapport structuré » ( Structured Report « SR ») et ne contiennent pas d'image mais un rapport des mesures effectuées par la machine ou le clinicien, organisé hiérarchiquement par des attributs de type « séquence » imbriqués les uns dans les autres sur plusieurs niveaux.

Les attributs DICOM et les rapports structurés étant variés (de nombreux « templates » de rapports existent et sont créés régulièrement[157]) nous avons fait le choix de ne pas les décrire à priori dans notre terminologie mais d'alimenter cette dernière au fur et à mesure. Pour conserver la hiérarchie et donc le contexte de chaque attribut, les codes de tous les attributs de séquence au-dessus de lui ainsi que le sien sont concaténés dans une chaine de caractères qui est ensuite hachée (en utilisant la fonction de hachage cryptographie sha256) pour obtenir une chaine dont on conserve les 50 premiers, cette dernière devient son code dans la terminologie DCMEHOP. On s'assure ainsi que chaque code est unique mais sera le même pour un même attribut à la même position dans le fichier DICOM, peu importe la provenance du fichier.

Ainsi, lors de l'intégration d'une série DICOM, le fichier d'intégration généré par le plugin eHOP contient aussi l'ensemble des nouveaux éléments à ajouter à la terminologie DCMEHOP.

### 2.3.4. Anonymisation

Afin de pouvoir afficher ou exporter les images provenant du PACS, les images sont anonymisées dans le module après avoir été récupérées. Les métadonnées sont anonymisées automatiquement selon les règles du "Basic Application Level Confidentiality Profile" défini par la norme DICOM [158]. Nous avons pour cela utilisé les capacités de la librairie PixelMed mais de nombreuses solutions existent pour mener ce travail de désidentification [159]. Pour le contenu des images, un système de profils d'anonymisation permet de définir quelles images vont être modifiées (des rectangles noirs sont incrustés dans l'image aux endroits susceptibles de contenir des information identifiantes). Ce système simple nécessite que l'utilisateur en charge des profils ait une bonne connaissance des images stockées dans le PACS et ne permet pas de se passer d'une revue manuelle des images lors d'un export.

Pour pouvoir créer facilement des filtres sur les métadonnées DICOM, nous avons créé une grammaire dédiée sous la forme d'une libraire Java. Cette grammaire permet d'exprimer des contraintes (numériques ou textuelles avec des expressions régulières) sur les attributs DICOM et des relations logiques entre ces contraintes. Les attributs DICOM peuvent y être identifiés par leur code ou leur nom (e.g. « (0008,1030) » ou « #StudyDescription » sont équivalent) afin de faciliter l'utilisation par des non spécialistes. La documentation de cette librairie est en Annexe 4 de ce document.

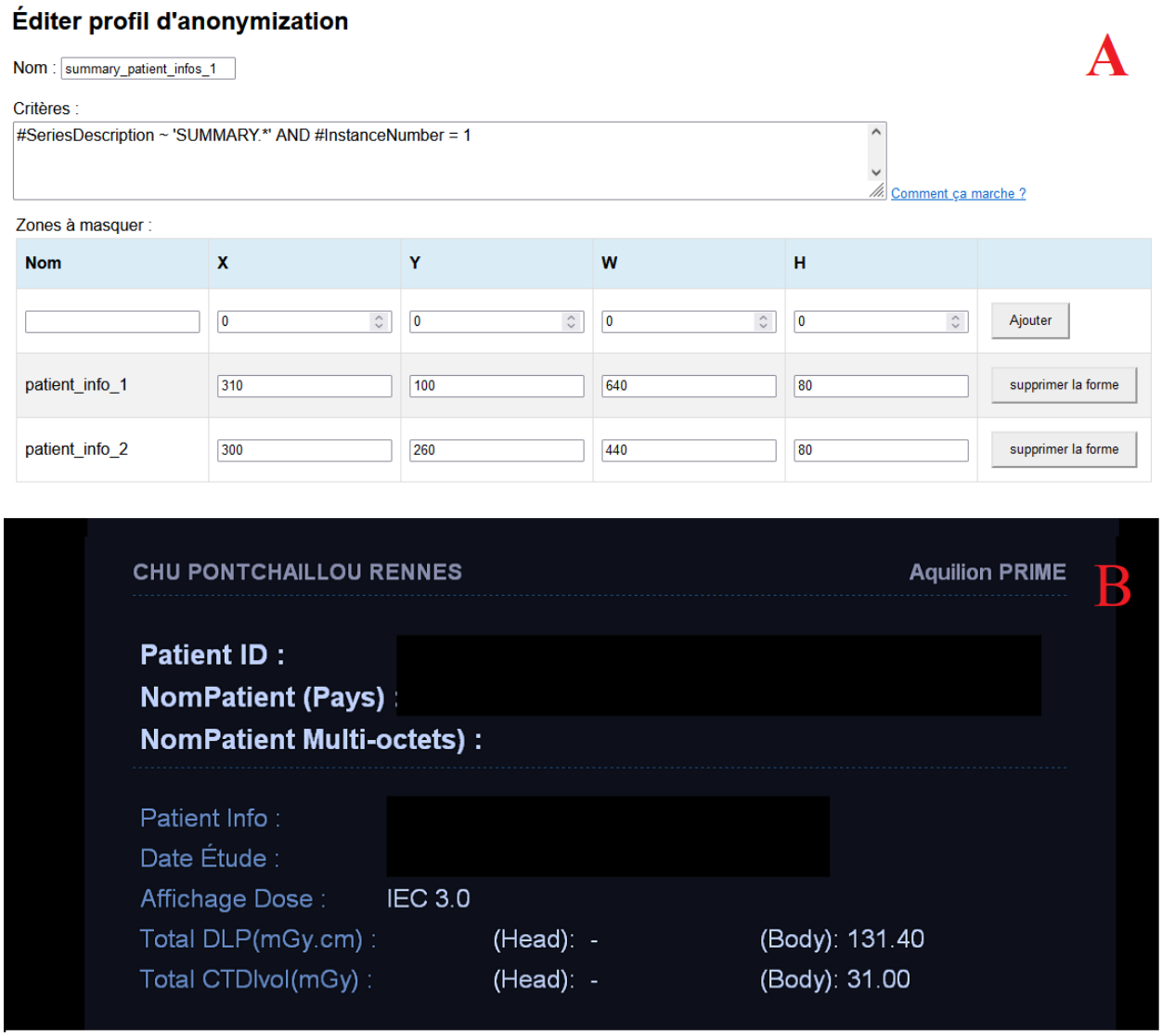



*Figure 28 : Un utilisateur du module peut définir des profils d'anonymisation en appliquant des filtres sur les valeurs d'attributs DICOM (Image A). Ici les fichiers DICOM dont la valeur de l'attribut "Series Description" concorde avec l'expression régulière 'SUMMARY.*' et dont la valeur de l'attribut "Instance Number" vaut 1 sont obfusquées aux endroits prédéfinis dans le profil, ici deux zones sont concernées (Image B).*

### 2.3.5. Performances

Pour évaluer la viabilité de notre système, nous avons mesuré les durées des requêtes effectuées par le module sur le PACS pour retrouver des métadonnées (un fichier par série) ou des acquisitions entières (tous les fichiers DICOM constituant une série). Avec moins de 8 secondes en moyenne pour récupérer une série DICOM et un peu plus de 2 secondes pour

les métadonnées uniquement, le système est adapté pour la visualisation ponctuelle d'examens ainsi que pour la récupération en routine des métadonnées.

Enfin, nous avons évalué la capacité de notre système à aider les utilisateurs à réaliser des études sur les données d'imagerie. Pour cela, dans le cadre d'une étude sur le cancer de la prostate, nous avons comparé les ensembles d'examens retrouvés par un clinicien en utilisant directement les outils de l'hôpital ou en utilisant eHOP avec le module imagerie. Tous les examens trouvés par le clinicien ont pu être retrouvés via eHOP ce qui en fait une solution fiable pour la mise en œuvre d'étude d'imagerie.

### 2.3.6. Récupération des données d'historique

Si la récupération des données d'imagerie en routine est viable, comme vu précédemment, le chargement des données d'historique dans l'entrepôt présente un nouveau verrou. En effet, le PACS utilisé dans notre environnement ne maintient plus l'indexation des examens au niveau "IMAGE" quand ces derniers ont plus de 2 mois. Il en résulte que notre système doit requêter l'ensemble des fichiers d'une série pour pouvoir récupérer ses métadonnées ce qui augmente considérablement la quantité de données échangée et les temps de traitements. Par ailleurs, ces données archivées ne sont pas non plus simplement accessibles sous la forme de fichiers DICOM car elles sont chiffrées par le logiciel de PACS.

Des discussions sont en cours avec l'éditeur de cette solution afin que nous puissions déchiffrer les archives et charger directement les données archivées, plus rapidement et sans solliciter le PACS.

## 2.5. Cas d'usage

Afin de valider le fonctionnement général de la récupération des métadonnées et la capacité d'eHOP à constituer des cohortes de patient sur des données d'imagerie, nous nous somme basé sur un cas d'usage portant sur le cancer de la prostate. Les responsables de l'étude avaient préalablement créé leur cohorte en utilisant directement les applications de gestion de données de l'hôpital, l'objectif était ici de comparer le contenu de leur cohorte avec une cohorte conçue via eHOP. La cohorte créée avec eHOP contenait plus de patients que la cohorte d'origine et l'incluait entièrement, le système eHOP est donc plus efficace que l'utilisation directe des applications de soins. Sur la cohorte ainsi créée, les responsables de l'étude s'intéressaient aux IRM utilisant un champ magnétique de 3 Teslas. Alors que cette information avait dû être retrouvée au cas par cas avec leurs données, une requête eHOP sur la métadonnée adéquate (l'attribut DICOM *Magnetic Field strength* (0018,0087)) a permis de filtrer immédiatement les examens ciblés.

Le module a déjà pu être utilisé en situation réelle dans le cadre d'une étude en rhumatologie. L'objectif de l'étude était de savoir si un biomarqueur d'imagerie (la densité Hounsfield au niveau de la vertèbre L1, extraire d'un scanner) permet d'estimer la densité minérale osseuse chez des patients réalisant une immunothérapie. Les patients ont été identifiés via eHOP et leurs données d'imagerie pertinentes pour le cas d'usage ont pu être extraites du PACS de façon anonymisée, il s'agissait des scanners thoraco-abdomino-pelviens (permettant de voir la colonne vertébrale). Après avoir reçu les données, les responsables de l'étude on extrait des features, la densité Hounsfield sur la vertèbre L1, manuellement en utilisant le logiciel LifEx. D'autres données clinico-biologiques extraites d'eHOP via ses outils et l'intervention d'un data scientist expert ont été fournies aux auteurs de l'étude qui ont eu alors à leur disposition tout le nécessaire pour leurs travaux.

Ce cas d'usage a pu prouver la capacité du système à mobiliser toutes les données pertinentes disponibles au sein d'eHOP et à les exporter de façon anonyme pour les mettre à disposition des chercheurs.

## 2.6. Discussion et Perspectives de l'imagerie avec eHOP

Nous avons mis en place l'intégration sémantique de l'imagerie dans eHOP en concevant une solution d'interface avec le PACS utilisant des technologies modernes et standardisées (i.e. le protocole http) et en intervenant sur le développement d'eHOP, au niveau du modèle de données et de son interface.

Le module imagerie lui-même reste indépendant d'eHOP et pourrait être utilisé par une autre solution pour faire l'interface avec le PACS. Le langage Java et le framework utilisé, Spring, permettent un déploiement très simple puisqu'il suffit d'avoir une base de données puis de configurer et lancer le logiciel. Son utilisation du protocole DICOM le rend compatible avec tous les PACS et le système de plugin permet d'ajouter de nouveaux traitements automatiques sur les métadonnées sans modifier le module lui-même. Ce module reste un prototype qui doit être amélioré. L'anonymisation pourrait par exemple reposer sur une approche OCR plus proche de l'état de l'art que l'approche actuelle par règles, même si cela ne dispensera pas d'une revue manuelle des examens exportés.

L'adaptation du modèle eHOP à l'organisation des documents d'imagerie permet de garder une vue globale d'un examen tout en représentant chaque série par un document eHOP. La terminologie DCMEHOP est efficace pour la recherche dans les rapports structurés mais contient beaucoup d'éléments et peut être difficile à visualiser et manipuler pour des recherches sur des attributs DICOM communs, elle peut être améliorée, par exemple en donnant la possibilité de masquer les termes les moins utiles en pratique dans l'interface.

Les évolutions faites sur l'interface d'eHOP permettent de naviguer entre les documents d'un même examen, d'afficher les données d'imagerie et de visualiser les rapports structurés. De nombreux points peuvent être améliorés. Dans l'affichage des résultats d'une requête sur eHOP, les documents sont groupés par séjour mais pas par examen. Ainsi toutes les séries de tous les examens d'un même séjour sont affichées ensemble, sans que l'on puisse distinguer quel document appartient à quel examen. Les vues peuvent aussi être améliorées pour être plus ergonomiques. Enfin, un système de vignette pourrait permettre de rendre l'interface plus lisible et permettrait potentiellement aux utilisateurs de savoir si une acquisition correspond à ce qu'ils cherchent sans avoir à ouvrir la vue du viewer DICOM.

Le système conçu reste un prototype mais peut déjà mener des cas d'usages et son développement, concernant les modifications portées à eHOP, est intégré avec le processus de développement de la solution eHOP. Ainsi les développements effectués sur eHOP dans le cadre de ma thèse sont intégrés dans le suivi de la roadmap et des évolutions d'eHOP, comme toute autre évolution du logiciel. Le système de plugin laisse de nombreuses possibilités d'amélioration, comme par exemple un plugin capable de détecter la séquence d'IRM utilisée à partir des métadonnées. Les cas d'usages à venir continueront de consolider notre approche et demanderont sans doute de nouvelles évolutions. Ainsi on pourrait par exemple intégrer à eHOP ou à son écosystème les outils d'extraction des features (telles que celles extraites dans le cas d'usage présenté précédemment) de sorte que les images n'aient pas à être exportées mais seulement ces features.

# Discussion

L'imagerie est devenue un outil majeur du soin et par conséquent un objet d'étude primordial pour la recherche en santé. Nos travaux ont eu pour objectif de proposer des solutions méthodologiques et techniques contribuant à la réutilisation des données d'imagerie de vie réelles pour la recherche dans la perspective de rendre ces données « FAIR ».

A cette fin, nous nous sommes attachés à prendre en compte l'ensemble de la chaine de traitement des données d'imagerie pour les rendre intégrables et réutilisable dans une solution d'entrepôt de données de santé.

Au terme de ces travaux, nous considérons que les 2 principaux verrous explorés concernent l'interopérabilité des données d'imagerie dans une perspective de réutilisation secondaire des données et l'urbanisation d'un EDS au regard du système d'information de radiologie. Dans cette partie, nous discutons des apports et limites de notre contribution et des perspectives au regard de ces deux verrous.

# 1. Interopérabilité des données d'imagerie pour leurs réutilisations secondaires

## 1.1. Apports

### 1.1.1. Alignement terminologique

Nos travaux ont permis d'identifier les obstacles et les prérequis à l'alignement sémantique des données d'imagerie vers une représentation adaptée à la recherche. Nous avons ainsi relevé des problèmes dans la construction de la terminologie d'interface (ici les titres des rapports d'imagerie, redondants et porteurs de peu d'information), dans le contenu limité de ses termes (acquisitions d'images de différentes zones anatomiques non citées dans le titre de l'examen), dans la construction des terminologies de référence (limites de l'approche pré-coordonnée, couverture limitée du domaine), dans les ressources terminologiques utilisées pour la traduction (absence de certains termes composés dans l'UMLS, comme les "troncs supra aortiques").

### 1.1.2. Classification ontologique des examens à partir des métadonnées

Nous avons par ailleurs développé une approche de classification automatique se basant uniquement sur les métadonnées d'imagerie. Cette méthode consiste à traduire les informations (essentiellement non structurées) depuis les métadonnées en utilisant l'UMLS comme pivot pour la traduction vers l'anglais et faire ainsi le lien avec les concepts de l'ontologie RadLex. Le raisonneur HermiT est utilisé pour classifier l'examen dans un des termes définis par le playbook RadLex au sein d'une ontologie RadLex préalablement enrichie de ces termes tirés du playbook. Ici plusieurs obstacles ont pu être identifiés comme l'utilisation du standard DICOM, dans lequel finalement peu d'attributs sont obligatoires et dont les constructeurs de matériel ne respectent pas toujours les recommandations. Ainsi même pour des attributs dont la valeur devrait être structurée, on retrouve des phrases conçues localement, peu parlantes, ou tronquées par des processus automatiques.

## 1.2. Limites

Notre étude de couverture de RadLex sur les titres de comptes-rendus est limitée aux données d'un seul établissement et limitée au mapping manuel des 200 codes les plus fréquents. Par conséquent, il est probable que d'autres obstacles à la représentation sémantique des données d'imagerie apparaissent lorsque l'on traite des données d'examens moins commun ou d'autres établissements.

Notre classifieur ontologique est améliorable sur plusieurs aspects. La traduction des termes Français utilisant l'UMLS pourrait être probablement améliorée par l'utilisation de méthode de NLP comme cela a pu être fait dans d'autres travaux [160]. La construction de l'ontologie RadLex est parfois un obstacle à la représentation de certaines informations, par exemple l'absence d'un objet ou d'une action (absence d'injection de produit de contraste par exemple) n'est pas représentée ce qui limite les capacités de notre classifieur.

## 1.3. Perspectives

Nos travaux sur le mapping terminologique de notre terminologie locale sur les playbook a permis d'avoir une vision globale de ce processus dans l'objectif d'implémenter une solution de mapping automatique.

Dans le contexte de l'entrepôt de données de santé, où les rapports des examens d'imagerie sont accessibles avec les métadonnées de l'examen, notre méthode de classification ontologique pourrait être améliorée en étant alimentée par des concepts extraits à la fois des métadonnées mais aussi et des rapports d'examen. Dans [161] par exemple, les auteurs ont pu effectuer l'extraction de concepts RadLex depuis les rapports de radiologie via des méthodes de reconnaissance d'entités nommées.

# 2. Urbanisation d'un EDS au regard d'un système d'information de radiologie

## 2.1. Apports

Les formats de stockage et d'échange de données de santé évoluent pour s'adapter à des représentations des données servant des usages nouveaux et à des protocoles d'échanges qui se normalisent, au-delà du domaine de la santé (formats XML ou JSON ou le protocole http pour lequel on peut même parler d'interopérabilité structurelle [162]). Le standard DICOM et son protocole de communication n'étant pas adaptés à la manipulation des données pour la recherche [101,155], les solutions d'utilisation secondaire des données d'imageries doivent faire l'interface avec les outils DICOM et extraire les données (au moins les métadonnées) pour les manipuler dans un environnement adapté.

Dans ce domaine, la contribution de ce travail a été la conception d'un prototype pour l'intégration des données d'imagerie dans l'entrepôt de données eHOP que nous avons pu utiliser en situation réelle. Ce prototype sera amené à évoluer pour répondre toujours mieux aux besoins des chercheurs mais permet de présenter concrètement aux utilisateurs les prérequis d'un tel système et des opportunités qu'il offre. Les retours positifs et les nouvelles demandent témoignent ainsi de son potentiel. Enfin, son intégration simple avec

le système ETL (par l'utilisation du protocole http) et le logiciel eHOP est un atout pour sa pérennité et son déploiement plus large.

Le choix de ne pas répliquer les PACS et de conserver un identifiant de référence permet un passage à l'échelle simple. Ainsi même dans les cas où le contenu du PACS serait plus volumineux, le chargement en routine des données et l'utilisation du module par eHOP ne serait pas fortement impacté.

## 2.2. Limites

Notre prototype nécessite des améliorations pour être plus utile aux utilisateur d'eHOP. La terminologie DCMEHOP, si elle est très complète, peut être difficile à manipuler dans l'interface pour les examens d'imagerie classiques car toutes les métadonnées y sont représentées au même niveau (comme dans les fichiers d'origine). Un filtre dans l'interface sur les métadonnées les plus utiles améliorerait l'expérience utilisateur. De même, la représentation d'un examen complet (rassemblant le rapport, les rapports structurés et les séries d'acquisition) est possible lorsque l'on consulte un document mais pas encore dans la vue des résultats d'une requête ce qui peut compliquer la recherche d'un document.

## 2.3. Perspective

Des discussions sont en cours avec le constructeur du PACS pour pouvoir accéder directement aux fichiers DICOM archivés et notre module est déjà prêt pour charger ces données quand il y aura accès.

Notre approche modulaire, basée sur un système de plugin, permet d'envisager de nouvelles applications simplement, en permettant d'extraire des données agrégées spécifiques depuis les métadonnées. Ce système rend de plus I4DW indépendant d'eHOP et permet de l'utiliser dans d'autres contextes où il est nécessaire d'avoir un lien avec un PACS.

# Conclusion

La qualité des données est un facteur essentiel pour la recherche en santé, leur représentation sémantique en particulier est importante pour la mise en œuvre de méthodes d'intelligence artificielle. Les sujets abordés dans ce travail sont l'alignement des données d'imagerie de santé sur des référentiels permettant le partage et les études multicentriques et les méthodes de représentation sémantique de ces données, permettant par exemple de les lier avec des données d'autres domaines [163] ou d'effectuer dessus des raisonnements logiques automatisés. Parallèlement à cela, nous avons développé, évalué et mis en œuvre un système d'intégration des données d'imagerie dans un entrepôt de données de santé.

Comme on a pu le voir, l'imagerie est un domaine très étendu, dont les outils, méthodes et applications sont nombreux, variés et en constante évolution. De nombreux systèmes de représentations de ces données existent [164], sont de formes différentes (vocabulaire, ontologie, terminologies) et couvrent des aspects différents, parfois très spécifiques du domaine et peuvent se recouvrir mutuellement. Des initiatives sont apparues pour les réunifier largement ou au moins les rendre compatibles entre elles, comme l'UMLS ou la fonderie OBO. Nous avons pu constater à travers nos travaux que les axes d'améliorations pour la représentation des données sont nombreux [4] : Localement, par un contrôle plus strict de la construction de la terminologie d'interface et l'utilisation plus poussée des métadonnées et plus largement par l'amélioration des terminologies de référence en utilisant par exemple des systèmes de post coordination [165,166] et en capitalisant sur des ontologies fondatrices. L'ontologie RadLex, que nous avons largement utilisée dans ce travail, car elle correspondait le mieux à nos besoins, présente des failles dans sa conception.

A travers le monde, des projets de grandes envergures nécessitant la gestion de données de santé voient le jour. L'espace européen des données de santé [167], lancé le 3 Mai 2022, est un écosystème spécifique à la santé, composé d'un cadre de gouvernance, d'infrastructures, de règles et de standards et de pratiques communes. Cet espace concerne toutes les données des patients, dont l'imagerie. Les initiatives de standardisation de la réutilisation des données se déploient de plus en plus largement [168] comme le modèle de donnée standard OMOP [169] ou la plateforme d'entrepôt et d'analyse de données i2b2, et s'adaptent aux données d'imagerie [155,170].

Au niveau national, des projets nécessitant un partage à large échelle vont devoir faire des choix d'implémentation et ainsi favoriser l'apparition de standards de représentation des données. Le HdH par exemple, ou les projets de niveau régional comme le Ouest Data hub vont faire avancer la standardisation des données en France, du côté de leur utilisation

secondaire dans un premier temps. L'adhésion récente de la France à SNOMED International [171] est un exemple de la volonté nationale de standardisation qui simplifiera le partage et les échanges de données au niveau national et au-delà. Le secteur est encore bouillonnant et les choix des grands industriels ou des projets à large échelle (nationaux, européens) obligeront sans doute à un tri et à la promotion des solutions plébiscitées au rang de standard *de facto*.

# Références


1. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *J Big Data*. 2019;6(1):54. doi:10.1186/s40537-019-0217-0

2. Cirillo D, Valencia A. Big data analytics for personalized medicine. *Curr Opin Biotechnol*. 2019;58:161-167. doi:10.1016/j.copbio.2019.03.004

3. Rosenkrantz AB, Mendiratta-Lala M, Bartholmai BJ, et al. Clinical Utility of Quantitative Imaging. *Acad Radiol*. 2015;22(1):33-49. doi:10.1016/j.acra.2014.08.011

4. Leung KYE, van der Lijn F, Vrooman HA, Sturkenboom MCJM, Niessen WJ. IT Infrastructure to Support the Secondary Use of Routinely Acquired Clinical Imaging Data for Research. *Neuroinformatics*. 2015;13(1):65-81. doi:10.1007/s12021-014-9240-7

5. Ko CC, Yeh LR, Kuo YT, Chen JH. Imaging biomarkers for evaluating tumor response: RECIST and beyond. *Biomark Res*. 2021;9(1):52. doi:10.1186/s40364-021-00306-8

6. Safran C, Bloomrosen M, Hammond WE, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *J Am Med Inform Assoc JAMIA*. 2007;14(1):1-9. doi:10.1197/jamia.M2273

7. Schlegel DR, Ficheur G. Secondary Use of Patient Data: Review of the Literature Published in 2016. *Yearb Med Inform*. 2017;26(1):68-71. doi:10.15265/IY-2017-032

8. Sarwal D, Gupta V. Personal Health Record. In: *StatPearls*. StatPearls Publishing; 2022. Accessed July 5, 2022. http://www.ncbi.nlm.nih.gov/books/NBK557757/

9. Bouzillé G. *Enjeux et Place Des Data Sciences Dans Le Champ de La Réutilisation Secondaire Des Données Massives Cliniques : Une Approche Basée Sur Des Cas d'usage*. These de doctorat. Rennes 1; 2019. Accessed July 4, 2022. http://www.theses.fr/2019REN1B023

10. Sarkans U, Chiu W, Collinson L, et al. REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology. *Nat Methods*. 2021;18(12):1418-1422. doi:10.1038/s41592-021-01166-8

11. Deslandes M, Chave L, Pommier M, et al. État de l'art en imagerie médicale. *IRBM News*. 2019;40(2):45-61. doi:10.1016/j.irbmnw.2019.02.001

12. Diagnostic assisté par ordinateur. Accessed July 4, 2022. http://stringfixer.com/fr/Automated_medical_diagnosis

13. Dwivedi K, Sharkey M, Condliffe R, et al. Pulmonary Hypertension in Association with Lung Disease: Quantitative CT and Artificial Intelligence to the Rescue? State-of-the-Art Review. *Diagn Basel Switz*. 2021;11(4):679. doi:10.3390/diagnostics11040679

14. Shapiro LG, Stockman GC. *Computer Vision*. Prentice Hall; 2001.

15. What Is Medical Image Segmentation and How Does It Work? | Synopsys. Accessed July 5, 2022. https://www.synopsys.com/glossary/what-is-medical-image-segmentation.html

16. Taghanaki SA, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G. Deep Semantic Segmentation of Natural and Medical Images: A Review. Published online June 3, 2020. doi:10.48550/arXiv.1910.07655

17. Willemink MJ, Koszek WA, Hardell C, et al. Preparing Medical Imaging Data for Machine Learning. *Radiology*. 2020;295(1):4-15. doi:10.1148/radiol.2020192224

18. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang J, Wu Z, Ding X. Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation. Published online February 11, 2020. doi:10.48550/arXiv.1908.10454

19. Frontiers | Deep Network for the Automatic Segmentation and Quantification of Intracranial Hemorrhage on CT. Accessed July 4, 2022. https://www.frontiersin.org/articles/10.3389/fnins.2020.541817/full

20. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. *Invest Radiol*. 2017;52(7):434-440. doi:10.1097/RLI.0000000000000358

21. Hirsch L, Huang Y, Luo S, et al. Radiologist-Level Performance by Using Deep Learning for Segmentation of Breast Cancers on MRI Scans. *Radiol Artif Intell*. 2022;4(1):e200231. doi:10.1148/ryai.200231

22. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal*. 2017;35:303-312. doi:10.1016/j.media.2016.07.007

23. Pacilè S, Lopez J, Chone P, Bertinotti T, Grouin JM, Fillard P. Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool. *Radiol Artif Intell*. 2020;2(6):e190208. doi:10.1148/ryai.2020190208

24. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology*. 2019;290(2):305-314. doi:10.1148/radiol.2018181371

25. Sahran S, Qasem A, Omar K, et al. *Machine Learning Methods for Breast Cancer Diagnostic*. IntechOpen; 2018. doi:10.5772/intechopen.79446

26. Wu N, Phang J, Park J, et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Trans Med Imaging*. 2020;39(4):1184-1194. doi:10.1109/TMI.2019.2945514

27. Hardy M, Harvey H. Artificial intelligence in diagnostic imaging: impact on the radiography profession. *Br J Radiol*. 2020;93(1108):20190840. doi:10.1259/bjr.20190840

28. Li Y, Chen HJ. A Survey of Computer-aided Detection of Breast Cancer with Mammography. *J Health Med Inform*. 2016;7. doi:10.4172/2157-7420.1000238

29. Flores M, Glusman G, Brogaard K, Price ND, Hood L. P4 medicine: how systems medicine will transform the healthcare sector and society. *Pers Med*. 2013;10(6):565-576. doi:10.2217/PME.13.57

30. Crommelin DJA, Storm G, Luijten P. 'Personalised medicine' through 'personalised medicines': Time to integrate advanced, non-invasive imaging approaches and smart drug delivery systems. *Int J Pharm*. 2011;415(1):5-8. doi:10.1016/j.ijpharm.2011.02.010

31. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur J Cancer Oxf Engl 1990*. 2012;48(4):441-446. doi:10.1016/j.ejca.2011.11.036

32. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*. 2001;69(3):89-95. doi:10.1067/mcp.2001.113989

33. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer Oxf Engl 1990*. 2009;45(2):228-247. doi:10.1016/j.ejca.2008.10.026

34. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. 2000;100(1):57-70. doi:10.1016/s0092-8674(00)81683-9

35. Buckler AJ, Paik D, Ouellette M, Danagoulian J, Wernsing G, Suzek BE. A novel knowledge representation framework for the statistical validation of quantitative imaging biomarkers. *J Digit Imaging*. 2013;26(4):614-629. doi:10.1007/s10278-013-9598-3

36. Amdouni E. *Représentation Sémantique Des Biomarqueurs d'imagerie Dans Le Domaine Médical*. These de doctorat. Rennes 1; 2017. Accessed July 4, 2022. http://www.theses.fr/2017REN1S124

37. Colquitt RB, Colquhoun DA, Thiele RH. In silico modelling of physiologic systems. *Best Pract Res Clin Anaesthesiol*. 2011;25(4):499-510. doi:10.1016/j.bpa.2011.08.006

38. Kolla L, Gruber FK, Khalid O, Hill C, Parikh RB. The case for AI-driven cancer clinical trials – The efficacy arm in silico. *Biochim Biophys Acta BBA - Rev Cancer*. 2021;1876(1):188572. doi:10.1016/j.bbcan.2021.188572

39. PRIMAGE Project -. PRIMAGE Project. Accessed July 4, 2022. https://www.primageproject.eu/

40. Martí-Bonmatí L, Alberich-Bayarri Á, Ladenstein R, et al. PRIMAGE project: predictive in silico multiscale analytics to support childhood cancer personalised evaluation

empowered by imaging biomarkers. *Eur Radiol Exp*. 2020;4(1):22. doi:10.1186/s41747-020-00150-9

41. Badano A. In silico imaging clinical trials: cheaper, faster, better, safer, and more scalable. *Trials*. 2021;22(1):64. doi:10.1186/s13063-020-05002-w

42. Digital twin for real care: SIMBIOTX models the healthcare of the future | Inria. Accessed July 4, 2022. https://www.inria.fr/en/digital-twin-simbiotx-healthcare-future

43. Kamel Boulos MN, Zhang P. Digital Twins: From Personalised Medicine to Precision Public Health. *J Pers Med*. 2021;11(8):745. doi:10.3390/jpm11080745

44. Solutions for individual patients. Accessed July 4, 2022. https://www.siemens-healthineers.com/perspectives/mso-solutions-for-individual-patients.html

45. Update: Now featuring a podcast – Taking a look at Digital Twin Technology: a new frontier in personalised healthcare. AI Blog. Published January 20, 2020. Accessed July 4, 2022. https://ai.myesr.org/articles/taking-a-look-at-digital-twin-technology-a-new-frontier-in-personalised-healthcare/

46. Seymour K, Benyahia N, Hérent P, Malhaire C. Exploitation des données pour la recherche et l'intelligence artificielle : enjeux médicaux, éthiques, juridiques, techniques. *Imag Femme*. 2019;29(2):62-71. doi:10.1016/j.femme.2019.04.004

47. Gong E, Pauly JM, Wintermark M, Zaharchuk G. Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. *J Magn Reson Imaging JMRI*. 2018;48(2):330-340. doi:10.1002/jmri.25970

48. Krazinski AW, Meinel FG, Schoepf UJ, et al. Reduced radiation dose and improved image quality at cardiovascular CT angiography by automated attenuation-based tube voltage selection: intra-individual comparison. *Eur Radiol*. 2014;24(11):2677-2684. doi:10.1007/s00330-014-3312-9

49. Moorin R, Forsyth R, Fox R. Evaluating the impact of dose reduction software on Computed Tomography radiation dosimetry using radiology information systems meta data: An example of the use of novel linkable data. *Int J Popul Data Sci*. 1(1):65. doi:10.23889/ijpds.v1i1.65

50. GOLDBERG M, ZINS M, GUEGUEN A, et al. Apport des cohortes à la connaissance de la santé. *Actual Doss En Santé Publique*. 2012;(n° 78):13-52.

51. Cours. Accessed July 4, 2022. http://campus.cerimes.fr/maieutique/UE-sante-publique/epidemiologie/site/html/5.html

52. Fu L, Li Y, Cheng A, Pang P, Shu Z. A Novel Machine Learning-derived Radiomic Signature of the Whole Lung Differentiates Stable From Progressive COVID-19 Infection. *J Thorac Imaging*. 2020;35(6):361-368. doi:10.1097/RTI.0000000000000544

53. Les essais cliniques (Recherches interventionnelles portant sur un produit de santé) · Inserm, La science pour la santé. Inserm. Accessed July 4, 2022. https://www.inserm.fr/nos-recherches/recherche-clinique/essais-cliniques-recherches-interventionnelles-portant-sur-produit-sante/

54. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics Initiatives: Two National Projects to Promote Data Sharing in Healthcare. *Yearb Med Inform*. 2019;28(1):195-202. doi:10.1055/s-0039-1677917

55. Plateforme des données de santé | Direction de la recherche, des études, de l'évaluation et des statistiques. Accessed February 16, 2022. https://drees.solidarites-sante.gouv.fr/article/plateforme-des-donnees-de-sante

56. Les data challenges au service de la santé. Health Data Hub. Accessed July 5, 2022. https://www.health-data-hub.fr/nos-data-challenges-en-2022

57. France Life Imaging — Cat OPIDoR. Accessed July 5, 2022. https://cat.opidor.fr/index.php/France_Life_Imaging

58. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak*. 2012;12(1):8. doi:10.1186/1472-6947-12-8

59. Ouyang W, Zimmer C. The imaging tsunami: Computational opportunities and challenges. *Curr Opin Syst Biol*. 2017;4:105-113. doi:10.1016/j.coisb.2017.07.011

60. Kohli MD, Summers RM, Geis JR. Medical Image Data and Datasets in the Era of Machine Learning—Whitepaper from the 2016 C-MIMI Meeting Dataset Session. *J Digit Imaging*. 2017;30(4):392-399. doi:10.1007/s10278-017-9976-3

61. Aiello M, Cavaliere C, D'Albore A, Salvatore M. The Challenges of Diagnostic Imaging in the Era of Big Data. *J Clin Med*. 2019;8:316. doi:10.3390/jcm8030316

62. Safran C. Update on Data Reuse in Health Care. *Yearb Med Inform*. 2017;26(1):24-27. doi:10.15265/IY-2017-013

63. Medical image annotation: a challenge for improving care and research. Alcimed. Published December 21, 2021. Accessed July 4, 2022. https://www.alcimed.com/en/alcim-articles/medical-image-annotation-improving-care-research/

64. F2: Data are described with rich metadata. GO FAIR. Accessed July 4, 2022. https://www.go-fair.org/fair-principles/f2-data-described-rich-metadata/

65. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. *J Biomed Semant*. 2018;9:12. doi:10.1186/s13326-018-0179-8

66. RGPD : Traitement des données de santé. Data Legal Drive. Published February 7, 2021. Accessed February 16, 2022. https://datalegaldrive.com/donnees-sante-rgpd/

67. CHAPITRE I - Dispositions générales | CNIL. Accessed July 4, 2022. https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre1#Article4

68. L'anonymisation de données personnelles | CNIL. Accessed July 4, 2022. https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

69. Pseudonymization according to the GDPR [definitions and examples]. Data Privacy Manager. Published November 2, 2021. Accessed July 4, 2022. https://dataprivacymanager.net/pseudonymization-according-to-the-gdpr/

70. Casey A, Davidson E, Poon M, et al. A systematic review of natural language processing applied to radiology reports. *BMC Med Inform Decis Mak*. 2021;21(1):179. doi:10.1186/s12911-021-01533-7

71. Grouin C, Névéol A. De-identification of clinical notes in French: towards a protocol for reference corpus development. *J Biomed Inform*. 2014;50:151-161. doi:10.1016/j.jbi.2013.12.014

72. Monteiro E, Costa C, Oliveira JL. A De-Identification Pipeline for Ultrasound Medical Images in DICOM Format. *J Med Syst*. 2017;41(5):89. doi:10.1007/s10916-017-0736-1

73. Newhauser W, Jones T, Swerdloff S, et al. Anonymization of DICOM Electronic Medical Records for Radiation Therapy. *Comput Biol Med*. 2014;0:134-140. doi:10.1016/j.compbiomed.2014.07.010

74. Schwarz CG, Kremers WK, Wiste HJ, et al. Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives. *NeuroImage*. 2021;231:117845. doi:10.1016/j.neuroimage.2021.117845

75. Eke D, Aasebø IEJ, Akintoye S, et al. Pseudonymisation of neuroimages and data protection: Increasing access to data while retaining scientific utility. *Neuroimage Rep*. 2021;1(4):100053. doi:10.1016/j.ynirp.2021.100053

76. Connectome. In: *Wikipédia*. ; 2022. Accessed July 4, 2022. https://fr.wikipedia.org/w/index.php?title=Connectome&oldid=193679798

77. Ravindra V, Grama A. De-anonymization Attacks on Neuroimaging Datasets. Published online August 8, 2019. doi:10.48550/arXiv.1908.03260

78. Valizadeh SA, Liem F, Mérillat S, Hänggi J, Jäncke L. Identification of individual subjects on the basis of their brain anatomical features. *Sci Rep*. 2018;8(1):5611. doi:10.1038/s41598-018-23696-6

79. Goldberg M, Zins M. Le Health Data Hub (suite) - Pourquoi ? Comment ? *médecine/sciences*. 2021;37(3):271-276. doi:10.1051/medsci/2021016

80. Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol*. 2008;61(11):1085-1094. doi:10.1016/j.jclinepi.2008.04.008

81. Gauriau R, Bridge C, Chen L, et al. Using DICOM Metadata for Radiological Image Series Categorization: a Feasibility Study on Large Clinical Brain MRI Datasets. *J Digit Imaging*. 2020;33(3):747-762. doi:10.1007/s10278-019-00308-x

82. G□uld M, Kohnen M, Keysers D, Schubert H, Wein B, Lehmann T. Quality of DICOM header information for image categorization. *Proc SPIE - Int Soc Opt Eng*. 2002;4685. doi:10.1117/12.467017

83. The Open Microscopy Environment. Accessed September 27, 2022. https://www.openmicroscopy.org/about/

84. Alzheimer's disease ontology - Summary | NCBO BioPortal. Accessed September 27, 2022. https://bioportal.bioontology.org/ontologies/ADO

85. Quantitative Imaging Biomarkers Alliance. Accessed September 27, 2022. https://www.rsna.org/research/quantitative-imaging-biomarkers-alliance

86. Manzoor JD Werner Ceusters, Sager, Shahid. Werner Ceusters, MD Ontology Research Group - ppt download. Accessed July 5, 2022. https://slideplayer.com/slide/13274890/

87. GRIFFON N, SAVOYE-COLLET C, MASSARI P, DANIEL C, DARMONI SJ. An interface terminology for medical imaging ordering purposes. *AMIA Annu Symp Proc*. 2012;2012:1237-1243.

88. Schulz S, Rodrigues JM, Rector A, Chute CG. Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration. *Stud Health Technol Inform*. 2017;245:940-944.

89. Delbecque T, Jacquemart P, Zweigenbaum P. Utilisation du réseau sémantique de l'UMLS pour la définition de types d'entités nommées médicales. :15.

90. UMLS Metathesaurus Vocabulary Documentation. Accessed July 5, 2022. https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html

91. Rosenbloom ST, Brown SH, Froehling D, et al. Using SNOMED CT to Represent Two Interface Terminologies. *J Am Med Inform Assoc JAMIA*. 2009;16(1):81-88. doi:10.1197/jamia.M2694

92. SNOMED CT - Body structure - Classes | NCBO BioPortal. Accessed July 5, 2022. https://bioportal.bioontology.org/ontologies/SNOMEDCT/?p=classes&conceptid=root

93. Spackman KA, Cote RA. SNOMED RT: A Reference Terminology for Health Care. :5.

94. LOINC Document Ontology OWL File. LOINC. Accessed July 5, 2022. https://loinc.org/document-ontology/owl/

95. Cornet R, Abu-Hanna A. Usability of expressive description logics--a case study in UMLS. *Proc AMIA Symp*. Published online 2002:180-184.

96. Studer R, Benjamins VR, Fensel D. Knowledge engineering: Principles and methods. *Data Knowl Eng*. 1998;25(1):161-197. doi:10.1016/S0169-023X(97)00056-6

97. Gruber TR. A translation approach to portable ontology specifications. *Knowl Acquis*. 1993;5(2):199-220. doi:10.1006/knac.1993.1008

98. Borst WN, Borst WN. Construction of Engineering Ontologies for Knowledge Sharing and Reuse. Published online September 5, 1997. Accessed July 6, 2022. https://research.utwente.nl/en/publications/construction-of-engineering-ontologies-for-knowledge-sharing-and-

99. Bouaud J, Bachimont B, Charlet J, Zweigenbaum P. Methodological Principles for Structuring an "Ontology." Published online August 21, 1998.

100. Guarino N. Formal Ontology and Information Systems. :13.

101. Van Soest J, Lustberg T, Grittner D, et al. Towards a semantic PACS: Using Semantic Web technology to represent imaging data. *Stud Health Technol Inform*. 2014;205:166-170.

102. BioPortal FAQ - NCBO Wiki. Accessed July 4, 2022. https://www.bioontology.org/wiki/BioPortal_FAQ#What_is_the_OBO_Foundry.3F

103. Smith B, Ceusters W, Klagges B, et al. [No title found]. *Genome Biol*. 2005;6(5):R46. doi:10.1186/gb-2005-6-5-r46

104. Jackson R, Matentzoglu N, Overton JA, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database J Biol Databases Curation*. 2021;2021:baab069. doi:10.1093/database/baab069

105. Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 2007;25(11):1251. doi:10.1038/nbt1346

106. Furini F, Rai R, Smith B, Colombo G, Krovi V. Development of a Manufacturing Ontology for Functionally Graded Materials. In: American Society of Mechanical Engineers Digital Collection; 2016. doi:10.1115/DETC2016-59964

107. About Us | BioPortal. Accessed July 5, 2022. https://www.bioontology.org/about-us/

108. Wang KC. Standard Lexicons, Coding Systems and Ontologies for Interoperability and Semantic Computation in Imaging. *J Digit Imaging*. 2018;31(3):353-360. doi:10.1007/s10278-018-0069-8

109. LOINC. In: *Wikipédia*. ; 2022. Accessed July 4, 2022. https://fr.wikipedia.org/w/index.php?title=LOINC&oldid=191984151

110. Smith B, Arabandi S, Brochhausen M, et al. Biomedical imaging ontologies: A survey and proposal for future work. *J Pathol Inform*. 2015;6:37. doi:10.4103/2153-3539.159214

111. Burger A, Davidson D, Baldock R. *Anatomy Ontologies for Bioinformatics: Principles and Practice*. Vol 6.; 2008. doi:10.1007/978-1-84628-885-2

112. Wang KC, Patel JB, Vyas B, et al. Use of Radiology Procedure Codes in Health Care: The Need for Standardization and Structure. *RadioGraphics*. 2017;37(4):1099-1110. doi:10.1148/rg.2017160188

113. ACR NRDR Homepage. Accessed July 4, 2022. https://nrdr.acr.org/Portal/Nrdr/Main/page.aspx

114. Mabotuwana T, Lee MC, Cohen-Solal EV, Chang P. Mapping Institution-Specific Study Descriptions to RadLex Playbook Entries. *J Digit Imaging*. 2014;27(3):321-330. doi:10.1007/s10278-013-9663-y

115. Parisot C. The basic structure of DICOM. :39.

116. DICOM Attributes. Accessed July 4, 2022. https://www.l3harrisgeospatial.com/docs/dicomattributes.html

117. PS3.3. Accessed July 4, 2022. https://dicom.nema.org/medical/dicom/current/output/html/part03.html#table_A.4-1

118. Griffon N. *Modélisation, Création et Évaluation de Flux de Terminologies et de Terminologies d'interface : Application à La Production d'examens Complémentaires de Biologie et d'imagerie Médicale*. These de doctorat. Rouen; 2013. Accessed July 5, 2022. http://www.theses.fr/2013ROUES008

119. Beitia AO, Kuperman G, Delman BN, Shapiro JS. Assessing the performance of LOINC® and RadLex for coverage of CT scans across three sites in a health information exchange. *AMIA Annu Symp Proc AMIA Symp*. 2013;2013:94-102.

120. Sandhu RS, Shin J, Wang KC, Shih G. Single-Center Experience Implementing the LOINC-RSNA Radiology Playbook for Adult Abdomen/Pelvis CT and MR Procedures Using a Semi-Automated Method. *J Digit Imaging*. 2018;31(1):124-132. doi:10.1007/s10278-017-0016-0

121. Peng P, Beitia AO, Vreeman DJ, et al. Mapping of HIE CT terms to LOINC®: analysis of content-dependent coverage and coverage improvement through new term creation. *J Am Med Inform Assoc JAMIA*. 2019;26(1):19-27. doi:10.1093/jamia/ocy135

122. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. *J Am Med Inform Assoc JAMIA*. 1997;4(6):484-500. doi:10.1136/jamia.1997.0040484

123. Rodrigues JM, Robinson D, Della Mea V, et al. Semantic Alignment between ICD-11 and SNOMED CT. *Stud Health Technol Inform*. 2015;216:790-794.

124. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. *J Am Med Inform Assoc JAMIA*. 2006;13(3):277-288. doi:10.1197/jamia.M1957

125. Apache Jena - Home. Accessed September 30, 2022. https://jena.apache.org/

126. Glimm B, Horrocks I, Motik B, Stoilos G, Wang Z. HermiT: An OWL 2 Reasoner. *J Autom Reason*. 2014;53(3):245-269. doi:10.1007/s10817-014-9305-1

127. Mejino JLV, Rubin DL, Brinkley JF. FMA-RadLex: An application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology. *AMIA Annu Symp Proc AMIA Symp*. Published online November 6, 2008:465-469.

128. Chervenak AL, van Erp TGM, Kesselman C, et al. A System Architecture for Sharing De-Identified, Research-Ready Brain Scans and Health Information Across Clinical Imaging Centers. *Stud Health Technol Inform*. 2012;175:19-28.

129. Doran SJ, d'Arcy J, Collins DJ, et al. Informatics in radiology: development of a research PACS for analysis of functional imaging data in clinical research and clinical trials. *Radiogr Rev Publ Radiol Soc N Am Inc*. 2012;32(7):2135-2150. doi:10.1148/rg.327115138

130. Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics*. 2007;5(1):11-34. doi:10.1385/ni:5:1:11

131. Minati L, Ghielmetti F, Ciobanu V, et al. Bio-Image Warehouse System: Concept and Implementation of a Diagnosis-Based Data Warehouse for Advanced Imaging Modalities in Neuroradiology. *J Digit Imaging*. 2007;20(1):32-41. doi:10.1007/s10278-006-0859-2

132. Wong STC, Hoo KS, Knowlton RC, et al. Design and Applications of a Multimodality Image Data Warehouse Framework. *J Am Med Inform Assoc JAMIA*. 2002;9(3):239-254. doi:10.1197/jamia.M0988

133. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018. doi:10.1038/sdata.2016.18

134. Interopérabilité en informatique. In: *Wikipédia*. ; 2021. Accessed July 8, 2022. https://fr.wikipedia.org/w/index.php?title=Interop%C3%A9rabilit%C3%A9_en_informatique&oldid=184065807

135. Wg A. Semantic Interoperability. :14.

136. Achieving Semantic Interoperability Using RDF and OWL — v10. Accessed July 8, 2022. https://www.w3.org/2001/sw/BestPractices/OEP/SemInt/

137. Open Access and Fair Principles. Published November 28, 2018. Accessed July 4, 2022. https://ccafs.cgiar.org/open-access-and-fair-principles

138. Fair data. In: *Wikipédia*. ; 2021. Accessed July 4, 2022. https://fr.wikipedia.org/w/index.php?title=Fair_data&oldid=186437479

139. Principes FAIR | CCSD. Accessed July 4, 2022. https://www.ccsd.cnrs.fr/principes-fair/

140. Clunie DA. *DICOM Structured Reporting*. PixelMed Pub; 2000.

141. Noumeir R. DICOM structured report document type definition. *IEEE Trans Inf Technol Biomed Publ IEEE Eng Med Biol Soc*. 2003;7(4):318-328. doi:10.1109/titb.2003.821334

142. Bidgood WD, Horii SC, Prior FW, Van Syckle DE. Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging. *J Am Med Inform Assoc*. 1997;4(3):199-212.

143. Madec J, Bouzillé G, Riou C, et al. eHOP Clinical Data Warehouse From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. *Stud Health Technol Inform*. 2019;264:1536-1537. doi:10.3233/SHTI190522

144. Here is the first publication about the ArchiMed platform! Centre d'Investigation Clinique - Innovations Technologique du CHRU de Nancy. Published January 4, 2017. Accessed July 4, 2022. http://www.cic-it-nancy.fr/en/2017/01/04/premiere-publication-au-sujet-de-notre-plateforme-archimed-copy/

145. Frontiers | Shanoir: Applying the Software as a Service Distribution Model to Manage Brain Imaging Research Repositories. Accessed July 4, 2022. https://www.frontiersin.org/articles/10.3389/fict.2016.00025/full

146. Temal L, Lando P, Dojat M, Fürst F, Gibaud B, Kassel G. OntoNeuroLOG : une ontologie modulaire et multi-niveaux pour gérer l'hétérogénéité sémantique des métadonnées. :5.

147. and the CATI Consortium, Operto G, Chupin M, et al. CATI: A Large Distributed Infrastructure for the Neuroimaging of Cohorts. *Neuroinformatics*. 2016;14(3):253-264. doi:10.1007/s12021-016-9295-8

148. Langer SG. A Flexible Database Architecture for Mining DICOM Objects: the DICOM Data Warehouse. *J Digit Imaging*. 2012;25(2):206-212. doi:10.1007/s10278-011-9434-6

149. Langer SG. DICOM Data Warehouse: Part 2. *J Digit Imaging*. 2016;29(3):309-313. doi:10.1007/s10278-015-9830-4

150. Cohen S, Gilboa F, Shani U. PACS and electronic health records. In: *Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation*. Vol 4685. SPIE; 2002:288-298. doi:10.1117/12.467019

151. Rubin DL, Desser TS. A Data Warehouse for Integrating Radiologic and Pathologic Data. *J Am Coll Radiol*. 2008;5(3):210-217. doi:10.1016/j.jacr.2007.09.004

152. Seymour K, Payoux P. Radiomics Enabler R , an ETL (Extract-Transform-Load) for biomedical imaging in big-data projects. :6.

153. Medexprim | Real-World Imaging Evidence. Medexprim. Accessed July 4, 2022. https://www.medexprim.com/

154. Rajala T, Savio S, Penttinen J, et al. Development of a Research Dedicated Archival System (TARAS) in a University Hospital. *J Digit Imaging*. 2011;24(5):864-873. doi:10.1007/s10278-010-9350-1

155. Murphy SN, Herrick C, Wang Y, et al. High Throughput Tools to Access Images from Clinical Archives for Research. *J Digit Imaging*. 2015;28(2):194-204. doi:10.1007/s10278-014-9733-9

156. Kaspar M, Liman L, Ertl M, et al. Unlocking the PACS DICOM Domain for its Use in Clinical Research Data Warehouses. *J Digit Imaging*. 2020;33(4):1016-1025. doi:10.1007/s10278-020-00334-0

157. A Structured Reporting Templates (Normative). Accessed September 28, 2022. https://dicom.nema.org/medical/dicom/current/output/chtml/part16/chapter_A.html

158. E.2 Basic Application Level Confidentiality Profile. Accessed July 6, 2022. https://dicom.nema.org/medical/dicom/current/output/chtml/part15/sect_E.2.html

159. Aryanto KYE, Oudkerk M, van Ooijen PMA. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. *Eur Radiol*. 2015;25(12):3685-3695. doi:10.1007/s00330-015-3794-0

160. Deléger L, Merabti T, Lecrocq T, Joubert M, Zweigenbaum P, Darmoni S. A Twofold Strategy for Translating a Medical Terminology into French. *AMIA Annu Symp Proc*. 2010;2010:152-156.

161. Tsuji S, Wen A, Takahashi N, Zhang H, Ogasawara K, Jiang G. Developing a RadLex-Based Named Entity Recognition Tool for Mining Textual Radiology Reports: Development and Performance Evaluation Study. *J Med Internet Res*. 2021;23(10):e25378. doi:10.2196/25378

162. The role of interoperability – part one: Communication and understanding. Accessed July 6, 2022. https://www.cerner.com/ishmed/news/the-role-of-interoperability-communication-and-understanding

163. Chennubhotla C, Clarke LP, Fedorov A, et al. An Assessment of Imaging Informatics for Precision Medicine in Cancer. *Yearb Med Inform*. 2017;26(1):110-119. doi:10.15265/IY-2017-041

164. Sansone SA, Rocca-Serra P. Review: Interoperability standards. Published online October 24, 2016. doi:10.6084/m9.figshare.4055496.v1

165. Mabon K, Steinum O, Chute CG. Postcoordination of codes in ICD-11. *BMC Med Inform Decis Mak*. 2022;21(6):379. doi:10.1186/s12911-022-01876-9

166. Semantic interoperability: SNOMED CT, post-coordination and the model. Accessed July 6, 2022. https://wardle.org/terminology/2018/10/27/snomed-postcoordination-1.html

167. European Health Data Space. Accessed July 12, 2022. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

168. Observational Medical Outcomes Partnership (OMOP). Center for Healthcare Improvement and Medical Effectiveness — CHIME. Accessed July 12, 2022. https://chime.ucsf.edu/observational-medical-outcomes-partnership-omop

169. OMOP Common Data Model – OHDSI. Accessed July 12, 2022. https://www.ohdsi.org/data-standardization/the-common-data-model/

170. Park C, You SC, Jeon H, Jeong CW, Choi JW, Park RW. Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. *Yonsei Med J*. 2022;63(Suppl):S74-S83. doi:10.3349/ymj.2022.63.S74

171. Interopérabilité sémantique : la France choisit la SNOMED CT pour la description des localisations anatomiques. Accessed July 13, 2022. https://esante.gouv.fr/espace-presse/interoperabilite-semantique-la-france-choisit-la-snomed-ct-pour-la-description-des-localisations-anatomiques

# Annexes

# 1. Article: “Indexing imaging reports for data sharing: A study of mapping using RadLex Playbook and LOINC”

L’article suivant a été présenté lors de la conférence MIE 2022 à Nice. Ma contribution dans ce travail a été d’interroger les professionnels cliniciens et radiologues pour recueillir et analyser leurs besoins, l’analyse des données présentes dans eHOP pour identifier les éventuels problèmes de qualité qui pourraient nuire au mapping, la définition de la méthode de mapping et la réalisation du mapping manuel des 200 titres de rapports d’examens d’imagerie les plus fréquents, que j’ai ensuite vérifié avec l’aide d’un radiologue expert.

# Indexing imaging reports for data sharing : A study of mapping using RadLex Playbook and LOINC

Pierre LEMORDANT [a,b,1], Fleur MOUGIN [d], Sandie CABON [a], Yves GANDON [c], Guillaume BOUZILLE[a], Marc CUGGIA[a]

[a] *Univ Rennes, CHU Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France ;*
[b] *Enovacom, Marseille, France ;* [c] *CHU Pontchaillou,F-35000 Rennes, France ;*
[d] *Univ. Bordeaux, INSERM, BPH, U1219, Bordeaux, France ;*

**Abstract.** New use cases and the need for quality control and imaging data sharing in health studies require the capacity to align them to reference terminologies. We are interested in mapping the local terminology used at our center to describe imaging procedures to reference terminologies for imaging procedures (RadLex Playbook and LOINC/RSNA Radiology Playbook). We performed a manual mapping of the 200 most frequent imaging report titles at our center (i.e. 73.2% of all imaging exams). The mapping method was based only on information explicitly stated in the titles. The results showed 57.5% and 68.8% of exact mapping to the RadLex and LOINC/RSNA Radiology Playbooks, respectively. We identified the reasons for the mapping failure and analyzed the issues encountered.

**Keywords.** RadLex, LOINC, Mapping, Imaging, Clinical Data Warehouse

## 1. Introduction

The secondary use of imaging data, with AI methods, requires large scale data pooling and consequently interoperability in order to allow sharing data from different sources.

Healthcare data are defined using Interface Terminologies (ITs). In the biomedical field, an IT is commonly defined as "a systematic collection of healthcare related phrases (terms) that supports clinicians' entry of patient-related information into computer programs" [1]. However, the semantic interoperability in multi-center studies requires common, semantically defined terminologies. These Reference Terminologies (RTs) are defined as "terminologies designed to provide exact and complete representations of a given domain's knowledge, including its entities and ideas, and their interrelationships, and are typically optimized to support the storage, retrieval, and classification of clinical data" [2].

In a previous work, we proposed a pipeline to allow the integration, indexing and presentation of imaging data in our Clinical Data Warehouse (CDW) eHOP [3] via their

[1] Corresponding Author, Pierre Lemordant, Univ Rennes 1, Inserm, LTSI UMR 1099, Rennes, France; E-mail: pierre.lemordant@univ-rennes1fr.

metadata. These data come from the Picture Archiving and Communication System (PACS). Now, we want to map our imaging data to RTs to allow data sharing among different centers. We consider the RTs of RadLex and LOINC, whose coverage of ITs has been the subject of several studies [4-6].

In this work, we describe the details of mapping these local exam labels to RadLex and LOINC/RSNA terminologies and the barriers encountered.

## 2. Materials

As IT, we considered the labels describing imaging reports used at Rennes academic hospital over the last 18 years. For instance, the label "Scanner - Thorax sans IV" (Chest Computed tomography without intravenous injection of contrast agent) has been locally defined by the imaging department staff and is used in the radiology information system and in the electronic health records as a metadata of the imaging report for healthcare purposes. We queried these labels in our eHOP CDW which contains 1,467,000 imaging reports from more than 486,000 patients.

The Radiological Society of North America (RSNA) has created the RadLex Playbook [7], a standard system for naming radiological procedures that includes a list of 4,374 imaging procedure labels (terms) formed with elements of the RadLex ontology and identified by a RadLex Procedure ID (RPID). The RSNA and the Regenstrief Institute have been working together to create the LOINC/RSNA Radiology Playbook (L/R Playbook) [8] using a new information model to describe 6289 terms. This playbook uses LOINC ID as identifiers and provides correspondences between RadLex Playbook codes and LOINC codes. The RadLex Playbook has not been updated since 2018. The RSNA and Regenstrief Institute continue their collaboration to further develop the L/R playbook by adding new procedure codes identified only by a LOINC ID and not an RPID. In this work, we used version 2.71 of the L/R Playbook and version 2.5 of the RadLex Playbook.

## 3. Methods

First, we extracted the 200 most frequent imaging labels. This set represented our IT and covered 1,073,886 imaging exams (i.e 73.2% of all imaging exams in eHOP CDW). After removal of duplicates (e.g the same label but in upper case or with words separated by dashes instead of spaces), only 106 labels remained.

We then manually mapped our local labels to the RadLex and L/R Playbooks by considering all explicit information contained in the label. For example, although the expert (Y.G) knows that the "Ultrasound, intracranial vessels" exam is a Doppler exam, we did not map it to a RT term stating "Doppler" because this label did not specify the technique. Similarly, if the local term specified a reason for the exam, this reason must be mentioned in the RT term. Finally, the mapping was done with a single RT term, we did not combine several RT terms to describe a local term, unlike other approaches [5].

The mapping classification was rooted in previous classification proposals [9] and was as follows :

- **Exact match**: an RT code corresponded exactly to the procedure e.g. "Ultrasound - Abdomen-Kidney" perfectly matched "US ABD KIDNEY" (RPID1992);
- **Broader RT term issue**: the best RT candidate was broader in meaning than the local label. For instance, the L/R Playbook does not have a code containing all the elements from "CT - Chest Abdomen Pelvis Skull";
- **Narrower RT term issue**: some RT terms specify additional information that is not available in the local label, e.g. the RadLex Playbook always specifies information on the contrast agent in breast MRI and this did not allow finding an exact match for the local label "MRI - Breasts";
- **No exact match**: the local code used a concept that is not defined yet or never used in the RT. For instance "Hemosiderosis" which is never used in the two playbooks (but is defined in the RadLex ontology (RID5203));

## 4. Results

### *4.1.* Labels of the Interface Terminology

We observed much redundancy in our IT because the 200 most used labels represented 106 different imaging exams. This is explained by the fact that several teams of radiologists defined and modified this list of codes over the years. Among these labels, 80 specified only modalities and anatomic areas, 6 specified the contrast technique, 11 specified a procedure (e.g. "Densitometry"), 3 specified a reason (e.g. "Pulmonary embolism") and 2 specified a technique (e.g "Doppler").

By comparing these labels with the metadata that describe each acquisition made during the exam, we noted that the title was not always fully accurate. Indeed, sometimes, clinicians decide in the radiology room to make additional images than those scheduled, especially for X-rays exams in the context of trauma. For instance, procedures labeled as "X-ray - Wrist" contained a "wrist" acquisition but often also acquisitions targeting the "cervical spine", "elbow" or "clavicle".

### *4.2.* Mapping of the Interface Terminology to Reference Terminologies

Table 1 shows the results of our manual mapping to the RadLex and L/R Playbooks.

**Table 1.** Outcomes of the manual mapping of the 106 local labels.

| Mapping category | RadLex Playbook | LOINC/RSNA (L/R) Radiology Playbook |
|---|---|---|
| Exact match | 61(57.5%) | 73 (68.8 %) |
| Broader RT term issue | 18 (17.0 %) | 13 (12.3 %) |
| Narrower RT term issue | 17 (16.0 %) | 10 (9.4 %) |
| No exact match | 10 (9.4 %) | 10 (9.4 %) |

The "Broader RT term issues" outcome occurred when the closest terms in the RT did not include all words to match the local label. We observed that the level of specification of RT terms can vary according to the modality, among other things. For example, the code "MR Lower Extremity Joint" exists in L/R Playbook, but there is no exact equivalent for "Ultrasound - Lower Extremity Joint".

In most cases, the "Narrower RT terms issues" outcome was due to the mention of the contrast agent in the RT labels, while this information was not specified in the local label. However, this information is not provided homogeneously in the RTs e.g. in the L/R Playbook, the terms "CT Chest", "CT Abdomen" and "CT Chest and Abdomen and Pelvis" (resp. 24627-2, 41806-1, 87869-4) do not mention contrast agent, whereas terms describing CTs of "Chest and Abdomen" always specify the use of a contrast agent ("with", "without" or "without and with" resp. 42275-8, 42276-6, 42277-4). The expert review showed that in some cases, the imaging exam was done with and/or without a contrast agent in practice. However, as the local label did not explicitly specify this information, we could not perfectly match the label with an RT term, although the RT specifies the code that would allow the perfect match.

In the case of no match, the most common reason was the specification by the local code of the reason for the exam, or a procedure that was not mentioned in the two playbooks. For instance, "tuberculosis" or "cystography" are never used in the L/R Playbook (but exist in the RadLex ontology, RID34878 and RID29116). Another reason for the lack of match was the use of a new imaging technique that has not been added in the RT yet (e.g. the recently described EOS™ imaging system) [10]. Finally, in several cases, the French local label referred to an anatomic structure that was not referenced in the anatomical concepts of the RadLex ontology. For instance, the "Troncs supra aortiques" (supra-aortic trunk), which designates the brachio-cephalic artery, left common carotid artery and left subclavian artery has no exact equivalent in the RadLex ontology. Four of the ten "no match" cases identified (both playbooks) were explained by local labels referring to the "Troncs supra aortiques"

## 5. Discussion and conclusion

The aim of this work was to identify the issues encountered in mapping an IT to a RT in the medical imaging field, and focused on the coverage of the local terminology by the two RTs by taking into account the whole content of local terms. This work tried to identify areas for improvement in IT and RT for data reuse and is, to our knowledge, the first work of this type using local French terminology.

As a limit, our study was based on data from a single hospital and some of the identified issues may be specific to our institution. Moreover, we used a limited set of local terms and thus we might have missed problems linked to less common exams.

Other works about mapping of ITs to RTs in different domains mention the same problems of different granularity between ITs and RTs [4-6], or of terms that only have meaning within the local institution [6]. These observations show that ITs are primarily designed to be human-readable and highlight the importance of following good practice in creating ITs based on RTs [1,11]. This approach to creating ITs also permits the identification of terms that would be missing from the RT to efficiently evolve it [6].

Beitia *et al.* analyzed the CT procedure coverage by LOINC (before the unification of LOINC radiology procedures with RadLex) and RadLex terminologies [4] and obtained coverage rates of 70% and 75% respectively. The difference with our results can be explained by the stricter mapping method we used, the language barrier (e.g. "Supra aortic trunks"), the fact that we considered all modalities (by focusing only on CT, we obtained 75% and 66% of exact mapping to the L/R and RadLex playbooks

respectively) and possibly by the construction method of our IT (e.g. the mention of "reasons for exam", such as "endometriosis" or "hemosiderosis", in local labels led to a "no match" with our mapping method).

Our results show that the mapping coverage was higher with the L/R Playbook than with the RadLex Playbook. However, we could note that the unification process is not completely achieved (e.g. the RadLex term "XRAY BONE DENSITO" (RPID3335) does not yet have an equivalent in the L/R playbook).

In this study, we could identify the elements to be taken into account concerning IT, RT, and data workflow in the healthcare system to develop a classification system of imaging exams that will allow optimal data integration and sharing among centers.

## Acknowledgment

We would like to thank the French National Research Agency (ANR), for funding this work in the framework of the LabCom LITIS project (grant no. ANR-17-LCV1-0004).

## References

[1] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc. 2006 May-Jun;13(3):277-88. doi: 10.1197/jamia.M1957. PMID: 16501181; PMCID: PMC1513664.

[2] Rosenbloom ST, Brown SH, Froehling D, Bauer BA, Wahner-Roedler DL, Gregg WM, Elkin PL. Using SNOMED CT to Represent Two Interface Terminologies. Journal of the American Medical Informatics Association. 2009 Jan;16(1):81-8.

[3] Madec J, Bouzillé G, Riou C, Van Hille P, Merour C, Artigny ML, Delamarre D, Raimbert V, Lemordant P, Cuggia M. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. Stud Health Technol Inform. 2019 Aug 21;264:1536-1537. doi: 10.3233/SHTI190522. PMID: 31438219.

[4] Beitia AO, Kuperman G, Delman BN, Shapiro JS. Assessing the performance of LOINC® and RadLex for coverage of CT scans across three sites in a health information exchange. AMIA Annu Symp Proc. 2013 Nov 16;2013:94-102. PMID: 24551324; PMCID: PMC3900124.

[5] Sandhu RS, Shin J, Wang KC, Shih G. Single-Center Experience Implementing the LOINC-RSNA Radiology Playbook for Adult Abdomen/Pelvis CT and MR Procedures Using a Semi-Automated Method. J Digit Imaging. 2018 Feb;31(1):124-132. doi: 10.1007/s10278-017-0016-0. PMID: 28842816; PMCID: PMC5788822.

[6] Peng P, Beitia AO, Vreeman DJ, Loo GT, Delman BN, Thum F, Lowry T, Shapiro JS. Mapping of HIE CT terms to LOINC®: analysis of content-dependent coverage and coverage improvement through new term creation. J Am Med Inform Assoc. 2019 Jan 1;26(1):19-27. doi: 10.1093/jamia/ocy135. PMID: 30445562; PMCID: PMC7587153.

[7] Radiological Society of North America, *Radlex playbook*, http://playbook.radlex.org

[8] The LOINC®/RSNA Radiology playbook. https://loinc.org/collaboration/rsna. Accessed January 2022.

[9] Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc. 1997 Nov-Dec;4(6):484-500. doi: 10.1136/jamia.1997.0040484. PMID: 9391936; PMCID: PMC61267.

[10] Illés T, Somoskeöy S. The EOS™ imaging system and its uses in daily orthopaedic practice. Int Orthop. 2012 Jul;36(7):1325-31. doi: 10.1007/s00264-012-1512-y. PMID: 22371113; PMCID: PMC3385897.

[11] Griffon N, Savoye-Collet C, Massari P, Daniel C, Darmoni SJ. An interface terminology for medical imaging ordering purposes. AMIA Annu Symp Proc. 2012;2012:1237-43. PMID: 23304401; PMCID: PMC3540496.

# 2. Article: “Ontology-based classification of radiological procedures for consistent sharing in Clinical Data Warehouses”

L’article suivant a été présenté lors de la conférence ICBO 2020. L’article suivant a été présenté lors de la conférence ICBO 2020. Mes contributions dans ce travail ont été le développement des méthodes permettant (i) d’extraire un sous ensemble de l’ontologie RadLex pour obtenir une ontologie plus simple à manipuler tout en correspondant à nos besoins (ii) d’enrichir cette ontologie avec les synonymes français en mobilisant l’UMLS (iii) de fusionner le playbook RadLex dans l’ontologie en créant pour chaque terme du playbook une classe héritant de la classe “Procédure” et (iv) de parser un fichier DICOM et en extraire les données principales afin de créer dans l’ontologie un “individu” correspondant qui sera classé automatiquement par le raisonneur ontologique.

# Ontology-based classification of radiological procedures for consistent sharing in Clinical Data Warehouses

Pierre LEMORDANT[a,b,1], Bernard GIBAUD[a], Cyril GARDE[b], Sébastien DELARCHE[c], Didier GOUDET[d], Marc CUGGIA[a]

[a] *Univ Rennes, Inserm, LTSI UMR 1099, Rennes, France,*

[b] *Enovacom, Marseille, France,*

[c] *Centre Hospitalier Universitaire Pontchaillou, Rennes, France*

[d] *Centre Eugène Marquis, Rennes, France*

**Abstract.** Clinical data warehouses (CDW) allow the reuse of care data in a research context. Designing and operating CDWs require addressing interoperability, data enrichment and data modeling problems, among others. This work concerns the management of medical imaging data in CDWs. It proposes a data-driven approach for classifying radiological procedures using an ontology-based approach. This approach relies on the RadLex ontology and an imaging procedures terminology called RadLex Playbook, both developed by RSNA. We first created an ontology of the radiological procedures by merging the Playbook with the relevant extract of the RadLex ontology and enriched it with French terms using the UMLS meta thesaurus. Then, we developed a proof of concept of a radiological procedures data classifier that exploits the richness of RadLex ontology and the ontological reasoning and we assessed it using medical imaging data retrieved from two different facilities. Our results demonstrate feasibility and relevance of the approach. They also highlight differences in the methods of filling imaging procedure data in the two institutions, as well as some problems in the RadLex ontology. Based on this experience, this proof of concept will be refined to evolve towards a routinely usable classification tool supporting medical imaging data management in CDWs.

**Keywords.** ontology, RadLex, playbook, clinical data warehouse,

## 1. Introduction

Clinical data warehouses (CDW) allow reusing care data for research purposes, and some inter-institutional projects also propose to gather data from several CDWs at regional or national levels. CDW gather data from different poles of the hospital and allow researchers to conduct their studies by providing them with tools to select patient cohorts, perform analyses or import these data into their own analysis tools. The main challenges to be addressed are therefore the heterogeneity of health data and interoperability between health care institutions. The "Massive Health Data" team of the LTSI Laboratory develops and maintains eHOP, a CDW technology within the Rennes University Hospital [1]. In order to improve our CDW's versatility, we want to add to it

[1] Corresponding Author, Pierre Lemordant, Univ Rennes 1, Inserm, LTSI UMR 1099, Rennes, France; E-mail: pierre.lemordant@univ-rennes1.fr.

the ability to manage medical imaging information. Therefore, we plan to extract from the hospital's Picture Archiving and Communication Systems (PACS) the structured data that characterize imaging procedures and to use it in the hospital's datawarehouse, to index, classify and enrich this data. In the long run, we consider artificial intelligence or data mining study cases that will benefit from the alignment of our imaging data in a model describing in depth the entire field of radiology and accessible to algorithms. Ontologies provide a way to meet these needs and, in this work, we propose to categorize local imaging procedures using an ontology of radiology procedures.

The biomedical domain is a favorable ground for the development of ontologies. Many ontologies have been developed to describe different aspects of medicine : anatomy with FMA (Foundational Model of Anatomy), genomics with GO (Gene Ontology), radiation oncology with ROO (Radiation Oncology Ontology) as well as many others. The leading initiative to describe the field of radiology, which is of interest to us in this work, is RadLex [2][3]. The RadLex ontology has been produced by the Radiological Society of North America (RSNA) in order to provide radiologists with a common vocabulary. Started in 2005, RadLex now includes more than 30,000 terms and covers all fields of radiology like modalities (i.e. imaging techniques), anatomy, pharmacology and so on. In addition, the RSNA designed the RadLex Playbook [4], intended to be a terminology for radiology procedures. The current version of the Playbook is the result of a harmonization work between RadLex and the Logical Observation Identifiers Names & Codes (LOINC) terminology; it defines a list of radiological procedures, describing them through elements of the RadLex ontology. However, the procedures themselves are not integrated into RadLex.

Medical images and their related data are managed using the international DICOM standard [5]. This standard combines imaging data with relevant metadata such as patient demographics, examination information, device settings, etc. This is the format in which we receive the information about imaging procedures that we have to classify. Depending on the institution or modality manufacturer, DICOM metadata are more or less filled in, with information of variable quality. In our case our data come from two different institutes, which gives us the opportunity to make our solution more generic. DICOM defines a vocabulary for imaging metadata but does not define a standard terminology for procedures themselves.

In this work, we describe on the one hand how we developed an ontology of radiology procedures based on the RadLex Playbook and integrating the relevant classes of the RadLex ontology. On the other hand, we propose a proof of concept of the use of this procedures ontology by developing a tool to automatically classify imaging procedures from DICOM medical imaging metadata.

In this paper, we first describe the creation of our radiological procedures ontology and the development of our classification tool, explaining how we dealt with the challenges we faced. We then detail our results before discussing the limits of our design. Finally, we discuss the opportunities offered by this work, and particularly the ontological approach.

## 2. Material and methods

### *2.1. Creation of the ontology of radiological procedures*

The Playbook is a CSV file defining more than 4400 procedures identified by a unique RadLex Playbook ID (RPID). Procedures are described through some text fields (some created manually, others automatically) and by a set of variables whose values are references to RadLex ontology elements. Concretely, the last column named "RIDS" contains a list of RadLex identifiers separated by a vertical bar which allows to link fields from MODALITY to VIEW_4 with RadLex entities. Figure 1 shows a few columns of the Playbook, in this example the value "RID13060" in the set of values of column RIDS refers to the MODALITY column and represents the "magnetic resonance imaging" class of the RadLex ontology.

| RPID | SHORT_NAME | AUTOMATED_SHORT_NAME | MODALITY | BODY_REGION | MODALITY_MODIFIER | ANATOMIC_FOCUS | RIDS |
|---|---|---|---|---|---|---|---|
| RPID2599 | | XRAY LE 1-2VWS ANKLE BILAT | XR | LOWER EXTREMITY | 1 - 2 VIEWS | ANKLE | RID10345\|RID13060\|0\|RID2638\|0\|0\|0 |
| RPID2600 | XR Ankle 1-2V | XRAY LE 1-2VWS ANKLE | XR | LOWER EXTREMITY | 1 - 2 VIEWS | ANKLE | RID10345\|RID13060\|0\|RID2638\|0\|0\|0 |
| RPID2601 | | XRAY GUIDE MAJ JNT ASP | XR | | GUIDANCE | MAJOR JOINT | RID10345\|RID13060\|0\|0\|0\|0\|0\|0\|RID |

**Figure 1.** Sample of the RadLex Playbook.

First, to avoid working on the whole RadLex ontology, we extracted the subset of classes that are used in the Playbook. In order to maintain the hierarchical organization, we also extracted all the parents of each of these classes. We also extracted their descendants, since we believed that the latter could be useful during the process of matching DICOM metadata with the ontology classes.

To integrate the procedures into the ontology, we naturally chose to create a new class for each radiological procedure of the Playbook identified by a RPID. Text fields then became annotation properties while fields containing RPIDs became object properties. In order to conform to the RadLex ontology practice, we defined a Preferred_name annotation property. We assigned it with the value of the column SHORT_NAME (if not filled in, we used AUTOMATED_SHORT_NAME). All text columns from the Playbook were then converted to annotation properties with their original name (see Fig 2). For "RID columns" (from MODALITY to VIEW_4), filled in fields were converted to "owl:someValuesFrom" restriction on the "subClass of" definition of the new class (compare Fig 1 and 2). Some columns have several numbered versions. For example, 5 columns are defined for the body region, from BODY_REGION to BODY_REGION_4 (as seen on Fig 1) because an imaging procedure could cover several body regions. We have not considered a hierarchy within these columns and defined a single object property for these cases. Therefore, each field BODY_REGION_X being filled-in resulted in the definition of a new restriction has_BODY_REGION on the newly created class. The only specific item is the POPULATION field for which we also declared an "owl:allValuesFrom" restriction, as the patient can't be in several categories at once.

Finally, we defined an "owl:equivalentClass" clause gathering all these restrictions and adding the fact that the class is a "procedure" (already defined in RadLex). The MODALITY_MODIFIER column sometimes contains information on the number of views used for the procedure. We thought a data property would be more meaningful and we defined the "nb_views" data property. So, the values of MODALITY_MODIFIER

describing the number of views were specifically converted to constraints using this data property. Figures 1 and 2 show how the value "1 – 2 views" for MODALITY_MODIFIER becomes constraints on nb_views in the "owl:equivalentClass" clause in the resulting ontology. All these newly created classes become subclasses of "playbook_procedure" defined as a subclass of procedure.


**Figure 2.** Example of Playbook procedure integrated into our ontology.

As our use cases concern French data, we had to adapt our ontology by translating textual properties as well as possible.

The Unified Medical Language System (UMLS) is a meta thesaurus incorporating many biomedical terminologies. The elements of the various references are aligned and often benefit from translation efforts in many languages. The RSNA made the effort to align many classes from the RadLex ontology on the UMLS through the annotation property UMLS_ID which made our task easier. We used the freely available file MRCONSO.RRF containing all elements of all source vocabularies in the UMLS. A small extract of this file is provided in Fig 3, showing elements coming from different sources and translated in various languages. We tried to link RadLex classes to UMLS codes using the UMLS_ID property or to get an exact matching between the Preferred_name property and the UMLS label. For every matching code, we extracted all French labels described in the UMLS file and added them to the RadLex class using the Synonym_french annotation property that we created.

```
C0037303|ENG|S|L0037303|VC|S0425684|N|A0489194|||2715-5950|CSP|PT|2715-5950|skull|0|N|256|
C0037303|ENG|S|L0037303|VC|S1018639|Y|A23972546||C12789||NCI_CDISC|PT|SDTM-LOC|SKULL|0|N|256|
C0037303|ENG|S|L2545181|PF|S3005686|N|A2665209|||UWDA|PT|71325|Set of bones of cranium|0|N|256|
C0037303|FIN|P|L5680220|PF|S6513281|Y|A13411734||M0019947|D012886|MSHFIN|MH|D012886|Kallo|3|N||
C0037303|FRE|P|L3253461|PF|S3780993|Y|A7447501||M0019947|D012886|MSHFRE|MH|D012886|Crâne|3|N||
C0037303|GER|P|L3287242|PF|S3814835|Y|A7531090||M0019947|D012886|MSHGER|MH|D012886|Schädel|3|N||
C0037303|GER|S|L1243730|PF|S1485676|Y|A27687821||M0019947|D012886|MSHGER|ET|D012886|Cranium|3|N||
```

**Figure 3.** Sample from the UMLS.

### *2.2. Ontology-based classification of radiological procedures*

The DICOM standard organizes medical imaging information using a 4-level hierarchy. The PATIENT level contains the demographic data of the patient. The STUDY level describes an examination, this is the level we map to the codes of the Playbook. For the same imaging examination (called DICOM Study in the DICOM jargon), the practitioner may use several modalities in many configurations, the studies are therefore usually composed of several SERIES, each describing the results of each imaging operation (e.g. acquisition, processing). Finally, these SERIES are composed of elements of the IMAGE level each describing one image. Figure 4 shows this hierarchical organization, the blue part designates a DICOM study, i.e. a radiological procedure. At each level, the data is described by a sequence of key/value elements (in which the value can itself be a sequence of such elements) where keys are DICOM attributes.

In each healthcare facility, image data is managed differently and the DICOM data elements are filled with varying degrees of accuracy. The "Study Description" data element is the one most used by the operators and the whole radiology system. It is generally filled in with local codes, specific to the institution. These codes are known by the operators and allow an automatic pre-filling of the DICOM attributes. The RadLex RPIDs have been designed to be mapped with these local codes and thus allow radiology system from several institutes to communicate. The ACR Dose Index Registry project is an example of this use of the Playbook [6]. The process of mapping local codes on the playbook is a manual task and some initiatives aim to assist this mapping in a semi-automatic approach [7].

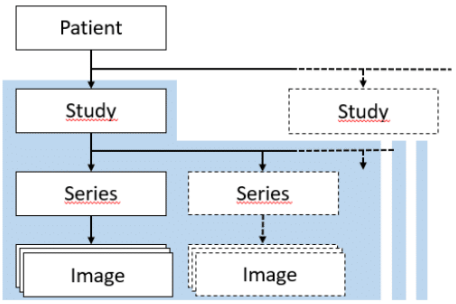



**Figure 4.** DICOM's hierarchical organization.

In this proof of concept, we worked from the raw data of each examination and tried to place it into the RadLex ontology. Our data-driven approach aims to enrich our data by getting the most out of each instance of imaging procedure.

We place ourselves at the moment of the extraction of DICOM data from the hospital PACS and before inserting data into our CDW. At this stage, each examination should result in an individual instantiating a class of radiological procedure from our ontology.

We first identified a list of DICOM attributes of interest. Among these, some -should- have normalized values, for instance "Image Laterality" has enumerated values: "R = right L = left U = unpaired B = both left and right". Therefore we designed our classifier to manage each case by associating the right class, e.g. when the value "B" is

encountered in attribute “Image Laterality”, we add a restriction on property has_LATERALITY aiming at RadLex class “bilateral” (RID5771). Using the “Study Date” and “Patient’s Birth Date”, we similarly set a has_POPULATION restriction depending on the patient’s age (for instance, targeting class “neonatal” if the patient is less than a month old).

The other data elements are filled in using free text data. By using defined attributes whose values are descriptive (“Study Description”, “Series Description” and so on), we extracted all words and groups of 2 or 3 words. From these samples, we removed those that appear to be in a negative context, checking if they were preceded by words like “without”, “non”, “no”, etc. We then enriched this list using a mapping of acronyms, abbreviations, and codes to keywords. This mapping is a json file that we populated manually, for instance associating “TAP” to keywords “thorax”, “abdomen” and “pelvis”. The next step was to find RadLex classes matching the elements of this enriched list of samples. To do so, we created an inverted dictionary that allowed to find RIDs from the values of the annotation properties “Preferred_name”, “Acronym”, “Synonym” and “Synonym_french”.

Once we had a set of RIDs, we checked which object properties were likely to point to them. For example, RID10312 “MRI” was only targeted by object property has_MODALITY. For each of such property, we added an object property assertion aiming at a created individual of the target class (IRI built by concatenating the Preferred_name and a generated UUID). Figure 5 shows how a procedure resulted in an individual with 9 object property assertions.

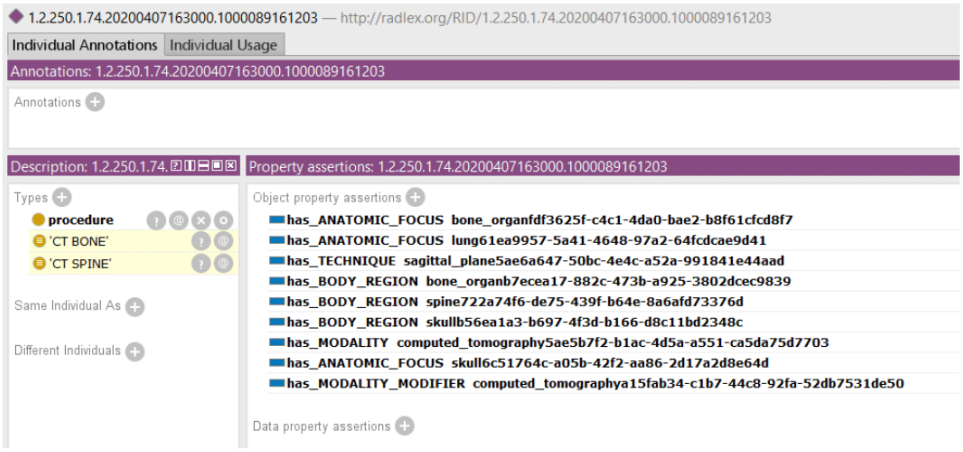



**Figure 5.** An individual in the procedures ontology, created from a real imaging examination.

Once the instances available, we can simply deduce the classes to which they belong using an OWL reasoner, to this end we used HermiT [8]. An example is provided in Fig 5, the individual has been considered as both a “CT BONE” procedure and a “CT SPINE” procedure, based on the assertions we produced with our extraction program. The development works presented in this part have been done with java and the Jena library.

## 3. Results

### *3.1. The procedures ontology*

The ontology of the procedures we produced contains 16506 classes, 4402 classes resulting from the integration of the Playbook, and 12104 classes extracted from the RadLex ontology. We defined eleven new object properties and one data property (nb_views) to describe these classes. Fig 6 shows the mapping between the columns of the Playbook and our new object properties.

| PlayBook Column | Playbook entries (RPID) using this column | Property assertion in "equivalent To" clause in the |
|---|---|---|
| MODALITY | 4402 | has_MODALITY |
| PLAYBOOK_TYPE | 4402 | N/A |
| POPULATION | 88 | has_POPULATION |
| BODY_REGION | 3735 | has_BODY_REGION |
| BODY_REGION_2 | 622 | has_BODY_REGION |
| BODY_REGION_3 | 216 | has_BODY_REGION |
| BODY_REGION_4 | 87 | has_BODY_REGION |
| BODY_REGION_5 | 45 | has_BODY_REGION |
| MODALITY_MODIFIER | 2178 | has_MODALITY_MODIFIER |
| MODALITY_MODIFIER_2 | 246 | has_MODALITY_MODIFIER |
| MODALITY_MODIFIER_3 | 8 | has_MODALITY_MODIFIER |
| PROCEDURE_MODIFIER | 333 | has_PROCEDURE_MODIFIER |
| PROCEDURE_MODIFIER_2 | 2 | has_PROCEDURE_MODIFIER |
| ANATOMIC_FOCUS | 2601 | has_ANATOMIC_FOCUS |
| ANATOMIC_FOCUS_2 | 202 | has_ANATOMIC_FOCUS |
| LATERALITY | 923 | has_LATERALITY |
| REASON_FOR_EXAM | 1104 | has_REASON_FOR_EXAM |
| REASON_FOR_EXAM_2 | 92 | has_REASON_FOR_EXAM |
| REASON_FOR_EXAM_3 | 0 | has_REASON_FOR_EXAM |
| TECHNIQUE | 246 | has_TECHNIQUE |
| PHARMACEUTICAL | 1874 | has_PHARMACEUTICAL |
| PHARMACEUTICAL_2 | 1 | has_PHARMACEUTICAL |
| VIEW | 142 | constraint on nb_views |
| VIEW_2 | 65 | constraint on nb_views |
| VIEW_3 | 24 | constraint on nb_views |
| VIEW_4 | 8 | constraint on nb_views |

**Figure 6.** Mapping of the Playbook columns onto the object properties of the ontology.

4623 French synonyms have been found and 1301 classes have been translated with at least one French synonym. However, we sometimes detected errors due to the matching that is done on acronyms. Thus, MRI was associated with "Mauritanie" (Mauritania) or “arriération mentale” (Mental Retardation). Our ontology is freely available here.

### *3.2. The proof of concept*

We tested our program with data from two different institutions, each time by picking up the characteristics of all the radiological procedures that had been produced in one day on modalities “US”, “CT” and “MR” (for ultrasound, computed tomography,

and magnetic resonance, respectively). Figure 7 shows how object properties were created on our radiological procedure individuals. For example, on the first institution, 4 individuals defined a has_REASON_FOR_EXAM property; these properties took 3 different values ("transplanted organ", "pregnancy" and "injection").

| | Institution 1 | | Institution 2 | |
|---|---|---|---|---|
| our Object Property | number of individuals using this Object Poperty | number of different Radlex classes targeted by this Object Property | number of individuals using this Object Poperty | number of different Radlex classes targeted by this Object Property |
| has_MODALITY | 122 | 4 | 75 | 5 |
| has_MODALITY_MODIFIER | 58 | 2 | 63 | 2 |
| has_BODY_REGION | 41 | 6 | 57 | 9 |
| has_ANATOMIC_FOCUS | 29 | 11 | 41 | 5 |
| has_REASON_FOR_EXAM | 4 | 3 | 8 | 5 |
| has_POPULATION | 4 | 1 | 2 | 1 |
| has_TECHNIQUE | 0 | 0 | 26 | 1 |

**Figure 7.** Detail of the created procedure individuals.

On the first one we retrieved 122 instances; on average these instances are described by more than 2 object properties. On the other we retrieved 75 instances, described by 5 object properties on average. In both cases the execution time of the request was less than 30 seconds and the creation time of the instances from the raw data was about five seconds. The reasoner HermiT takes 20 seconds to load and classify our individuals.

Figure 5 shows an example of classification where the individual was classified in "CT Bone" and "CT Spine". Of all the procedures received in both institutions, the program detected at least the modality. In the first institution, out of 122 individuals, 4 were classified as 2 different Playbook procedures, 92 were classified in 1 procedure and 26 were not classified. For those that were not classified, the only information collected was the procedure's modality worth "US" (ultrasound) but there is no "US" procedure that would describe a procedure that we simply know has an "ultrasound" modality (as is the case for MR). In the second institution, out of 75 individuals, 42 were classified in one class, 30 have two classes and 3 have 4 classes.

The last result of this work is our java program. To sum it up, it allows us to perform 4 tasks: 1) From an ontology and a list of class identifiers, extract the subset of these classes with their direct ascendants and descendants to form an ontology extract; 2) Merge the Playbook and the RadLex extract to form a procedures ontology; 3) From an ontology and the UMLS file, enrich the ontology with French synonyms (configurable for another language); 4) Run a DICOM query on a PACS, parsing the data using a keyword mapping file and producing an RDF file of individuals that a classifier can use to classify in our procedures ontology.

## 4. Discussion

### *4.1. The procedures ontology*

We believe that our radiological procedures ontology can be very useful and have demonstrated a use case for it. This work has been done with the idea of building on the extensive work done by many experts in the development of RadLex. However, while the advantages of RadLex are many, there are a few caveats. The description of anatomy

for example could be improved even if an effort was made to map some terms to the FMA [9]. More globally, it would be interesting to have an alignment of RadLex with formal ontologies such as BFO and other core ontologies from the Open Biological and Biomedical Ontologies Foundry (OBO) [10].

The richness of annotation properties “Preferred_name”, “Acronym”, “Synonym” and “Synonym_french” is important as we rely on it to match source data and RIDS, it however needs improving. For example, our program could not handle the word “épaule” (shoulder) because the Playbook does not use the RadLex class “shoulder” (RID39518) but “shoulder girdle”(RID1852) and the latter could not be translated.

Entries in the Playbook cover the different modalities differently. An “MR” entry makes it possible to assign a class to all MRI scans even if no other information is available, but such an entry does not exist for ultrasound. For CT, such an entry exists (RPID88) but its Preferred_name is “CT Guide needle place” which reveals that this procedure was intended to model CT Guidance for Needle Placement, although this motivation is not explicit in, e.g., the REASON_FOR_EXAM column of the Playbook.

The RadLex Playbook is not used in France, nor any other radiology procedures vocabulary, but inter-institutional grouping projects could gradually lead RadLex or other standard vocabulary to take hold.

As the Playbook is compiled manually and focuses more on classification than on description, the number of columns mapped remains limited (even if it is already very large). Also, it is sometimes difficult to have a clear definition of the role of a column, for example MODALITY_MODIFIER which points at classes of many different types. Similarly, in [2], REASON_FOR_EXAM is defined as: “Information about the clinical indication, patient diagnosis, clinical status (e.g., postoperative), an intended measurement, altered anatomy (e.g., endograft), or some other purpose of the study (eg, screening)”. There is obviously significant overlap between columns MODALITY_MODIFIER_X, REASON_FOR_EXAM_X and TECHNIQUE, possibly also resulting from the merging work with LOINC.

Our work being based on ontological reasoning for classification and on annotation properties to make the link between keywords and classes, it should rely on axioms definitions that are fully consistent, i.e. non-overlapping and where the use of subsumption is perfectly correct. An alignment with BFO, the use of methods based on OBO principles for the alignment with other ontologies and a restructuring work on the playbook could therefore be a way to improve our system and could be done as a continuation of this work.

A quick look at the results already showed the value of our approach. For example, an examination in which we found the term “mammo” was associated with the RID “mammography”. The latter is a sub-class of RadLex “radiography projection” (RID10345) which is a potential target class of the object property has_MODALITY, the newly created procedure individual has therefore been linked to a mammography individual via has_MODALITY. The reasoner classified the examination with the procedure code “X-RAY” (RPID2501) because it had a radiography projection as its modality.

By associating an institution's local codes with the RPIDs like it is done in the ACRDose project, one could situate the labeled exams within RadLex and thus take advantage of the power of this ontology. However, our approach is to take advantage of the power of the ontology for classification itself since we place the concrete objects in RadLex, which then allows classification. Thus, if RadLex changes, we simply need to regenerate our ontology.

*4.2. The proof of concept*

Further validation of this work will be required before it can be used in practice in the CDW, where its usefulness will be further evaluated. If the “ontologization” of the Playbook is rather successful, the use of the DICOM data raises challenges that are not completely solved by our solution and will still require work on the development of our algorithm and the parsing of the input data.

More information could be drawn from the DICOM metadata. DICOM attributes are not very filled in, it is mainly the description fields that enable classifying the procedure. Moreover, there is a difference of method between the two institutions: in one of them the DICOM data element “Body Part Examined” is filled in (with free text instead of DICOM enumerated values), in the other it is a simple copy of the free text data element “Study Description”, which is problematic because this data element is 16 characters long, so the value is often truncated. Beyond the management of acronyms, abbreviations and codes, we faced the problem of adjectives. The difference in classification between institutions depends mainly on the fact that the first one uses adjectives, for example "encéphalique" that does not match in our ontology. A possible solution would be to manage this case like the others by adding in our mapping entries such as "encéphalique" to the keyword "crâne" (skull), or to find an already existing resource providing this link.

One of the main problems is the management of classes that describe the absence of a thing or an action. To assert that an exam is “Without IV contrast”, the data should be analyzed in depth (perhaps down to the IMAGE level). If a local code defines this “lack of IV contrast”, we could associate it in our keyword mapping file to the label “imaging without IV contrast” so that the algorithm would correctly add an assertion on has_MODALITY_MODIFIER to RID28768 (“imaging without IV contrast”).

One of the advantages of our solution is that the refining axes are reduced to the constitution of configuration files. There is still mapping work on local codes, acronyms, abbreviations and adjectives, but this work is very generic and does not depend on a specific ontology or institution. In the long term, one could imagine replacing this mapping with a Natural Language Processing approach. It was also found that the use of DICOM could be optimized in general and is not the same everywhere. In practice most of the information is in the Study Description data element. Our approach would be sensitive to an improvement in the filling and quality of the source data.

By improving and validating the accuracy of our program we can envisage other uses. If our program is unable to classify a radiological procedure or if it classifies it in several very different classes, it may be that it is a new examination not covered by the Playbook yet, or that the DICOM file is filled out incorrectly. With improved accuracy, it could be used to help detecting new uses or errors.

## 5. Conclusion

Starting from the problem of classification and interoperability in our CDW, we have designed the first ontology of the RadLex Playbook in OWL. By merging the Playbook with the relevant extract of the RadLex ontology, we were able to capitalize on the great work already done by the actors of these projects. We have given an example of how we have enriched our ontology resource, translating it in another language via UMLS.

Finally, we created a concrete use case of our ontology of procedures that we were able to test with data from two different institutions.

There are many ways to improve our ontology, its translation and our proof of concept, but the data-driven approach is already giving interesting results. The use of such a system offers new opportunities within the framework of CDW given the importance of interoperability and the increasing place of imaging.

## Acknowlegement

We would like to thank the French National Research Agency (ANR), for funding this work inside the LabCom LITIS (Laboratoire d'Interopérabilité, de Traitement et d'Intégration des données de Santé) project (grant no. ANR-17-LCV1-0004).

## References

[1] Madec J, Bouzillé G, Riou C, et al. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. Stud Health Technol Inform. 2019;264:1536 - 1537. doi:10.3233/SHTI190522

[2] Wang KC, Patel JB, Vyas B, Toland M, Collins B, Vreeman DJ, Abhyankar S, Siegel EL, Rubin DL, Langlotz CP. Use of Radiology Procedure Codes in Health Care: The Need for Standardization and Structure. Radiographics. 2017 Jul-Aug;37(4):1099-1110. doi: 10.1148/rg.2017160188. PMID: 28696857; PMCID: PMC5548452.

[3] Wang KC. Standard Lexicons, Coding Systems and Ontologies for Interoperability and Semantic Computation in Imaging. J Digit Imaging. 2018 Jun;31(3):353-360. doi: 10.1007/s10278-018-0069-8. PMID: 29725962; PMCID: PMC5959830.

[4] Vreeman DJ, Abhyankar S, Wang KC, Carr C, Collins B, Rubin DL, Langlotz CP. The LOINC RSNA radiology playbook - a unified terminology for radiology procedures. J Am Med Inform Assoc. 2018 Jul 1;25(7):885-893. doi: 10.1093/jamia/ocy053. PMID: 29850823; PMCID: PMC6016707.

[5] NEMA PS3 / ISO 12052, Digital Imaging and Communications in Medicine (DICOM) Standard, National Electrical Manufacturers Association, Rosslyn, VA, USA (available free at http://medical.nema.org/)

[6] Morin RL, Coombs LP, Chatfield MB. ACR Dose Index Registry. J Am Coll Radiol. 2011;8(4):288 - 291. doi:10.1016/j.jacr.2010.12.022

[7] Mabotuwana T, Lee MC, Cohen-Solal EV, Chang P. Mapping institution-specific study descriptions to RadLex Playbook entries. J Digit Imaging. 2014 Jun;27(3):321-30. doi: 10.1007/s10278-013-9663-y. PMID: 24425187; PMCID: PMC4026460.

[8] Glimm B, Horrocks I, Motik B, Stoilos G., Wang Z. (2014). HermiT: An OWL 2 Reasoner. Journal of Automated Reasoning, 53, 245-269.

[9] Mejino JL, Rubin DL, Brinkley JF. FMA-RadLex: An application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology. AMIA Annu Symp Proc. 2008;2008:465 - 469. Published 2008 Nov 6.

[10] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251-5. doi: 10.1038/nbt1346. PMID: 17989687; PMCID: PMC28140

# 3. Article: “How to optimize connection between PACS and Clinical Data Warehouse: A web service approach based on full metadata integration”

L’article suivant a été présenté lors de la conférence MedInfo 2021. Ma contribution à ces travaux consistait tout d’abord à comparer les solutions logiciel permettant d’afficher du contenu DICOM dans une interface web (les “DICOM viewers”) pour sélectionner celle à intégrer à l’interface d’eHOP. J’ai ensuite développé le module imagerie (I4DW) ainsi que le plugin d’intégration eHOP qui crée un fichier “pivot” destiné à l’intégration dans le modèle eHOP à partir des fichiers DICOM récupérés depuis le PACS. J’ai mis en place les processus d'intégration dans notre solution d’ETL (Extract Transform and Load) ESV2. J’ai effectué les modifications nécessaires dans eHOP : nouvelles vues, fonctions de récupération et affichage des fichiers DICOM et fonction permettant de rassembler les documents d’un même examen. Enfin, j’ai conçu la terminologie DCMEHOP qui permet de préserver le contexte de chaque attribut DICOM indexé dans eHOP en donnant aux utilisateurs la possibilité de parcourir les métadonnées de la même façon qu’elles se présentent dans les fichiers DICOM d’origine, en particulier les rapports structurés (Structured Reports).

# How to optimize connexion between PACS and Clinical Data Warehouse : A web service approach based on full metadata integration

**Pierre Lemordant[a,b,1], Guillaume Bouzille[a], Bernard Gibaud[a], Cyril Garde[b], Romain Mathieu[ce], Boris Campillo-Gimenez[f], Didier Goudet[d], Sébastien Delarche[c], Yann Roland[f], Marc Cuggia[a]**

*[a] Univ Rennes, CHU Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France ; [b] Enovacom, Marseille, France ;*
*[c] CHU Pontchaillou,F-35000 Rennes, France ; [d] CLCC Eugène Marquis, F-35000 Rennes, France ;*
*[e] Univ Rennes, Inserm, EHESP, Irset – UMR_S 1085, F-35000 Rennes, France;*
*[f] Univ Rennes, CLCC Eugène Marquis, Inserm, LTSI - UMR 1099, F-35000 Rennes, France ;*

## Abstract

Clinical image data analysis is an active area of research. Integrating such data in a Clinical Data Warehouse (CDW) implies to unlock the PACS and RIS and to address interoperability and semantics issues. Based on specific functional and technical requirements, our goal was to propose a web service (I4DW) that allows users to query and access pixel data from a CDW by fully integrating and indexing imaging metadata. Here, we present the technical implementation of this workflow as well as the evaluation we carried out using a prostate cancer cohort use case. The query mechanism relies on a Dicom metadata hierarchy dynamically generated during the ETL Process. We evaluated the Dicom data transfer performance of I4DW, and found mean retrieval times of 5.94 seconds and 0.9 seconds to retrieve a complete DICOM series from the PACS and all metadata of a series. We could retrieve all patients and imaging tests of the prostate cancer cohort with a precision of 0.95 and a recall of 1. By leveraging the CMOVE method, our approach based on the Dicom protocol is scalable and domain-neutral. Future improvement will focus on performance optimization and de identification.

***Keywords:***

Medical Imaging, Clinical Data Warehouse, Health Information System Interoperability

## Introduction

Clinical images data analysis, especially with artificial intelligence methods, is an active area of research that holds promise for disease characterization, precision medicine, and early assessment of treatment response. However, a key challenge needs to be overcome: unlocking hospital imaging software components (i.e Picture Archiving and Communication Systems, PACS and Radiology Information Systems, RIS) to integrate imaging data with the other patient clinical data into Clinical Data Warehouses (CDW) or data lakes for research reuse [1,2]. Researchers have developed several solutions for reusing imaging data, such as Research PACS [3,4] that manages the whole process for imaging-oriented clinical trials. Other solutions exist to carry out imaging studies, but they are fully imaging oriented [5,6], or extract data from the PACS for a given predefined cohort and allow crossing imaging data with clinical data a posteriori [7]. Kaspar et al [8] described the implementation of a technical component to integrate imaging metadata from the clinical PACS into a CDW. They proved the feasibility of routine feeding via basic metadata queries (CFIND), or queries on the first image of a series (CMOVE) to retrieve all metadata for the identified patient subsets.

However, the metadata recovered via CFIND queries are very limited. To tackle this issue, we developed a prototype (Images for Data Warehouse, I4DW) that fully captures the semantics around the image and efficiently connects a PACS to a CDW. Here, we present the technical implementation of this workflow, and its evaluation using a prostate cancer cohort use case.

## Material and Methods

***Functional and technical requirements :*** To prioritize the R&D tasks, we interviewed a group of radiologists, clinicians and data scientists at our hospital to define the following list of functional and technical requirements :

- secondary reuse of data for the widest choice of purposes;
- combining queries on imaging data and other clinical data using the same interface;
- possibility to connect to the PACS and perform queries without jeopardizing the care processes; no duplication of the PACS data due to safety, GPRD, and IT resources reasons;
- possibility to browse and view clinical images on the CDW graphical user interface (GUI) to perform pre-screening and/or for data quality control;



128

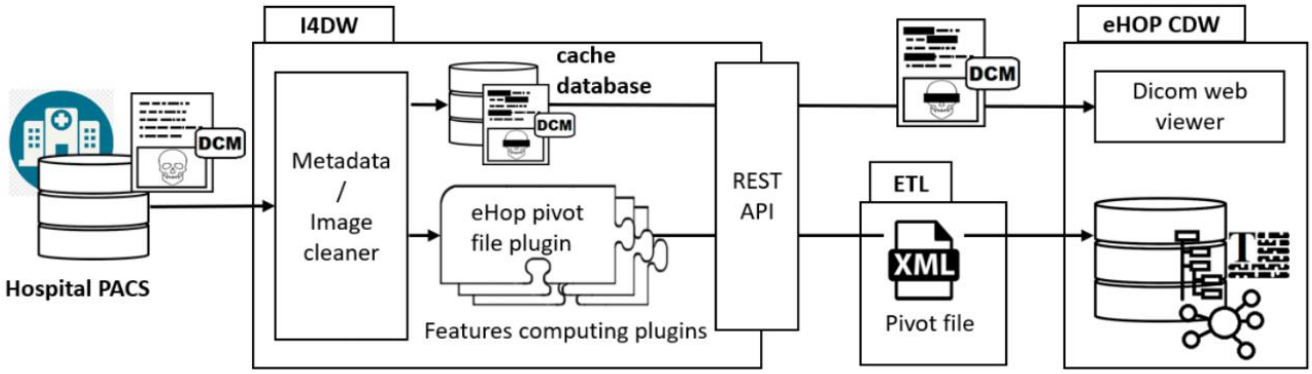



*Figure 1 - Global architecture : Flow of the imaging data through the I4DW server*

- leveraging the Digital Imaging and Communication in Medicine (DICOM) and terminology standards used in Dicom metadata (e.g. SNOMED, LOINC) to share data between CDWs and perform multi-center studies;
- possibility to enrich metadata during the Extraction, Transformation, Loading (ETL) process to improve image indexing and retrieval.

***System design :*** We designed a stand-alone JAVA server component called I4DW to connect the ETL component with the PACS. Figure 1 is an overview of the system architecture. I4DW relies on the DICOM protocol. This server, using the Spring Framework[1], provides a HTTP REST API (Application Programming Interface) on the CDW side and a Dicom standard interface on the other side, using the PixelMed Java library[2], to connect to one or more PACS.

To avoid unnecessary solicitations of the IT ressources, a DataBase cache is dedicated to store image collections (one or more cohorts) loaded from the PACS.

I4DW is based on a "data extraction/enrichment" plugin system that facilitates the addition of new functionalities without increasing the application burden and without any impact on the API. The ETL process uses a transitional "pivot" XML file. Each document to be integrated in the CDW goes through this format before its integration into the target data model. For more flexibility, we chose to create a dedicated data extraction plugin that produces this pivot format as a result. The main parameters of the REST API call performed by the ETL are the list of accession numbers to retrieve, the list of features needed (i.e the list of plugins to call) and a flag indicating the need to store the images in the cache.

Periodically, the ETL component calls the I4DW, passing as parameter a list of accession numbers obtained via the already integrated radiology reports data stream (coming from the RIS). This method allows retrieving the link between the targeted Dicom image data and the patient ID. The DICOM protocol allows to make queries on a limited set of fields indexed by the PACS (CFIND query) or queries requiring the retrieval of at least one image to obtain the whole header (CMOVE). As metadata extraction via CFIND is limited, we opted to routinely query imaging data with the CMOVE method.

***Integration in the CDW data model :*** eHOP is a CDW technology [9] developed by our team and currently used by 17 academic hospitals in France. eHOP data model is similar to that of the currently most recognized solutions (I2B2, OMOP). However, it introduces the notion of "document" entity that groups a set of atomic data elements in a specific context (e.g. a laboratory test report assembles all the measurements made, a drug prescription report lists all the drug prescriptions and administrations during a stay ). To integrate imaging data in this model, we considered the Dicom series as a document entity. During integration, each series from a Dicom study becomes a document. To keep track of the original study, all these documents and the radiology report document are linked by an accession number. This logical view of a Dicom study is computed at the application level and available in the dedicated view on the eHOP interface (Figure 3B).

eHOP allows querying all the patient's data (age, sex, etc), hospital stays (dates, medical unit, etc), and documents. Documents have a text field for text search and are linked to structured elements stored in a dedicated table. Structured elements are atomic values of different types (number, text, code, date) associated with terminologies. For instance, a document representing a surgery report can be associated with a structured element with the code 'JGFC001' (i.e prostatectomy) from the terminology of the French medical classification for clinical procedures (CCAM).

The Dicom standard describes images with attributes and can organize them hierarchically with "sequence attributes" grouping subsets of other attributes (e.g extensively used in Dicom structured reports). When a Dicom series is integrated into eHOP as a document, structured elements linked to this document are created

[1] The Spring framework https://spring.io/projects/spring-framework

[2] PixelMed™ Java Dicom Toolkit https://www.pixelmed.com/Dicomtoolkit.html



129

and rely on the “DCMEHOP” terminology, based on the Dicom attributes and their position in the hierarchy of sequences. This terminology is built dynamically in eHOP as Dicom data are progressively integrated, and thus users can search through a terminology organised like the data is organized in Dicom. The simple “Attribute Tag” top node of the terminology contains the attributes used in modalities other than Dicom SR. For each newly integrated type of SR modality, a new top node is created that contains the attributes organized according to the report template. The querybuilder allows setting constraints depending on the structured element type. Using the Value Representation (VR) of Dicom tags, each attribute type can be identified. eHOP also keeps track of all possible values for a given structured element, e.g when defining a constraint on modality the system suggests a set of possible values, “MR”, “CT”, etc. Figure 2 shows how this terminology is displayed in the query builder GUI.

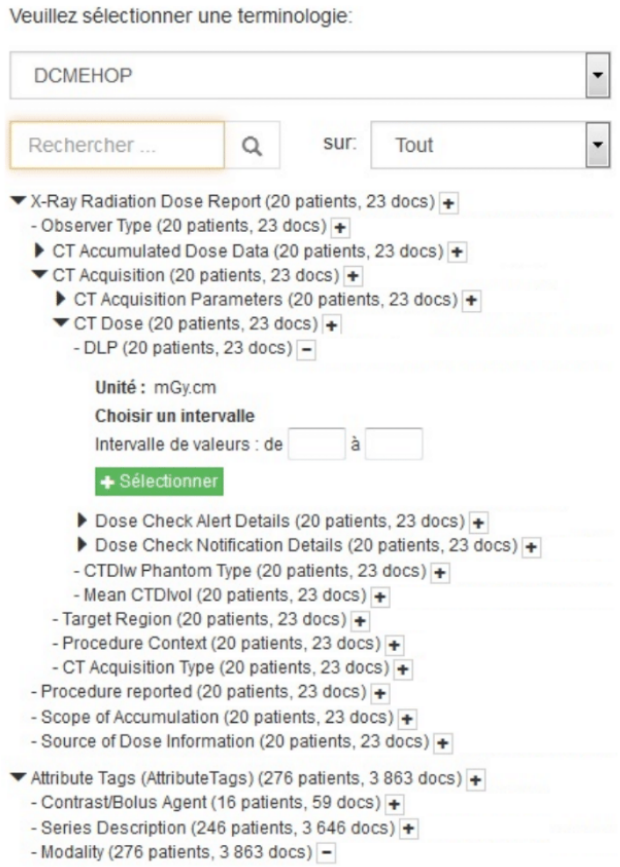



Figure 2 - Generated Dicom terminology displayed in the eHOP query builder

Moreover, some elements in the Dicom files are based on external terminologies such as LOINC and SNOMED. For instance, the attribute "Anatomic Region Sequence" could contain the attribute "coding scheme designator" with the value "SRT" (i.e SNOMED) and the attribute "code value" with the value "T-D3000" to designate the chest. As eHOP can do mapping between terminologies, when integrating these elements, the mapping must be maintained between the generated terminology and the SNOMED Clinical Terms or LOINC elements already in use in eHOP. eHOP documents created from the integration of imaging tests are also textually indexed based on a subset of Dicom attributes.

Before being cached or sent to the viewer, data passing through the I4DW are deidentified. Dicom headers are anonymized in a configurable way, by attribute, defaulting to the Dicom “Basic Application Level Confidentiality Profile” and retaining temporal and device identity information.

***GUI functionalities :*** Dicom unstructured and structured reports (“SR” modality) are displayed as documents, like any other eHOP document type. We added a set of specific functionalities to help users to easily navigate and visualize the imaging document :

- possibility to browse other documents belonging to the same Dicom study;

- possibility to view and handle images (Dicom series) through an integrated Dicom viewer (Papaya viewer library [3]);

- possibility to open a "Dicom Study View" that presents side by side the report and a viewer with a selector to display any study series (see Figure 3B).

***Evaluation:*** We evaluated the performance of our system for selecting a cohort and generating a dataset that was compared to an existing cohort of 271 patients with prostate cancer used as “gold standard”. The objective was to generate, from the data of 1.4 million patients available in our CDW, a datamart containing all clinical data including imaging reports and tests for this cohort. Inclusion criteria were (i) patients who underwent prostatectomy between 01/01/2014 and 31/12/2019, (ii) and patients with at least one pre-op MRI. Standard metrics of information retrieval (Precision, Recall, and F-measure) were computed to assess whether our system could retrieve all patients of the existing prostate cancer cohort and their imaging data.

We evaluated I4DW data transfer performance both in “routine” mode (i.e query a single image from the PACS for each Dicom series to obtain all metadata) and in “caching” mode (i.e all pixel data and metadata were retrieved from the PACS and images were cached in I4DW for future use).

[3] Papaya, JavaScript medical research image viewer https://github.com/rii-mango/Papaya



130

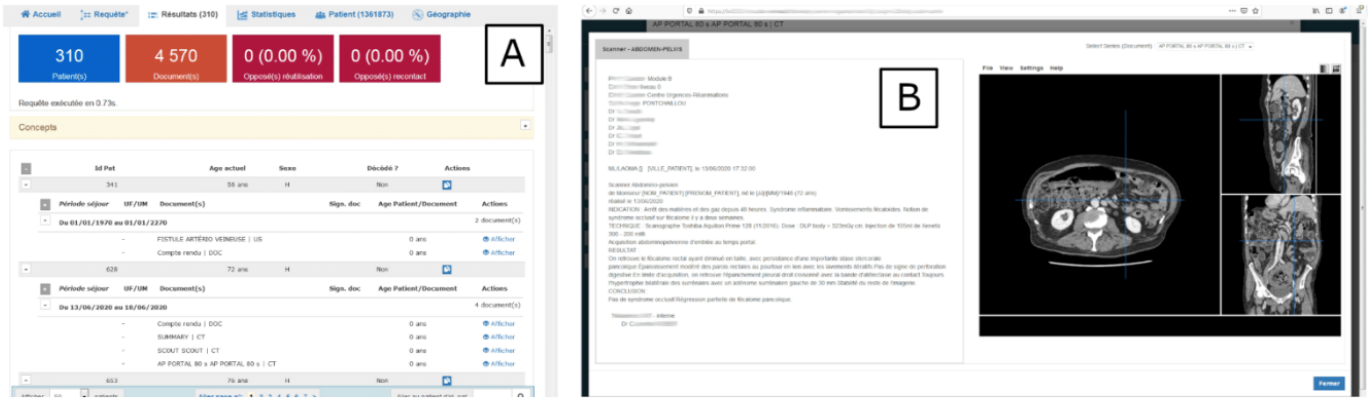



*Figure 3* - Screenshots of the eHop interface. **A** Result panel in eHOP. **B** The "Dicom Study" view.

## Results

***Data integration performance :*** The ETL sends queries to the I4DW server through a list of three accession numbers on each call, specifying whether to store the images in the cache (a CMOVE for all images is performed) or not (a CMOVE is performed for a single image per series). This limit of three accession numbers has been chosen because the I4DW response that contains the list of eHOP pivot files must not exceed a specific size limit. In function of the implied Dicom modalities (e.g MRI, CT, ultrasonography), each HTTP call will include different numbers of series.

Table 1 shows the mean durations (in seconds) of the image retrieval tasks for the whole series and for the first image of a series. For the first retrieval task, on a set of 1200 Dicom series, all images were retrieved and cached in the I4DW component. In routine mode, on a set of 300 Dicom series, only the first image of each DICOM series was retrieved to get all the metadata.

*Table 1* – I4DW query performance

| Tasks | Whole series | First image of each series |
|---|---|---|
| Total http request time per series (mean) | 7.74 s | 2.06 s |
| CMOVE Query time per series (mean) | 5.94 s | 0.9 s |

***Retrieval performance :***

We recreated the cohort in our CDW based on the inclusion criteria used for the "gold standard" prostate cancer cohort. Table 2 presents the precision, recall and F1-measure for this retrieval task.

*Table 2* - Retrieving an imaging cohort

| | Patients and their MRI |
|---|---|
| Precision | 0.95 |
| Recall | 1 |
| F1-measure | 0.975 |

All patients in the cohort were found and some patients who were not included in the cohort were found in addition in eHOP. Once this cohort is in our data warehouse, we need to classify the patients according to the magnetic field strength of the MRI. This is possible thanks to the integration and indexing of the attribute "Magnetic Field Strength" which comes from the dicom metadata.

***Query builder and data visualization :*** Users can query the integrated imaging data with all other data available in eHOP. The structured searching mode allows querying any attribute from the Dicom metadata. Figure 2 shows the query builder with the hierarchy of data elements organized according to the generated Dicom terminology. Each node is associated with the number of patients and documents available in the CDW or in the current datamart. According to the data element type, users can query using numerical or textual criteria. Relevant Dicom attributes ("Study Description", "Series Description" and "Body Part Examined") are indexed as text during the integration phase, and this allows the eHOP text searching mode to easily retrieve imaging tests. After query completion, results are sorted and displayed by patient and hospital stay (Figure 3A).

Users can browse, open and visualize clinical images to check the image quality or to ensure that the images in the cohort match his expectations, by viewing a document as they would view any other document in eHOP. It is also possible to directly check the consistency between the report and the Dicom study series thanks to the "Dicom study" view.

## Discussion

In this work, we presented the prototype we developed to integrate imaging data as a new information source for our CDW. The implemented system addresses the problem of exhaustive and semantic integration of imaging data. We tested the performance of the prototype by creating an imaging cohort and we demonstrated that this approach is feasible. The



131

integration of fine-grained data allowed the advanced query of imaging data with all the other data gathered in eHOP and leverages coded Dicom attributes (using LOINC or SNOMED). This metadata extraction method goes beyond in terms of data indexing than the CFIND based approach implemented in the study by Kaspar et al. [8]. We plan to keep the plugin approach to implement a set of services such as classification in the RadLex ontology [10].

As a limitation, our prototype was evaluated using a specific prostate cancer cohort. However, the system was designed independently of a specific clinical domain and can be used for many use cases.

The fine-grained integration of metadata generates a very rich collection of technical and clinical data elements. However, some are not useful to clinicians (e.g. Manufacturer), and some can be confusing (e.g. elements in depth in the hierarchy of a structured report). An improvement would be to customize the DCMEHOP terminology from a technical or clinical point of view.

Currently, metadata integration only covers the Dicom public attributes. The management of Dicom private attributes (vendor-defined attributes) is an important problem, as mentioned by Langer [5] and by Doran et al [6]. For some specific cases where these attributes are necessary, we could consider developing an I4DW plugin focused on the conciliation of private attributes from different vendors. This plugin could map attributes with the same meaning on a unique new DCMEHOP code.

Like in the study by Kaspar et al [8], our PACS does not implement the WADO and QIDO protocols [11], which seems still underused by vendors [12]. Theoretically, WADO might improve the ETL process performance because it does not require images to extract metadata. Although WADO is not designed to retrieve data in a bulk mode, it would be interesting to develop and evaluate a WADO-based retrieval approach.

We now plan to technically improve the ID4W component. Specifically, we want to optimize the ETL process by managing an asynchronous task system that will enable pooling more queries to the PACS and optimize the bulk metadata retrieval. We also want to improve the pixel data privacy using a ML-OCR method as recommended in good practice guidelines [13].

**Acknowledgments :**We would like to thank the French National Research Agency (ANR), for funding this work in the framework of the LabCom LITIS project (grant no. ANR-17-LCV1-0004) and the Canceropole GO for funding the Oncoshare project.

**References**

[1] Huang, SC., Pareek, A., Seyyedi, S. *et al.* Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit. Med.* **3,** 136 (2020). https://doi.org/10.1038/s41746-020-00341-z

[2] Murphy P, Koh DM. Imaging in clinical trials. *Cancer Imaging*. 2010;10 Spec no A(1A):S74-S82. Published 2010 Oct 4. doi:10.1102/1470-7330.2010.9027

[3] Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics*. 2007;5(1):11-34. doi:10.1385/ni:5:1:11

[4] E. Micard, D. Husson, and J. Felblinger, ArchiMed: A Data Management System for Clinical Research in Imaging, *Frontiers in ICT*. **3** (2016) 31. doi:10.3389/fict.2016.00031.

[5] Langer SG. Dicom Data Warehouse: Part 2. *J Digit Imaging*. 2016;29(3):309-313. doi:10.1007/s10278-015-9830-4

[6] Doran SJ, d'Arcy J, Collins DJ, et al. Informatics in radiology: development of a research PACS for analysis of functional imaging data in clinical research and clinical trials. *Radiographics*. 2012;32(7):2135-2150. doi:10.1148/rg.327115138

[7] Rajala T, Savio S, Penttinen J, et al. Development of a research dedicated archival system (TARAS) in a university hospital. *J Digit Imaging*. 2011;24(5):864-873. doi:10.1007/s10278-010-9350-1

[8] Kaspar M, Liman L, Ertl M, et al. Unlocking the PACS Dicom Domain for its Use in Clinical Research Data Warehouses. *J Digit Imaging*. 2020;33(4):1016-1025. doi:10.1007/s10278-020-00334-0

[9] Madec J, Bouzillé G, Riou C, et al. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. *Stud Health Technol Inform*. 2019;264:1536-1537. doi:10.3233/SHTI190522

[10] Lemordant P, Gibaud B, Garde C, et al. Ontology-based classification of radiological procedures for consistent sharing in Clinical Data Warehouses. 11th International Conference on Biomedical Ontologies (ICBO). Published in CEUR-WS Proceedings

[11] DICOMweb™ : https://www.dicomstandard.org/dicomweb (last accessed: 22 April 2021)

[12] Connectathon Results Browsing : https://connectathon-results.ihe.net/advanced.php (last accessed: 22 April 2021)

[13] Elngar, Ahmed, Ambika Pawar, and Prathamesh Churi, eds. *Data Protection and Privacy in Healthcare: Research and Innovations*. CRC Press, 2021.

**for correspondence**
Corresponding Author, Pierre Lemordant, Univ Rennes 1, Inserm, LTSI UMR 1099, Rennes, France; E-mail: pierre.lemordant@univ-rennes1.fr



132

# 4. Documentation de la librairie dicom-attribute-query

Cette librairie Java a été réalisée afin de faciliter la création de filtres sur les métadonnées DICOM. Elle a été conçue pour permettre à des utilisateurs peu habitués au standard DICOM de définir des contraintes sur les attributs DICOM avec une syntaxe simple. Cette librairie a été utilisée notamment dans I4DW pour la conception de profils d'anonymisation.

# DICOM Queries : Writing Rules

Rules are a set of conditions linked with boolean operators. These rules are meant to be applied to DICOM files (i.e lists of DICOM attributes) A String containing a defined rule can be parsed as a Criterion object. This criterion object holds the rule and allow to apply it to an attribute list to know if it matches the rule.

The rules have the following form, where keys are DICOM attributes :

```
key conditional_operator value [logical_operator keyN conditional_operatorN valueN]
```

Values can be mathematical expressions combining numbers and DICOM attributes.

## Keys

The keys used in these rules correspond to DICOM attributes. Find list of DICOM attributes here There are two ways to write them :

- Using **(dicom group number,dicom element number)** format :

```
(0008,1030) == 'this is a description'
```

- Using **#TagName** format :

```
#StudyDescription == 'this is a description'
```

Some Dicom attributes can have several values. The index concerned is defined using the syntax : key[index]

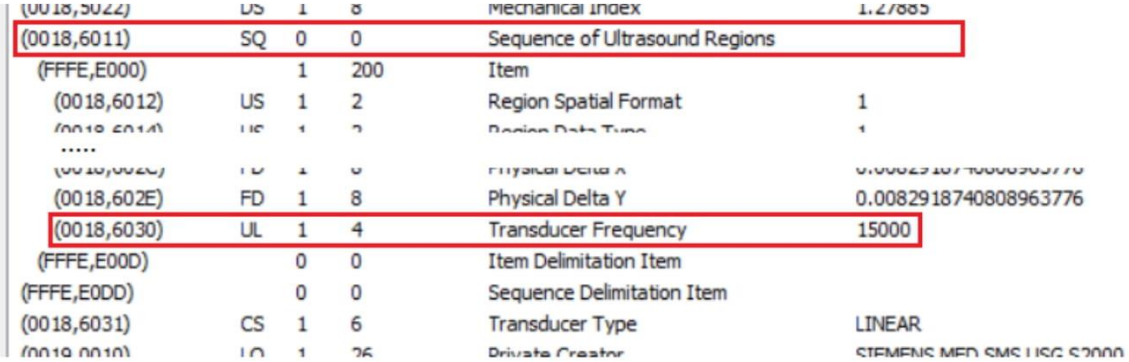
⚠ **index starts at 1**

```
(0008,0008)[3] == 'ANGIO'
#ImageType[3] == 'ANGIO'
```

It is possible to refer to fields contained in sequences by writing the following sub-attributes separated with a dot.

usage (format "TagName"): *#SequenceAttribute[.AnotherSequenceAttribute].Attribute operator value*

In the following example, a DICOM file contains a sequence of attribute "*Sequence of Ultrasound Regions*" (0018,6011) containing an item. This item contains the attribute "*Transducer Frequency*" (0018,6030) with value 15000.




To match files where the *Transducer Frequency* is greater than 12000, one would write this rule (both formats are allowed):

```
#SequenceOfUltrasoundRegions.TransducerFrequency > 12000
(0018,6011).(0018,6030) > 12000
```

⚠ if several items contain the targeted attribute, the rule will match if at least one of these attributes matches the value

# Logical Operators

There are two logical operators :

- **AND** (true if all sub elements are true)

usage : *key1 operator1 value1* ***AND*** *key2 operator2 value2 [* ***AND*** *keyα operatorα valueα ]*

```
#ImageType[3]=='ANGIO' AND #SOPClassUID=='1.2.840.10008.5.1.4.1.1.2'
```

- **OR** (true if one of the sub elements is true)

usage : *key1 operator1 value1* ***OR*** *key2 operator2 value2 [* ***OR*** *keyα operatorα valueα ]*

```
#ImageType[3] == 'ANGIO' OR #SOPClassUID == '1.2.840.10008.5.1.4.1.1.2' OR #Modality == 'CT'
```

⚠ Operators can't be mixed in the same group of elements. The following rules are incorrect and will produce a ParseError :

⛔ Parse Error : *key1 operator1 value1* ***AND*** *key2 operator2 value2* ***OR*** *key3 operator3 value3*

```
⛔ #ImageType[3] == 'ANGIO' AND #SOPClassUID == '1.2.840.10008.5.1.4.1.1.2' OR #Modality == 'CT'
```

⚠ Parenthesis are necessary when using logical operator between two groups

*( key1 operator1 value1 AND key2 operator2 value2 ) OR key3 operator3 value3*

```
✓ (#ImageType[3]=='ANGIO' AND #SOPClassUID == '1.2.840.10008.5.1.4.1.1.2') OR #Modality == 'CT'
```

*key1 operator1 value1 OR ( key2 operator2 value2 AND key3 operator3 value3 )*

```
✓ #ImageType[3]=='ANGIO' OR (#SOPClassUID == '1.2.840.10008.5.1.4.1.1.2' AND #Modality=='CT')
```

# Criterion Operators

there are different criterion operators depending on the type of the target field, *String* or *Number*.

⚠ The parsing operation do not check the coherence between DICOM tag VR and the operator's expected type. For instance, the following rule applies the number operator \"=\" on a text field, it will be parsed successfully but the classification will probably fail when trying to parse the value of StudyDescription into a number.

```
⛔ #StudyDescription = 3
```

## String Operators

Values used with String operators have to be quoted using '

There are two string operators :

- Equal (for String) with symbol : **==**

```
#StudyDescription == 'SCOUT'
```

- Regex with symbol : **~**

```
#StudyDescription ~ '\[Chir\].*'
```

## Number Operators

There are five number operators :

- Equal (for numbers) with symbol : **=**

```
#ImageType[4] = 1
```

- Superior with symbol : **>**

```
#PregnancyStatus > 3
```

- Inferior with symbol : **<**

```
#PregnancyStatus < 3
```

- Superior or equal with symbol : **>=**

```
#PregnancyStatus >= 4
```

- Inferior or equal with symbol : **<=**

```
#PregnancyStatus <= 2
```

## Mathematical expressions

Values used with number operators can be mathematical expressions. These expressions can contain numbers or attributes with compatible Value representations. The following VRs can be parsed as numbers : SL, DS, FL, FD, IS, OD, SS, UL, US.

Mathematical expressions can use the following operators :

- Multiplication with symbol : *

```
#RepetitionTime <= #EchoTime * 5
```

- Division with symbol : /

```
#EchoTime <= #RepetitionTime / 5
```

- Addition with symbol : **+**

```
#RepetitionTime <= #EchoTime + 500
```

- Substraction with symbol : **-**

```
#EchoTime < #RepetitionTime - 450
```

⚠ Operators with different priority levels cannot be mixed. The following rules are incorrect and will produce a ParseError :

⛔ Parse Error : *key operator element1 * element2 + element3*

```
⛔ #RepetitionTime <= #EchoTime * 2 + 75
```

⚠ Parenthesis are necessary when using different priority levels

*key operator **(** element1 * element2 **)** + element3*

```
✓ #RepetitionTime <= ( #EchoTime * 2 ) + 75
```

*key operator element1 - **(** element2 / element3 **)***

```
✓ #EchoTime <=  600 - ( #RepetitionTime / 2 )
```

# 5. Démonstration du module

Deux vidéos de démonstration sont accessibles sur les liens suivants :

La structure de données DCMEHOP : https://youtu.be/wO7Qc-MsVa4

L'interface pour l'imagerie : https://youtu.be/FBGMtfinIuc

## 5.1. Utilisation de la terminologie DCMEHOP

La première vidéo présente l'utilisation de la terminologie DCMEHOP. Un utilisateur réalise une requête simple en utilisant la terminologie DCMEHOP. Il ouvre le nœud correspondant au Rapport Structuré qui l'intéresse, ici un « X-Ray Radiation Dose Report » et parcoure la structure du rapport pour trouver le champ sur lequel il veut appliquer une contrainte, ici le champ « Total Number of Irradiation Events » dans la section « CT Accumulated Dose Data ». Cet attribut étant de type numérique, eHOP lui propose de choisir un intervalle de valeur. L'utilisateur cherche les rapports pour lesquelles cette valeur est comprise entre 6 et 8 inclus.

Après avoir exécuté la requête l'utilisateur parcoure les résultats et sélectionne un rapport retourné par la requête. On trouve dans ce document les attributs DICOM du premier niveau dans la hiérarchie du fichier, rassemblés dans un tableau, puis le contenu du rapport formatté pour être facilement lisible. L'attribut qui est l'objet de la requête est automatiquement surligné par eHOP. On constate que la valeur de l'attribut répond bien au critère défini.

## 5.2. Les vues accessibles pour les documents d'imagerie eHOP

Dans la deuxième vidéo, l'utilisateur utilise le requêteur eHOP pour chercher les documents de type « DICOM » dont le titre contient le mot « THORAX ». L'utilisateur ouvre un document retourné par la requête. On constate que le mot recherché est automatiquement surligné par eHOP. Dans le panneau supérieur 3 options sont disponibles :

**Voir les images** : Ouvre un viewer DICOM (Papaya) traite automatiquement les images pour reformer des vues en coupes. Ainsi même si l'acquisition a été faite sur un axe X, l'image peut être parcourue sur les axes Y et Z. il est possible de mettre chacune des trois vues sur la plus grande des trois zones d'affichage.

**Consulter un document de la même study DICOM** : permet d'ouvrir, à la place du document actuel, un autre document provenant du même examen d'imagerie, c'est-à-dire un document correspondant à une *DICOM series* ou le compte-rendu d'imagerie.

**Ouvrir la vue DICOM STUDY** : Ouvre une vue qui présente à gauche les éléments textuels de l'examen, soit le compte-rendu ou les rapports structurés DICOM, et à droite le viewer DICOM sur lequel il est possible d'ouvrir les acquisitions de l'examen. Cette vue permet de lire les informations textuelles et vérifier en même temps leur concordance avec les images.

UNIVERSITÉ DE RENNES 1

**Titre :** Méthodes d'intégrations sémantiques de données patients multi domaines dans des architectures d'entrepôts de données de santé : Une approche orientée cas d'usage

**Mots clés :** intégration sémantique, imagerie, entrepôt de données de santé

**Résumé :** Dans le domaine de l'imagerie médicale, les évolutions techniques (stockage de données, modalités d'imagerie, …) et méthodologiques (médecine personnalisée, évolution de l'imagerie diagnostique et interventionnelle, …) des dernières décennies ont fait apparaitre des enjeux majeurs concernant l'usage de l'imagerie dans le soin et dans la recherche, notamment de l'utilisation secondaire de cette source de données.
L'objectif de cette thèse a été de développer des méthodes afin de rendre intégrables et réutilisables ces données d'imagerie dans une solution d'entrepôt de données de santé et de mettre en œuvre cette intégration sémantique dans le cadre concret de la solution d'entrepôt développée dans notre laboratoire.
Dans ce travail, nous avons étudié les outils et méthodes d'alignement de terminologie locales et de références pour permettre la réutilisation et le partage des données d'imagerie. Nous avons conçu une preuve de concept d'outil de classification des examens d'imagerie utilisant le raisonnement ontologique. Enfin, nous avons développé et déployé un prototype de module d'intégration sémantique des données d'imagerie permettant de gérer le trajet des données depuis le PACS jusqu'à l'entrepôt.

**Title :** Methods for semantic integration of multi-domain patient data in clinical data warehouse architectures: A use case oriented approach

**Keywords :** semantic integration, imaging, clinical data warehouse

**Abstract :** In the field of medical imaging, technical (data storage, imaging modalities, etc.) and methodological (personalised medicine, evolution of diagnostic and interventional imaging, etc.) developments over the last few decades have raised major challenges concerning the use of imaging in care and research, particularly the secondary use of this data source.
The objective of this work was to develop methods to integrate and reuse this imaging data in a clinical data warehouse solution and to implement this semantic integration in the framework of the warehouse solution developed in our laboratory.
In this work, we studied tools and methods for local terminology alignment with reference terminology to enable the reuse and sharing of imaging data. We designed a proof of concept for an imaging classification tool using ontological reasoning. Finally, we developed and deployed a prototype module for semantic integration of imaging data to manage the data path from the PACS to the clinical data warehouse.